

Document d'accompagnement

Coralie REUTENAEUR

Transparent 2

Une langue vivante est un système dynamique. Tout mot, ou plus précisément toute unité lexicale, est susceptible d'évoluer sémantiquement. Par exemple, les emplois d'une unité lexicale peuvent se diffuser d'un domaine particulier vers des emplois plus généraux. Ainsi, avant la catastrophe de décembre 2004 en Asie du Sud-Est, les emplois de *tsunami* étaient rares et presque exclusivement en sciences de la terre, pour qualifier un phénomène ou une catastrophe naturels. Depuis, *tsunami* est employé massivement pour désigner des bouleversements majeurs, tels que le *tsunami bleu* désignant la victoire électorale de l'UMP aux législatives de 2007 ou encore le *tsunami financier* de la crise financière. D'autres unités lexicales connaissent de nouveaux emplois spécifiques à un domaine précis, comme *toxique* en finance pour qualifier certains types de produits ou instruments financiers.

Transparent 3

Chaque individu est conscient de tels phénomènes et pourrait citer l'un ou l'autre exemple. En revanche, de tels phénomènes restent délicats à mesurer à l'aide d'outils informatiques ou mathématiques.

Transparent 4

Deux problèmes se présentent au niveau des traitements automatiques : d'une part, le problème de la détection d'un nouveau sens, pour déterminer s'il existe un nouveau sens ; d'autre part, le problème de la qualification du nouveau sens, pour définir ce qu'est ce nouveau sens et en proposer une description. Le second problème, à savoir la qualification du nouveau sens, est au cœur de cette thèse.

Transparent 5

Je défends l'idée qu'il est possible de construire de façon semi-automatique une représentation nuancée et structurée d'un nouveau sens lexical.

Transparents 6-7

Pour le montrer, il est nécessaire de prendre position sur un certain nombre de points, à savoir :

- la définition du phénomène linguistique ciblé
- la définition de ce qu'on entend par "nouveau sens" et "nouveau sens"
- l'identification et le choix d'éléments d'information sur lesquels s'appuyer pour caractériser l'émergence d'un nouveau sens
- le cadre théorique et le modèle de représentation du sens
- l'élaboration d'un protocole en accord avec le cadre théorique et adapté à une approche applicative

Chacun de ces points sera abordé successivement.

Transparent 8

Le phénomène linguistique ciblé est celui de la néologie sémantique. Celle-ci se caractérise par l'apparition d'un nouveau signifié pour un signifiant préexistant.

En linguistique, la néologie sémantique est définie comme un processus, c'est-à-dire qu'elle s'inscrit dans le temps. Selon la perspective et les objectifs d'étude, elle peut être délimitée de façon plus ou moins large et le stade de diffusion qui lui est associé est variable : il peut aller de l'hapax, où la nouveauté est marquée mais la diffusion peu significative, à un grand nombre

d'emplois, où la diffusion du nouveau sens devient notable mais où le caractère nouveau s'estompe. La néologie sémantique est également associée à un sentiment de rupture, elle est perçue comme un écart entre un sens existant et de nouveaux sens en discours.

Transparent 9-10

Dans notre approche, l'objectif est à la fois de décrire un phénomène significatif, qui se traduit par un certain impact quantitativement, et d'être réactif face à l'émergence d'un nouveau sens. De ce fait, nous ne cibons ni un stade de diffusion précoce limité aux premiers emplois, où la diffusion n'est pas significative quantitativement, ni un stade trop avancé, où le nouveau sens s'est lexicalisé et où il est parvenu à une stabilisation. Le stade ciblé est un état transitoire, témoin d'une diffusion en cours d'implantation.

Pour pouvoir mesurer l'écart entre sens existant et nouveau sens, un traitement automatique doit disposer d'un support sur lequel s'appuyer, à savoir des données linguistiques. Celles-ci doivent fournir la matière pour représenter la signification avant l'évolution sémantique, d'autre part pour appréhender le sens associé aux nouveaux emplois. De ce fait, la nouveauté se caractérisera par un écart entre un sens codé dans une ressource de référence, en l'occurrence un dictionnaire de langue, et des emplois dans des ressources discursives, en l'occurrence des corpus textuels.

