

## PROPOSITIONS POUR UNE LEXICOLOGIE TEXTUELLE

Par MATHIEU VALETTE<sup>1</sup>

Notre propos est d'apporter quelques éléments de réflexion pour l'élaboration d'un programme de recherche en *lexicologie textuelle*. La première moitié de l'article est consacrée à la défense de l'hypothèse selon laquelle il existe une relation d'équivalence structurelle à déterminer entre le contenu sémantique d'une lexie et certaines formes sémantiques : signifié et forme sémantique seraient tous deux identifiables à un groupement de sèmes de compacité variable, associé à un signifiant stable et synthétique dans le cas du signifié, discontinu et sans lexicalisation privilégiée dans le cas de la forme sémantique. De cette équivalence, nous inférons que certaines formes sémantiques seraient potentiellement des signifiés en devenir, ou les signifiés de *proto-signes* sans signifiant stabilisé ni synthétique attiré. En bref, la lexie serait un cas particulier de forme sémantique et, corollairement, un fragment textuel. En adossant la description du lexique à une problématique des structures sémiques, nous adoptons l'outillage théorique d'une sémantique textuelle pour la lexicologie. La seconde moitié de l'article a pour objectif d'illustrer notre propos. Nous y relatons quelques récentes études sur corpus sémiques dont les conclusions sont de nature à étayer nos hypothèses.

### 1. INTRODUCTION

La sémantique interprétative (Rastier 1987 / 2001) propose d'étudier la structuration sémantique d'un texte par le biais de réseaux de traits sémantiques, ou *sèmes*, qui en assurent la cohésion. La théorie en détaille deux catégories générales : (i) l'isotopie, qui relève du *fond sémantique* (récurrence d'un même sème sur un empan de longueur variable, de la phrase au corpus) ; (ii) la molécule sémique, qui correspond à la *forme sémantique* (stabilisation et récurrence de sèmes hétérogènes). À l'interface entre le lexique et le texte, les réseaux sémiques permettent d'étudier de manière approfondie à la fois le lexique, le texte et leur relation cohésive. Un programme de recherche en lexicologie textuelle pourrait, par conséquent, s'articuler autour de deux objectifs conjoints : approfondir les connaissances actuelles sur ces objets sémantiques connus et notamment référencés par la théorie, en identifier de nouveaux qu'elle n'a pas su jusque-là reconnaître faute d'une instrumentation adéquate.

#### *Une approche textuelle du lexique*

Le mot est un concept linguistique fragile. À la fois imprécis de par ses frontières théoriques et matérielles, et ethnocentrique parce que les langues sans mot sapent

1 ATILF (CNRS, Nancy) ; mvalette@atilf.fr.

tout espoir d'en faire un concept universel, il demeure néanmoins un mode d'aperception du langage parmi les plus intuitifs et est peut-être le plus étudié en linguistique, en dépit du succès de la phrase et de l'énoncé depuis plus d'un demi-siècle. Une approche textuelle du mot pourrait, comme le suggère Rastier (2001 : 182 s.) à propos d'une reconception possible du signe, s'inspirer d'un texte où Saussure écrit :

[...] vous n'avez plus le droit de diviser, et d'admettre d'un côté le mot, de l'autre sa signification. Cela fait tout un. Vous pouvez seulement constater le kénôme  $\cap$  et le sème associatif  $\mathcal{C}$ . (Saussure 2002 : 93)

Rastier normalise le kénôme suivant le symbole mathématique de l'intersection «  $\cap$  » et l'interprète à l'aune de la représentation saussurienne du signe (Saussure 1972 : 158) comme un « signifié ouvert vers des signifiants indéterminés ». Quant au sème associatif (sème signifiant < signe > pour Saussure), Rastier le restitue avec deux symboles d'inclusion accolés «  $\supset\subset$  », neutralisant de la sorte la petite intersection visible dans le texte saussurien. Le sème associatif est, selon Rastier, « le signe linguistique contextuellement défini » et lui permet de développer son concept de *passage* (Rastier 2007). Le signe (ou le mot) est vu comme un passage vide entre deux contextes, gauche et droit.

Même si nous ne verrions aucun inconvénient à conserver le modeste espace intersectif de la représentation originale du sème associatif saussurien, la radicalité de cette conception du signe nous agrée et nous l'étendons à celle du mot. La lexicologie a recours en maints lieux théoriques et pratiques au contexte pour définir le contenu sémantique du mot (par la collocation notamment)<sup>2</sup>, mais ce contexte est perçu d'un point de vue syntaxique ou micro-syntaxique. Il importe selon nous de l'aborder comme un objet et donc, dans la perspective de son objectivation. Dans ce cadre général, nous présentons, dans un premier paragraphe, un ensemble de propositions adossé à la sémantique interprétative visant à situer l'étude du lexique dans le paradigme textuel. Plus précisément, notre projet est d'étudier les déterminations textuelles de la création et de la lexicalisation des concepts. Nous l'illustrerons dans un second paragraphe par une méta-étude rendant compte de différents travaux menés dans cette perspective.

