

Vers une formalisation des décompositions sémantiques dans la Grammaire d'Unification Sens-Texte

François LAREAU

Lattice - U. Paris 7, UFRJL, Case 7003, 2 pl. Jussieu, 75251 Paris cedex 5
OLST - U. de Montréal, Ling., CP 6128 succ C.-V., Montréal QC, H3C 3J7
francois.lareau@umontreal.ca

Résumé. Nous proposons une formalisation de la décomposition du sens dans le cadre de la Grammaire d'Unification Sens-Texte. Cette formalisation vise une meilleure intégration des décompositions sémantiques dans un modèle global de la langue. Elle repose sur un jeu de saturation de polarités qui permet de contrôler la construction des représentations décomposées ainsi que leur mise en correspondance avec des arbres syntaxiques qui les expriment. Le formalisme proposé est illustré ici dans une perspective de synthèse, mais il s'applique également en analyse.

Abstract. We propose a formal representation of meaning decomposition in the framework of the Meaning-Text Unification Grammar. The proposed technique aims at offering a better integration of such semantic decompositions into a global model of the language. It relies on the saturation of polarities to control the construction of decomposed representations as well as their mapping to the syntactic trees that express them. The proposed formalism is discussed from the viewpoint of generation, but it applies to analysis as well.

Mots-clés : Grammaire d'Unification Sens-Texte, Théorie Sens-Texte, sémantique, représentation du sens, paraphrasage.

Keywords: Meaning-Text Unification Grammar, Meaning-Text Theory, semantics, representation of meaning, paraphrasing.

1 Introduction

Dans le cadre de la théorie Sens-Texte (Mel'čuk, 1997; Kahane, 2001) (ci-après, TST), le sens d'un énoncé est représenté par un réseau de sémantèmes. Ces sémantèmes peuvent être décomposés (c'est-à-dire remplacés par leur définition) mais ce sont en général des représentations réduites (c'est-à-dire dont les sémantèmes ne sont pas décomposés) qui sont utilisées pour représenter le sens d'un énoncé. La principale raison pour laquelle on ne décompose pas les représentations sémantiques est qu'on ne saurait pas où s'arrêter. En effet, jusqu'à quel point faut-il décomposer ? Il n'y a à cette question que deux réponses possibles si on veut éviter l'arbitraire : soit on décompose jusqu'aux sens primitifs, ce que fait notamment (Wierzbicka, 1996), soit on ne décompose pas du tout, ce qui est la position de la TST (Mel'čuk, 1988; Mel'čuk, 1989). Ainsi, bien qu'ils soient des paraphrases, les trois énoncés suivants n'ont pas la même représentation sémantique, comme le montre la Figure 1.

- (1) a. *Une Torontoise rousse.*
 b. *Une rousse qui habite Toronto.*
 c. *Une femme rousse qui habite Toronto.*

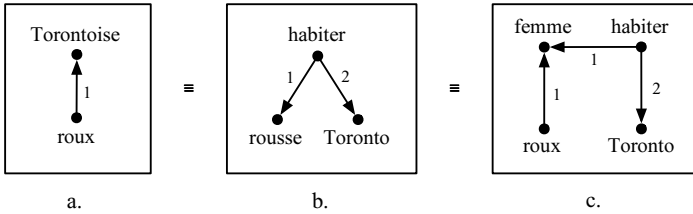


FIG. 1 – Trois paraphrases qui n’ont pas la même représentation sémantique

Ces trois énoncés expriment exactement les mêmes sens, soit ceux visibles à la Figure 1c, mais ils diffèrent par la façon dont ils les regroupent. Dans le premier les sens ‘femme’, ‘habiter’ et ‘Toronto’ sont exprimés par un même nom, $TORONTOISE_{(N)}$ (cf. Figure 1a), alors que dans le second ce sont ‘femme’ et ‘roux_(Adj)’ qui sont regroupés et exprimés par le nom $ROUSSE_{(N)}$ (cf. Figure 1b), tandis que dans le troisième énoncé à chaque sémantème correspond un lexème distinct (cf. Figure 1c). Bien entendu, on peut démontrer que leurs représentations sémantiques respectives sont strictement équivalentes, tout simplement par le biais de ce que (Milićević, sous presse) appelle des règles d’expansion / réduction ordinaires, c’est-à-dire des règles qui mettent en équivalence les sémantèmes et leur définition. La Figure 2 montre un exemple d’une telle règle.