Transparent 11

La néologie sémantique est abordée ici sous un angle particulier. Notre objet d'étude est donc moins la néologie sémantique que la *néosémie*, que (Rastier et Valette, 2009) définissent comme une « innovation sémantique dont témoignent les emplois ». Le phénomène étudié est donc la néosémie, c'est-à-dire la néologie sémantique telle qu'elle apparaît à travers les traces discursives, ou encore à travers les indices délivrés par le contexte linguistique.

Transparent 12

Les indices peuvent se répartir en deux sous-ensembles :

- ceux qui contribuent principalement à la détection du nouveau sens, tels que les guillemets, les expressions introductives ou consécutives ou les empreintes de fréquence ;
- ceux qui contribuent principalement à la qualification du nouveau sens. Parmi ces indices, les cooccurrents jouent un rôle central.

Nous privilégions la qualification du nouveau sens, enjeu généralement éclipsé par celui de la détection de la néologie sémantique. L'indice que nous choisissons d'exploiter sont les cooccurrents de l'unité lexicale ciblée.

L'intérêt des cooccurrents est triple. Il se situe sur le plan qualitatif, car chaque cooccurrent est porteur d'information sémantique ; sur le plan quantitatif, car ils sont présents en grand nombre et se prêtent à des mesures, notamment statistiques ; enfin, de par leur présence massive et leur diversité, ils sont difficiles à appréhender immédiatement et la mise en place d'une approche outillée est particulièrement utile pour accéder à l'information qu'ils apportent.

Transparent 13

Pour représenter le sens et analyser les phénomènes en jeu, je me situe dans un cadre théorique inspiré de la sémantique textuelle (Rastier *et al.*, 1994 ; Rastier et Valette, 2009). La signification d'une unité lexicale est représentée sous forme d'un ensemble structuré de sèmes ou traits sémantiques, qui s'organise sous forme de molécule sémique.

Par exemple, la signification de *tsunami* pourrait comporter des sèmes associés aux éléments naturels, tels que /mer/, /vague/, /côte/, des sèmes évoquant un caractère négatif, comme /danger/ ou /catastrophe/, ou encore des sèmes liés à l'ampleur du phénomène (/démésure/ par exemple). Ces sèmes peuvent avoir un degré de généralité variable. La représentation du sens peut notamment comporter des sèmes domaniaux, tels que //Sciences de la terre// dans le cas de *tsunami*.

Transparent 14

La représentation du sens n'est pas figée, elle est susceptible d'être modulée en contexte à travers trois mécanismes d'actualisation : l'activation, l'inhibition et l'enrichissement.

Les mécanismes d'actualisation reposent sur des phénomènes de récurrence de sèmes dans le cotexte (contexte linguistique) de la cible lexicale.

Transparent 15

Chacun des mécanismes sera illustré par l'exemple d'un emploi en discours de *tsunami*. Dans cet exemple, le cotexte (ou contexte linguistique) est un paragraphe extrait d'un article de presse sur la crise financière de fin 2008. Le sens de *tsunami* n'est pas celui de phénomène naturel. Le *tsunami* désigne ici la crise financière, ou plus précisément la finance comme source de bouleversements de grande ampleur du système économique.

Transparent 16-17

Le premier mécanisme d'actualisation est l'activation de sèmes. Si au voisinage de la cible lexicale, il y a récurrence d'un sème déjà associé à la cible lexicale, ce sème sera activé. Par exemple, des sèmes tels que /danger/ ou /catastrophe/ préalablement associés à la signification de *tsunami* sont présents de façon notable dans le voisinage de *tsunami*. Ils s'expriment par exemple à travers des unités lexicales telles que *ruine*, *menacer* ou encore *étrangler*. La présence massive des sèmes les fait ressortir au niveau du sens de *tsunami*, autrement dit, il y a activation de ces sèmes.

Transparent 18

A l'inverse, une absence remarquable de certains sèmes préalablement associés à la signification de la cible favorisera leur inhibition, c'est-à-dire qu'ils ne s'exprimeront pas dans le contexte d'emploi considéré. Ainsi, les sèmes relatifs aux phénomènes et éléments naturels, qu'il s'agisse du sème domanial //Sciences de la terre// ou des sèmes tels que /mer/ ou /côte/, ne sont pas repris par le contexte linguistique et sont ainsi inhibés.