## 2. UNIFIER LA DESCRIPTION DES MOTS ET DES TEXTES

Adoptons un empirisme de méthode : les mots apparaissent dans deux types d'objet matériel ; le texte, objet construit de façon syntagmatique, où ils sont actualisés dans un état qu'on pourrait qualifier de dynamique, et le dictionnaire, objet construit de façon paradigmatique, où ils sont dans un état passif, en attente d'une actualisation. Qu'est-ce qu'un mot, par-delà ces types d'objet ? Il est un signe con-

2 Cf. Blumenthal / Hausmann (éds., 2006).

stitué d'une forme et d'un contenu. La forme est acoustique ou graphique, on l'appelle le « signifiant », le contenu est, dans le paradigme structuraliste une collection de propriétés sémantiques plus ou moins articulées entre elles, qui constituent le « signifié ». Certaines traditions distinguent toutefois le signifié à proprement parler de son contenu, suivant l'opposition langue / parole. Ainsi, sur le modèle du lexème et de la lexie, on distingue le *sémème* en langue et la *sémie* en discours. Cette distinction, en première approximation, outrepassé notre ambition descriptive, même si l'on pourrait assimiler la relation du dictionnaire au texte à l'opposition langue / discours. Pour des raisons de clarté, mais aussi parce que la pratique du TAL dans laquelle nous inscrirons la seconde partie de l'article, rend peu pertinente ladite opposition, nous retiendrons le seul signifié<sup>3</sup>. Les propriétés sémantiques qu'il contient sont des sèmes. Elles sont d'ordre métalinguistique et résultent d'une analyse ou d'une validation humaine effectuées, par exemple, par un linguiste. Cohérentes, les collections de sèmes relèvent de la catégorie générale des objectivations sémantiques. On peut en distinguer deux types, selon que l'on se trouve dans une problématique du texte ou dans une problématique du dictionnaire.

### 2.1. Les signifiés

Composé d'un signifié et d'un signifiant, nous représenterons un mot de la façon suivante :

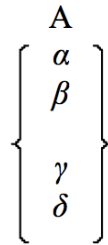


Figure 1 : un signifié composé des sèmes  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$  associé à un signifiant A

A désigne le signifiant (c'est-à-dire, en pratique, le mot graphique) et  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$  entre accolades désignent le signifié proprement dit (composé des sèmes regroupés). Tous ces traits n'ont pas nécessairement la même qualité ni la même valeur sémantique, comme nous le verrons dans le paragraphe 3, ce que figure le saut de

3 L'opposition taxème vs taxémie, suggérée par Rastier (2001, 154 s.) et discutée par Duteil (2004) pose un problème supplémentaire dans la mesure où ce type de classe sémantique ne peut être, sauf erreur de notre part, qu'établi dans les textes, c'est-à-dire en discours. En fait de taxèmes, il n'y a que des taxémies. La notion de < taxie > proposée par Missire (2006, 73) pose le même problème.

ligne entre  $\beta$  et  $\gamma$ . Certains sèmes, par exemple, sont génériques d'une classe donnée. Ainsi, le signifié du mot *chien* peut être décrit comme la collection de sèmes énumérés dans la partie gauche de la figure 2. Mais dans la mesure où les sèmes sont purement métalinguistiques, la liste n'est ni exhaustive, ni fermée, et ne relève pas d'une quelconque valeur de vérité. Dès lors, le contenu sémantique de *chien* construit par un informateur de quatre ans pourrait fort bien correspondre au signifié de droite, toujours sur la figure 2.

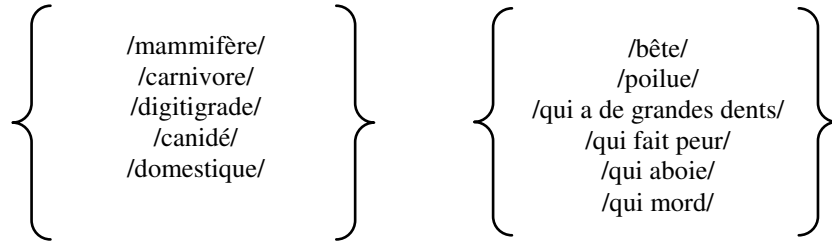


Figure 2 : deux signifiés possibles pour *chien*

Aucun de ces signifiés, qui sous-tendraient deux définitions lexicographiques parfaitement distinctes, n'est meilleur qu'un autre. Les propriétés sémantiques diffèrent par leur seul contexte de construction. Le premier est savant, le second se fonde sur une ou plusieurs expérience(s) sensible(s) ou fantasmée(s)<sup>4</sup>.