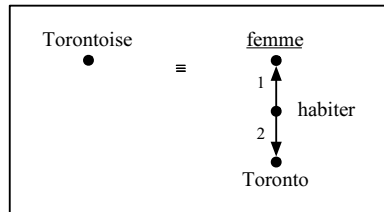


FIG. 2 – Une règle d’équivalence sémantique de la TST

Toutefois, ces règles soulèvent au moins trois questions quant à leur formalisation :

1. Où se situent les règles d’équivalence sémantique par rapport aux différents modules d’un modèle linguistique ?
2. Comment peuvent-elles interagir avec les autres règles du modèle ?
3. Comment peuvent-elles être formalisées ?

Nous tenterons d’y répondre en nous situant dans le cadre de la Grammaire d’Unification Sens-Texte, dont nous proposons d’abord un bref rappel. Nous suggérons une solution qui ne fait appel à aucun mécanisme extérieur et qui ne nécessite aucun changement au formalisme. Nous

verrons également brièvement comment notre solution peut servir à typer les actants sémantiques, si désiré.

2 La Grammaire d'Unification Sens-Texte

La Grammaire d'Unification Sens-Texte (ci-après, GUST) a d'abord été proposée par (Kahane, 2002), puis formalisée sous forme de grammaire d'unification polarisée (Kahane, 2004) par (Kahane & Lareau, 2005). Puisqu'elle est basée sur la TST, GUST considère plusieurs niveaux de représentation (sémantique, syntaxique, morphotopologique et phonologique). À chaque niveau correspond une grammaire de bonne formation qui décrit les structures possibles (ces grammaires sont appelées respectivement $\mathcal{G}_{\text{sém}}$, $\mathcal{G}_{\text{synt}}$, $\mathcal{G}_{\text{morph}}$ et $\mathcal{G}_{\text{phon}}$). Les niveaux sont ordonnés et la correspondance entre deux structures de niveaux adjacents est assurée par des grammaires d'interface ($\mathcal{I}_{\text{sém-synt}}$, $\mathcal{I}_{\text{synt-morph}}$ et $\mathcal{I}_{\text{morph-phon}}$) qui mettent en correspondance des fragments élémentaires de structures.

Tous les nœuds, arcs et autres objets des grammaires de bonne formation ou d'interface sont associés à un certain nombre de polarités. Contrairement à (Kahane & Lareau, 2005), nous utilisons un système à deux valeurs de polarités (plutôt que trois), soit \square et \blacksquare . Parmi ces deux valeurs, la noire est dite neutre, c'est-à-dire qu'une structure dont tous les objets ne portent que des polarités de valeur noire est considérée comme saturée, alors qu'une polarité blanche déclenche obligatoirement l'unification de son porteur avec d'autres objets afin d'être neutralisée. L'unification de deux objets entraîne en effet le calcul d'une nouvelle valeur pour chacune des polarités qu'ils portent de la façon suivante : $\square \cdot \square = \square$ et $\square \cdot \blacksquare = \blacksquare \cdot \square = \blacksquare$ (l'unification de deux objets noirs n'est pas permise). C'est ainsi que, par unification, les structures seaturent. C'est en fait la neutralisation des polarités qui contrôle tout le processus de synthèse ou d'analyse linguistique. Cela s'opère notamment par les polarités d'interface, c'est-à-dire des polarités non neutres qui sont ajoutées aux objets d'une grammaire dans le but de déclencher le module suivant dans le modèle linguistique. Par exemple, les objets de $\mathcal{G}_{\text{sém}}$, en plus de porter une polarité $p_{\text{sém}}$, qui contrôle la saturation des structures sémantiques, portent tous une polarité $p_{\text{sém-synt}}$ blanche. Les règles de $\mathcal{I}_{\text{sém-synt}}$ étant les seules à contenir des objets de polarité $p_{\text{sém-synt}}$ noire, ce module sera automatiquement appliqué, de façon à neutraliser les structures produites par $\mathcal{G}_{\text{sém}}$. Nous ne pouvons pas illustrer en détail ce mécanisme dans le présent article, aussi renvoyons-nous le lecteur à (Kahane & Lareau, 2005). Nous verrons qu'il est au cœur de la formalisation que nous proposons ici pour les règles d'équivalence sémantique.