Transparent 19-20

Le troisième et dernier mécanisme d'actualisation est l'enrichissement. Des sèmes absents de la signification de la cible lexicale et massivement présents dans le contexte linguistique auront tendance à s'intégrer au sens de la cible, autrement dit, ils pourront participer à l'enrichissement de son sens. C'est par exemple le cas du sème domanial //Finance//, massivement récurrent (par exemple, à travers *capitalisme*, *économie*, *zone euro*, *dollar*, *banque*, *crédit*).

Pour les mécanismes évoqués (activation, inhibition, enrichissement), un cotexte d'emploi peut donner lieu à une interprétation qui s'écarte de la signification, mais ce seul cotexte ne suffit pas à amender durablement la signification de l'unité observée. En revanche, la présence répétée dans le temps et sur un grand nombre d'emplois pourra témoigner de l'implantation d'un nouveau sens et pourra amener à intégrer le changement sémantique à la signification.

Le modèle de représentation du sens qui vient d'être décrit est celui qui sous-tend le protocole d'allocation de signifié dont nous dessinons les grandes lignes.

Transparent 21

La structure de base du protocole repose sur une représentation du sens en langue et en discours. Le paramètre temps est fondamental, pour contraster les nouveaux emplois avec les emplois antérieurs. De même, il est nécessaire de mettre en évidence les interactions entre langue et discours.

Transparent 22

En pratique, la représentation du sens en langue a été obtenue à partir d'un dictionnaire de langue, le *Trésor de la Langue Française informatisé (TLFi)*. L'exploitation du TLFi sera détaillée dans les transparents qui suivent.

Le sens en discours a été étudié à partir de trois corpus de presse datant des années 2000. Chaque corpus a permis d'étudier un aspect particulier de l'évolution de sens :

- Le corpus 'Factiva' a servi à étudier un ensemble de cibles prototypiques. Il a été utilisé pour observer l'influence des domaines comme indices et descripteurs de l'évolution de sens.
- Le corpus 'Crise financière' a été utilisé pour l'étude d'une cible particulière, *toxique* en contexte de crise financière. Il a été exploité afin de mettre en relation le sens en discours avec le sens codé dans le dictionnaire.
- Le corpus 'Outreau' a servi à étudier l'évolution de sens d'*Outreau*. Ce corpus, fourni par sa conceptrice Michelle Lecolle, a fait l'évolution d'une étude antérieure de (Lecolle, 2007) sur l'évolution progressive du sens de Outreau-ville à *Outreau* comme fiasco judiciaire. Le corpus, segmenté en périodes de temps, a principalement servi à étudier la progression de la diffusion d'un nouveau sens au cours du temps.

La méthode utilisée pour observer les variations de sens en corpus repose sur le calcul d'un indice statistique à partir d'une représentation en sac de mots. Nous y reviendrons par la suite, après avoir précisé le rôle du dictionnaire dans la représentation du sens.

Transparent 23

Le dictionnaire a permis d'affecter un contenu sémantique à toute unité lexicale, en particulier à la cible lexicale (par exemple, *toxique*, *Outreau*, *tsunami*). Il a également servi à enrichir les corpus textuels en traits sémantiques, pour que le formalisme puisse traduire les mécanismes d'actualisation évoqués précédemment (activation de traits sémantiques, inhibition de traits sémantiques et enrichissement en traits sémantiques).

L'appariement du corpus et du dictionnaire et l'affectation de signifié a été réalisée par la plateforme Semy, une ressource développée en interne au sein du laboratoire. Après lemmatisation et élimination des mots-outils, les unités lexicales extraites du corpus sont appariées aux entrées du dictionnaire correspondantes. De ces définitions sont extraits les noms, verbes, adjectifs et adverbes, qui sont considérés comme les traits sémantiques de l'unité lexicale considérée. L'ensemble de ces traits sémantiques constitue le sens codé associé à la cible. Ce dernier n'a pas de structure interne, il se présente sous forme de sac de traits. Les traits sémantiques sont ensuite réinjectés dans le corpus par annotation : toute unité lexicale se voit associer l'ensemble des traits sémantiques issus des définitions lexicographiques.

Transparent 24

Après annotation, trois niveaux de description du sens sont disponibles, le niveau supra-lexical, le niveau lexical et le niveau infra-lexical. Ces niveaux permettent d'aller au-delà du niveau lexical et ils offrent la possibilité de procéder par niveaux de granularité sémantique de plus en plus fins.