3 La place des règles d'équivalence sémantique dans GUST

Dans le cadre de la TST, les règles d'équivalence sémantique ne font partie d'aucun des modules de transition qui constituent un modèle Sens-Texte, puisqu'elles encodent des phénomènes orthogonaux aux processus de synthèse et d'analyse simulés par les règles de correspondance. Pour les mêmes raisons, elles ne peuvent pas constituer une grammaire d'interface dans GUST, puisqu'elles ne mettent pas en relation des objets de niveaux adjacents. En fait, ces règles ne peuvent même pas constituer un module distinct, puisqu'on ne sait pas a priori jusqu'à quel niveau on doit décomposer (ou réduire) un sémantème (ou une configuration de sémantèmes). Il

faudrait donc que ce module puisse s'appliquer à son propre résultat de façon réursive, ce qui est difficilement compatible avec le système de polarités utilisé dans GUST. Ainsi, nous proposons plutôt d'intégrer les règles d'expansion / réduction à la grammaire de bonne formation sémantique de GUST.

4 Une représentation formelle du sens en GUST

Afin d'intégrer les règles d'expansion / réduction à $\mathcal{G}_{\text{sém}}$, nous devons légèrement modifier la caractérisation de celle-ci donnée par (Kahane & Lareau, 2005), sans pour autant en modifier le formalisme. Les nœuds sémantiques seront ici conçus comme des ensembles qui peuvent contenir d'autres nœuds. Les nœuds inclus à l'intérieur d'un autre nœud sont reliés par des arcs et forment de petites structures sémantiques qui représentent les constituants du sémantème ainsi décomposé. Les nœuds qui décomposent un autre nœud sont reliés à ce dernier par la fonction *réduction*, qui associe à un nœud l'ensemble dont il fait partie (c'est-à-dire un autre nœud). Parmi les nœuds qui forment la décomposition d'un sémantème, un seul est identifié comme le sens principal. Il s'agit du genre prochain et il est identifié par la fonction *générique*, qui prend comme argument un sémantème (un nœud) et retourne son genre prochain (un autre nœud).¹ Graphiquement, nous représentons les décompositions à l'intérieur d'une bulle en pointillés liée au nœud réduit, dans laquelle le genre prochain est souligné. Par souci de lisibilité, nous réduisons la taille des caractères et des traits à l'intérieur de ces bulles. La Figure 3 encode la règle d'équivalence ('Torontoise') \equiv 'femme qui habite Toronto', tant sous forme explicite² que graphique.³

```

{ 1:nœudsém, 2:nœudsém, 3:nœudsém,
  4:nœudsém, 5:arcsém, 6:arcsém },
{ source(5) = 2, cible(5) = 3,
  source(6) = 2, cible(6) = 4,
  étiquette(1) = "Torontoise",
  étiquette(2) = "habiter", étiquette(3) = "femme",
  étiquette(4) = "Toronto", étiquette(5) = "I",
  étiquette(6) = "2",
  psém(1) = N, psém(2) = B, psém(3) = B,
  psém(4) = B, psém(5) = B, psém(6) = B,
  psém-synt(1) = B, psém-synt(2) = N, psém-synt(3) = N,
  psém-synt(4) = N, psém-synt(5) = N, psém-synt(6) = N,
  générique(1) = 3, réduction(2) = 1,
  réduction(3) = 1, réduction(4) = 1 }

```

\equiv

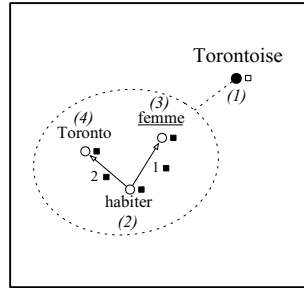


FIG. 3 – La décomposition sémantique de ('Torontoise')

¹Il n'est pas évident que l'on puisse toujours identifier un seul sémantème comme genre prochain pour tous les sens de la langue. Nous ne savons pas pour l'instant comment encoder les cas où le genre prochain est une configuration de sémantèmes ou où il y a plusieurs genres prochains.