La description supra-lexicale repose sur les étiquettes de domaine. Celles-ci sont certes extraites des définitions du dictionnaire et pourraient de ce fait être considérées comme infra-lexicales. Cependant, elles sont potentiellement vecteurs d'une information caractéristique à des paliers de textualité larges, susceptibles d'interagir avec des informations caractéristiques d'un palier de textualité global et d'être amendées par ces informations. De ce fait, nous les assimilons à des descripteurs supra-lexicaux.

La description lexicale repose sur les unités telles qu'elles apparaissent dans les discours.

La description infra-lexicale repose sur les traits sémantiques extraits des définitions autres que les domaines.

A partir de ce panel de descripteurs, l'enjeu est de définir comment sélectionner des unités pertinentes pour qualifier le nouveau sens. La sélection des unités repose sur un critère statistique en accord avec le principe de récurrence et qui associe significativité statistique et significativité sémantique.

Transparent 25

La sélection d'unités pertinentes repose sur l'hypothèse de distributionnalité de Harris : des mots qui partagent les mêmes contextes sont de sens proche. Autrement dit, l'affinité distributionnelle de eux mots reflète une affinité sémantique. Cette affinité peut être calculée entre toute paire d'unités (domaines, unités lexicales, traits sémantiques).

Le calcul de l'affinité entre deux unités repose sur une analogie des phénomènes en corpus avec le schéma d'urne. La représentation du corpus sous forme de sac de mots correspond à l'urne contenant des boules vertes et blanches. Le sous-corpus du voisinage de la cible lexicale correspond à l'échantillon tiré de l'urne. L'unité (domaine, trait sémantique ou unité lexicale) dont on cherche à calculer l'affinité avec la cible lexicale (par exemple, /danger/ au voisinage de *tsunami*) est associée aux boules vertes. L'affinité sémantique est obtenue à partir de la probabilité d'avoir le nombre observé de boules vertes dans l'échantillon tiré (probabilité d'avoir k fois l'unité observée au voisinage de la cible). Cette probabilité est interprétée comme un degré de surreprésentation ou de sous-représentation au voisinage de la cible, lui-même interprété comme une affinité sémantique forte ou faible de l'unité considérée avec la cible.

En pratique, l'indice retenu pour quantifier l'affinité sémantique est l'indice de spécificité de (Lafon, 1984), issu de la loi hypergéométrique, tel qu'implémenté dans le logiciel Lexico 3 (Salem *et al.*, 2003).

Transparent 26

Le calcul des spécificités dépend d'une part du choix d'observables (les domaines, les unités lexicales et les traits sémantiques dans notre cadre), ainsi que sur la définition d'un corpus et d'un sous-corpus. Dans notre cadre, il est possible d'effectuer de multiples découpages de corpus et sous-corpus, en fonction de différents paramètres. Le jeu sur de multiples découpages jouera un rôle fondamental dans notre démarche.

Transparent 27-29

La définition du sous-corpus repose sur trois paramètres :

- le voisinage de la cible lexicale défini de façon plus ou moins larges (dans notre cas, palier textuel : texte ou paragraphe), qui offre la possibilité d'effectuer des comparaisons entre différents voisinages ou entre un voisinage et le reste du corpus ;
- le temps, avec un découpage en un nombre variable de périodes, qui permet de confronter de périodes de temps ou plusieurs périodes de temps successives ;
- les domaines, avec la possibilité d'effectuer un découpage selon de grands ensembles textuels sémantiquement cohérent, puis de contraster différents domaines entre eux ou d'effectuer des analyses au sein d'un domaine donnée.

Transparent 30

La procédure proposée repose sur des jeux de contrastes organisés en fonction des observables, des ressources et des paramètres.

La clé d'organisation principale de la procédure est la granularité sémantique : on procède par granularité sémantique décroissante, avec des affinements successifs de l'information (d'abord sont observés les domaines, ensuite les unités lexicales et enfin les traits sémantiques).

La seconde clé d'organisation de l'information est réalisée en fonction du lien des unités avec le sens codé. On distingue d'une part les unités associées au sens codé (propre aux sens préexistant au changement), susceptibles de participer à sa reconfiguration (activation et inhibition des traits

sémantiques) ; d'autre part les unités qui n'appartiennent pas au sens codé (unités proprement nouvelles), susceptibles de participer à l'enrichissement du sens codé.