²Les valeurs « N » et « B » des fonctions de polarités renvoient respectivement aux polarités ■ et □.

³Toutes les figures qui suivent montrent deux polarités pour chaque objet : $p_{\text{sém}}$ (qui contrôle la saturation des structures sémantiques), indiquée par la couleur de l'objet, et $p_{\text{sém-synt}}$ (qui assure l'interface avec $\mathcal{I}_{\text{sém-synt}}$), indiquée par un petit carré à côté de l'objet. Par ailleurs, dans la Figure 3, nous ajoutons aux nœuds de la version graphique (entre parenthèses et en italique) les numéros d'identification utilisés dans la version textuelle afin que le lecteur puisse plus facilement mettre en relation les deux formes de représentation.

Les sémantèmes ‘rousse_(N)’ et ‘femme’ peuvent se décomposer de la même façon, comme le montre la Figure 4. On remarquera que les nœuds à l’intérieur des décompositions portent toujours une polarité $p_{sém}$ blanche. Ils forcent ainsi l’application d’autres règles de $\mathcal{G}_{sém}$ pour vérifier leur bonne formation. Par exemple, le nœud ‘femme’ apparaissant dans la décomposition de ‘rousse_(N)’ forcera l’application de la règle qui en vérifie la bonne formation (à droite dans la Figure 4), introduisant par le fait même la décomposition de ce sémantème.

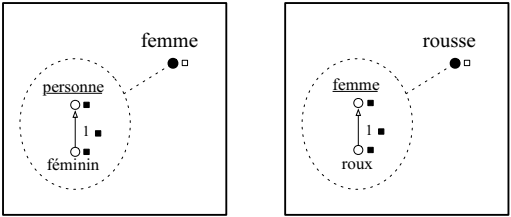


FIG. 4 – La décomposition sémantique de ‘femme’ et de ‘rousse_(N)’

On remarquera également que les nœuds à l’intérieur des décompositions sont toujours saturés en $p_{sém-synt}$. Cela a pour but d’éviter de réaliser à la fois un sens et sa définition dans le même énoncé. Lorsqu’un fragment de la structure sémantique a été réduit, tous ses éléments deviennent invisibles pour l’interface $\mathcal{I}_{sém-synt}$, qui ne peut plus voir que le sémantème réduit.

Ainsi, toutes les règles de $\mathcal{G}_{sém}$ qui construisent un nœud devront également en fournir la décomposition. Seuls les primitifs sémantiques⁴ font exception, puisqu’ils ne peuvent pas être décomposés. Pour les besoins de notre exposé, nous allons considérer que ‘(de sexe) féminin’, ‘(aux cheveux) roux_(Adj)’, ‘habiter’, ‘Toronto’ et ‘personne’ sont des sens indécomposables.

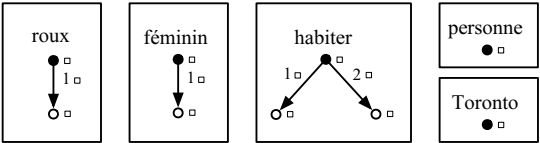


FIG. 5 – Les sens primitifs dans la grammaire sémantique

Voyons maintenant comment ces règles s’appliquent en synthèse. D’abord, les sens « primitifs » décrits à la Figure 5 nous permettent de construire la représentation de la Figure 6.

⁴Nous partons du postulat, hérité de la TST, que toute langue comporte un certain nombre de sens indécomposables, dits « primitifs » (qui ne sont pas forcément universels). Notre formalisme ne peut pas fonctionner si on admet la circularité dans les descriptions sémantiques. Cette propriété est d’ailleurs souhaitable à nos yeux puisqu’elle permet de vérifier la non-circularité de la grammaire sémantique.

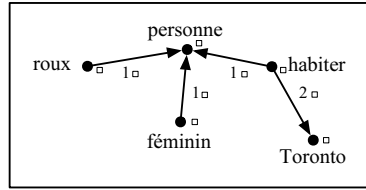


FIG. 6 – Une représentation sémantique décomposée à réaliser

Il est parfaitement possible d'utiliser cette représentation sémantique telle quelle et de la mettre en correspondance avec un arbre syntaxique sans la réduire aucunement. On obtient alors l'énoncé (2).

(2) *Une personne rousse de sexe féminin qui habite Toronto.*

On peut également opérer une première réduction en appliquant la définition de 'femme' (voir la Figure 4), ce qui donne la structure de la Figure 7.

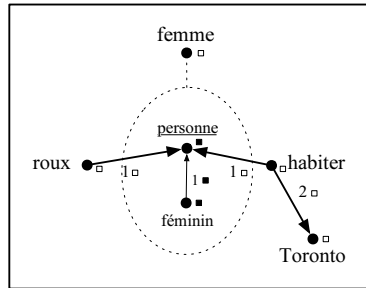


FIG. 7 – La réduction de 'personne (de sexe) féminin'

Cette représentation ne peut toutefois pas être mise en correspondance avec un arbre syntaxique, puisque du point de vue de l'interface sémantique-syntaxe elle n'est pas connexe ('personne' étant saturé en $p_{\text{sém-synt}}$, il n'est plus visible pour l'interface). La structure qui en résulterait ne serait pas un arbre et serait donc rejetée (Kahane & Lareau, 2005). Pour rendre cette structure connexe, il faut en quelque sorte « déplacer » les relations sémantiques qui pointaient vers 'personne' pour qu'elles portent maintenant sur 'femme'. Nous devons donc introduire une règle de propagation qui recopie les relations sémantiques pointant vers un nœud pour qu'elles pointent vers le sémantème réduit dont ce nœud est le genre prochain⁵ (voir Figure 8). Une fois cette règle appliquée, on obtient la structure de la Figure 9.

⁵Cette règle ne concerne que le genre prochain. En effet, si une relation pointe vers un autre sémantème, la réduction n'est pas possible. Par exemple, 'femme très rousse' ne peut pas être réduit sans perte d'information [**une très rousse_(N)*], alors que 'grande femme rousse' ne pose aucun problème [*une grande rousse_(N)*].

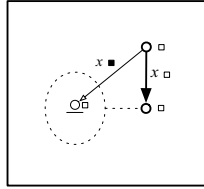


FIG. 8 – Une règle de propagation du gouverneur

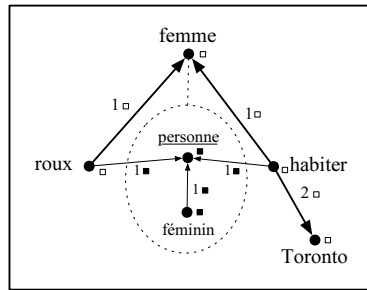


FIG. 9 – La réduction de 'personne (de sexe) féminin' avec propagation des gouverneurs

Les relations pointant vers 'personne' ont été neutralisées du point de vue de l'interface sémantique-syntaxe par la saturation de leur polarité $p_{\text{sém-synt}}$. La structure résultante peut tout de suite être mise en correspondance avec un arbre syntaxique, ce qui donne l'énoncé (3).

(3) *Une femme rousse qui habite Toronto.*

Elle peut aussi être réduite encore une fois en appliquant la règle qui construit 'rousse_(N)'. Cela nous donne la structure de la Figure 10a, puis celle en 10b une fois la règle de propagation des gouverneurs appliquée. Il résulte que les seuls sémantèmes accessibles à $\mathcal{I}_{\text{sém-synt}}$ (donc demandant à être exprimés), sont 'rousse_(N)', 'habiter' et 'Toronto' ce qui permet de former l'énoncé (4).

(4) *Une rousse qui habite Toronto.*

On aurait pu réduire les sémantèmes autrement, en utilisant d'autres règles d'équivalence sémantique. Dans tous les cas, on retrouverait toujours la configuration sémantique de la Figure 6, mais de nouveaux sémantèmes réduits apparaîtraient et les sens accessibles à $\mathcal{I}_{\text{sém-synt}}$ ne seraient pas les mêmes.