La dernière clé d'organisation est l'ensemble des paramètres textuels. Ils sont agencés pour évoluer vers une information de plus en plus précise, notamment avec la confrontation dans un premier temps de deux périodes, pour l'observation de l'évolution de sens sur plusieurs périodes successives. Ils sont également combinés pour prendre en compte différents types d'information, par exemple, pour prendre en compte à la fois ce qui distingue le voisinage de la cible au sein du domaine considéré et ce qui distingue le voisinage de la cible dans un domaine par rapport à d'autres domaines d'emploi.

Les différentes étapes ont été étudiées à travers une série d'expérience. Deux niveaux d'information qui font la spécificité de notre approche sont les domaines et les traits sémantiques. Nous présentons brièvement les expériences réalisées à chacun de ces niveaux pour observer l'évolution de sens.

Transparent 31

La première étape de l'étude de l'évolution de sens se fait à partir d'un niveau de granularité sémantique grossier, à travers l'étude des domaines d'emploi. Ceux-ci constituent des descripteurs génériques propres à des paliers de textualité larges.

L'analyse des domaines a d'abord été menée en corpus, par confrontation de deux puis plusieurs périodes de temps. Elle a été réalisée sur un panel de cibles lexicales prototypiques, avec des profils variés : ce panel comporte des cibles à caractère néologique (*toxique* en finances, *tsunami*, *tablette* dans le champ des nouvelles technologies, *moléculaire* en gastronomie), une cible à statut intermédiaire (*numérique*, *tempête*) et des cibles témoins non néologiques (*dangereux*, *délétère*, *raz-de-marée*). Ces cibles ont été étudiées à partir du corpus 'Factiva', un corpus multidomaines.

Transparent 32

L'analyse des domaines a servi dans un premier temps à effectuer un tri des cibles. Ce tri a été réalisé par confrontation de deux périodes de temps, sur critère d'accroissement minimal de la présence de la cible dans au moins un domaine. Le tri généré est dans l'ensemble en accord avec les connaissances préalables sur le caractère néologique des cibles : les cibles néologiques et à statut intermédiaire sont conservées, à l'exception de *moléculaire*, et les cibles non néologiques *dangereux*, *délétère* et *raz-de-marée* sont éliminées.

Dans un second temps, l'évolution de la présence de la cible a permis une première qualification du nouveau sens : les domaines dans lesquels était vérifié le critère d'accroissement minimal de présence la cible ont été considérés comme candidats pour qualifier le nouveau sens.

Transparent 33

Enfin, l'évolution des domaines a été observée sur plusieurs périodes de temps. Ceci a permis de faire émerger différents types d'évolution, avec des pics de diffusion dans un domaine (*toxique* en économie) ou plusieurs domaines (*tsunami*), ou des diffusions plus progressives (croissance progressive de *numérique* simultanément dans plusieurs domaines).

Transparent 34

La première sous-étape de l'étude des domaines a été effectuée exclusivement à partir des ressources discursives et elle a permis d'identifier des domaines où pouvait s'implanter le nouveau sens. La deuxième sous-étape a pour objectif de mettre en relation le sens en discours avec le sens codé du dictionnaire au sein des domaines identifiés. Les expériences illustratives de cette sous-étape ont été réalisées à partir du corpus 'Crise financière', associé au domaine économique. L'étude a porté sur la cible *toxique*.

Dans le sens codé extrait du TLFi, cinq domaines sont affectés à *toxique* : BIOLOGIE, CHIMIE, MEDECINE, PHYSIOLOGIE et PHARMACOLOGIE. D'abord, le devenir de ces domaines dans le corpus est étudié, autrement dit, on cherche à observer leur éventuelle reconfiguration. Les analyses de spécificités ont mis en évidence une inhibition massive des domaines associés au sens codé, reflet d'un sens textuel différent du sens codé. Ensuite, nous avons recherché des domaines saillants, susceptibles de participer à l'enrichissement du sens. Dans l'étude de *toxique*, le domaine FINANCE a ainsi été identifié comme candidat à l'enrichissement du nouveau sens.

Transparent 35

Après les domaines, le deuxième niveau d'information propre à notre approche est celui des traits sémantiques. Ceux-ci ont été étudiés afin de préciser l'information apportée par les domaines.

Dans un premier temps, ils ont été étudiés en termes de reconfiguration du sens codé, afin d'identifier les traits activés et les traits inhibés dans les nouveaux contextes d'emploi. Les expériences ont été réalisées sur *toxique* en contexte de crise financière, ce qui a mis en évidence une inhibition massive des traits du sens codé, puis sur *Outreau* dans le corpus 'Outreau' pour observer la modulation au cours du temps des traits du sens codé. L'évolution des traits sémantiques d'Outreau a été globalement conforme à l'analyse manuelle de (Lecolle, 2007).