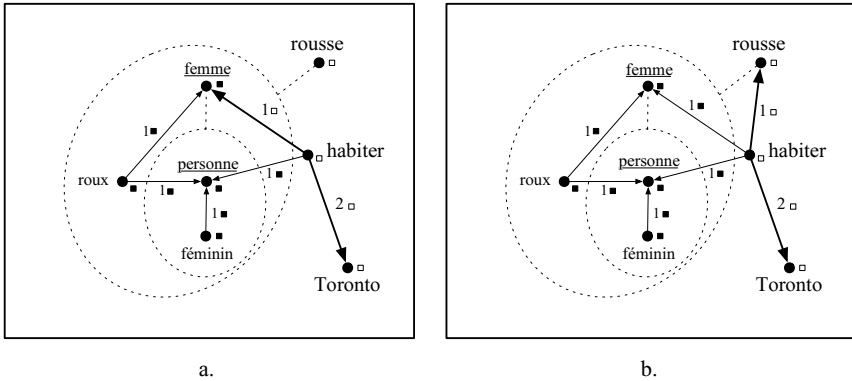


FIG. 10 – La réduction de ‘femme rousse’ avec propagation des gouverneurs

5 Le typage des actants sémantiques

Le type de représentation sémantique que nous suggérons ici permet de représenter de façon assez naturelle le typage des actants sémantiques. Si on souhaite restreindre la combinatoire sémantique d’un prédicat, on peut en étiqueter les actants dans la règle de $\mathcal{G}_{\text{sém}}$ qui le construit. Par exemple, si on veut que le prédicat ‘habiter’ ne puisse prendre comme premier actant qu’un sémantème dénotant une personne et comme deuxième actant un sémantème qui dénote un lieu, il suffit de modifier la règle (que nous avons déjà vue à la Figure 5) qui introduit ce sémantème de la façon illustrée à la Figure 11.

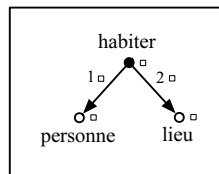


FIG. 11 – Le typage des actants sémantiques en GUST

Sans avoir accès à la décomposition du sens, le premier actant de ‘habiter’ ne pourrait pas, par exemple, être ‘rousse_(N)’ puisque l’étiquette ‘personne’ du nœud en Figure 11 ne peut pas directement s’unifier avec l’étiquette ‘rousse_(N)’. Par contre, avec les règles que nous avons présentées, puisque ‘personne’ est le genre prochain de ‘femme’, qui lui-même est le genre prochain de ‘rousse_(N)’, ce dernier pourra, grâce à la règle de propagation des gouverneurs, être le premier actant de ‘habiter’, sans qu’il ne soit nécessaire d’activer une quelconque procédure additionnelle ou de faire appel à une structure de donnée séparée qui encode une hiérarchie des types sémantiques. Comme on peut l’observer à la Figure 10b, le formalisme permet en effet de vérifier que ‘personne’ est une composante centrale du premier actant de ‘habiter’, et ce peu

importe à quelle « profondeur » se trouve enfouie cette composante dans le sens de l'actant. Bien entendu, puisque nous n'avons pas fourni dans nos règles la décomposition sémantique de 'Toronto', le mécanisme ne pourrait pas fonctionner pour le deuxième actant et 'habiter', tel que défini ici, ne pourrait pas se combiner à 'Toronto'. Mais si nous complétons notre grammaire pour fournir la décomposition de ce sémantème, 'habiter' pourra l'accepter comme deuxième actant puisque 'Toronto' \supset 'ville' \supset 'lieu'.