Transparent 36

Dans un second temps, les traits sémantiques ont été étudiés dans une perspective d'enrichissement du sens codé. Nous avons cherché à faire émerger des traits susceptibles de préciser le domaine identifié comme nouveau. Dans le cas de *toxique*, les expériences ont permis de faire émerger des traits candidats à l'enrichissement tels que /bilan/, /prêt/ ou /banque/ sur critère de significativité statistique et de dépendance au nouveau domaine.

Transparent 37

En conclusion, nous avons répondu point par point à chacune des questions soulevées par l'idée défendue, à travers des propositions théoriques et méthodologiques illustrées par des expériences illustratives pour chaque grande étape.

Transparent 38-40

Le phénomène observé est la néosémie, définie comme de la néologie sémantique telle qu'elle apparaît au prisme des indices discursifs.

La notion de nouveauté a été définie comme un écart à la fois dans le temps et entre sens codé et emplois en discours.

Les indices exploités pour qualifier le nouveau sens sont les cooccurrents. Ceux-ci délivrent des informations globales et locales et permettent d'intégrer des paliers de textualité large.

Le modèle de représentation du sens est inspiré de la sémantique textuelle. Il fait intervenir différents niveaux de granularité sémantique et permet d'aller au-delà du niveau lexical pour exploiter des niveaux d'information infra-lexical et supra-lexical. Les mécanismes sémantiques qui sous-tendent la néosémie reposent sur de la récurrence de traits sémantiques, qui donnent lieu soit à une reconfiguration du sens par inhibition et activation de traits sémantiques, soit à son enrichissement en nouveaux traits.

Nous avons défini les grandes lignes d'un protocole à partir d'une approche mixte de statistique et de sémantique. Ce protocole repose sur des affinements successifs de l'information sémantique et sur des interactions entre différentes ressources et dans le temps.

Transparent 41

Au niveau des perspectives, les lignes de force proposées pourraient être appliquées à d'autres types de néologie et le protocole pourrait être consolidé à partir d'une exploitation plus poussée de la structure du dictionnaire.

Transparent 42

Les autres types de néologie auxquels appliquer le protocole sont notamment les néologismes morphologiques et les nouveaux phrasèmes.

Pour les néologismes de forme par procédé morphologique, il pourrait y avoir un prétraiement, avec l'affectation préalable d'un signifié à partir d'une décomposition morphologique à l'aide d'outils tels que DériF (Namer, 2009). Un sens codé pourrait alors être affecté à partir de cette décomposition et de la définition du dictionnaire associée à la base morphologique. Puis ce sens codé pourrait être testé en corpus selon notre protocole, afin de tester la validité du sens affecté. Ainsi, pour *hameçonnage*, la procédure permettrait de tester en corpus les traits sémantiques récupérés dans les définitions de *hameçonner* et elle serait susceptible de montrer une inhibition massive des traits relatifs à la pêche, avec un enrichissement en traits sémantiques propres au domaine de l'informatique

De même, pour les nouveaux phrasèmes tels que *réalité augmentée*, un sens codé pourrait être affecté dans un premier temps par réunion des sens codés associés à chaque constituant du phrasème (traits sémantique de *réalité* et traits sémantiques d'*augmenté*), sous réserve de validation du sens codé ainsi généré selon les emplois en corpus. La procédure que nous proposons permettrait précisément cette validation, et l'amendement du sens codé obtenu par réunion, avec élimination des traits non pertinents et enrichissement en traits sémantiques absents des traits propres à chaque composante.

Transparent 43

Pour améliorer la procédure en renforçant le principe d'affinements successifs et l'adéquation avec le cadre théorique, notamment la notion de molécule sémique, il conviendrait d'exploiter la macrostructure et la microstructure du dictionnaire.

Au niveau de la macrostructure, exploiter la hiérarchie des domaines, aussi bien au niveau des subdivisions successives que des dépendances entre domaines et sous-domaines.

Au niveau de la microstructure, un développement serait de prendre en compte l'organisation interne des entrées, notamment le découpage en définitions et les dépendances de traits sémantiques définitoires à des domaines d'emplois.