Insistons sur le fait que le formalisme n'oblige pas à typer les actants d'un sémantème. On peut très bien ne pas étiqueter les actants dans les règles qui introduisent les sémantèmes prédicatifs sans que n'en souffre la grammaire. Le modèle permet alors de mettre en correspondance des phrases absurdes avec des représentations sémantiques formellement correctes, mais difficiles à interpréter. Revient-il à la grammaire de rejeter de tels énoncés ? D'un côté, il est vrai que la langue permet d'exprimer même des idées absurdes ou contradictoires. Si la grammaire n'est qu'un modèle fonctionnel de la langue, alors elle doit permettre de telles phrases. Le typage des actants risque, sans mécanisme approprié⁶, de rendre la grammaire trop rigide pour permettre les écarts et glissements sémantiques. D'un autre côté, le typage des actants semble modéliser une certaine connaissance que les locuteurs ont des sens de la langue. De fait, dans le cadre de la TST, Mel'čuk et Polguère n'hésitent pas à y avoir recours dans leur base de données lexicales *DiCo* (Polguère, 2003). En outre, cela peut s'avérer utile pour diverses applications, notamment en linguistique computationnelle (par exemple pour valider une analyse incertaine). Enfin, la question dépasse le cadre de notre exposé, mais quoi qu'il en soit, le formalisme permet d'implémenter, au moins dans une certaine mesure, les deux solutions : le linguiste choisira à sa guise de typer ou non les actants des prédicats qu'il décrit dans la grammaire de bonne formation sémantique.

6 Conclusion

Le type de règle que nous proposons permet, sans aucune procédure particulière et sans avoir à ajouter ni nouveau module ni nouvelle polarité au modèle, de trouver automatiquement toutes les réductions possibles pour une représentation sémantique donnée, et donc toutes ses différentes expressions. Nous avons illustré le fonctionnement de ces règles en synthèse, mais elles fonctionnent également en analyse. Elles fournissent alors, pour l'énoncé analysé, non seulement sa représentation sémantique réduite, mais également sa décomposition complète ainsi que toutes les décompositions intermédiaires. Cette caractéristique du formalisme proposé peut s'avérer fort intéressante notamment en traduction automatique, puisque la traduction des lexèmes et des sens grammaticaux doit souvent être approximative, ce qui demande d'avoir accès à la décomposition des sens.

Remerciements

Nous tenons à remercier les relecteurs anonymes pour leurs commentaires constructifs ainsi que Dominique Longin pour son aide précieuse lors de l'édition de la version finale de cet article.

⁶(Pustejovsky, 1995) propose justement de tels mécanismes.

Références

- KAHANE S. (2001). Grammaires de dépendance formelles et théorie sens-texte : Tutoriel. In D. MAUREL, Ed., *Actes de TALN 2001 (Traitement automatique des langues naturelles)*, volume 2, p. 17–76, Tours : ATALA.
- KAHANE S. (2002). *Grammaire d'Unification Sens-Texte : vers un modèle mathématique articulé de la langue*. Thèse d'habilitation à diriger des recherches, Université Paris 7.
- KAHANE S. (2004). Grammaires d'unification polarisées. In P. BLACHE, Ed., *Actes de TALN 2004 (Traitement automatique des langues naturelles)*, p. 233–242 : ATALA LPL.
- KAHANE S. & LAREAU F. (2005). Grammaire d'unification sens-texte : modularité et polarisation. In *Actes de TALN 2005 (Traitement automatique des langues naturelles)*, p. 23–32, Dourdan : ATALA LIMSI.
- MEL'ČUK I. (1988). Semantic description of lexical units in an explanatory combinatorial dictionary : basic principles and heuristic criteria. *International Journal of Lexicography*, **1**, 165–88.
- MEL'ČUK I. (1989). Semantic primitives from the viewpoint of the meaning-text linguistic theory. *Quaderni di semantica*, **10**, 65–102.
- MEL'ČUK I. (1997). *Vers une Linguistique Sens-Texte : Leçon inaugurale au Collège de France*. Paris : Collège de France.
- MILIĆEVIĆ J. (sous presse). *Modélisation de la paraphrase langagière*. Bern : Peter Lang.
- POLGUÈRE A. (2003). Étiquetage sémantique des lexies dans la base de données dico. *Traitement automatique des langues*, **44**(2), 39–68.
- PUSTEJOVSKY J. (1995). *The generative lexicon*. Cambridge / London : The MIT Press.
- WIERZBICKA A. (1996). *Semantics : primes and universals*. Oxford : Oxford University Press.