Un texte est, formellement parlant, un alignement de signes (de mots) suivant des règles de construction syntaxiques. Ignorant provisoirement la syntaxe, nous le représenterons de la façon schématique suivante :

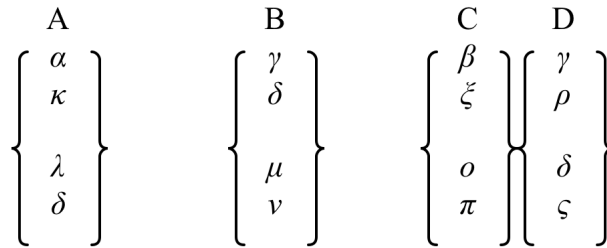


Figure 3 : Un texte (alignement de mots constitués chacun d'un signifiant (*A, B, C, D*) et d'un signifié)

4 Dans une classe des //animaux terrifiants//, il est possible que *loup* côtoie le *chien* à une courte encablure sémique : {/bête/, /poilue/, /qui a de grandes dents/, /qui hurle/, /qui dévore/}.

## 2.2. Les réseaux sémiques

La mise en relation dans un texte de plusieurs signifiés donne lieu à de nouveaux regroupements de sèmes, syntagmatiques cette fois-ci, c'est-à-dire des groupements entre sèmes appartenant à des signifiés différents. Ces groupements syntagmatiques sont beaucoup plus ponctuels, puisque spécifiques à un texte, ou à un ensemble de textes. L'interprétation d'un texte repose sur la reconnaissance et l'identification de ces groupements syntagmatiques. On distingue deux types de groupements syntagmatiques.

### 2.2.1. Les fonds sémantiques (isotopies)

Un sème donné peut se retrouver à plusieurs endroits, dans un même texte. On appelle cette récurrence une isotopie. Les isotopies constituent ce que l'on appelle le fond sémantique. Dans la figure 4 ci-dessous, nous avons un exemple d'isotopie due à la récurrence d'un trait  $\delta$ .

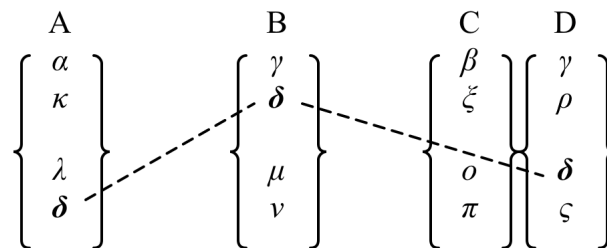


Figure 4 : Une isotopie (fond sémantique) – récurrence du sème  $\delta$  dans le texte

Dans le texte qui suit, on observe une isotopie simple, par la récurrence du sème /ville/ :

Violence urbaine<sup>/ville/</sup> : l'expression est née récemment, il y a plus ou moins dix ans. Une expression qui s'applique plus exactement à certaines villes<sup>/ville/</sup> de banlieue<sup>/ville/</sup>, celles qui rassemblent un grand nombre d'habitants d'origine étrangère, au point d'être devenues de véritables ghettos<sup>/ville/</sup>.<sup>5</sup>

Minimalement, cela signifie que ce texte traite de la ville. Mais les isotopies peuvent relever d'une description plus fine. Par exemple, elles peuvent être génériques, c'est-à-dire correspondre aux sèmes structurants de classes sémantiques telles que les domaines et les taxèmes<sup>6</sup>. Dans le fragment ci-dessous, s'actualise

5 <http://www.lacathode.org/cqfs/viol.htm>.

6 Un taxème est une petite classe sémantique correspondant à une situation d'usage précise. La cohésion de la classe est assurée par les sèmes génériques. Un domaine est composé d'un ensemble de taxèmes correspondant à une pratique déterminée.

une *isotopie taxémique* à partir de la classe sémantique des //animaux de compagnie// :

Des sociétés ou associations mettent à votre disposition une équipe d'assistants chargés de venir faire une ou plusieurs visites régulières à votre domicile, pour nourrir votre chien<sup>/dom/</sup>, chat<sup>/dom/</sup>, canari<sup>/dom/</sup>, ou hamster<sup>/dom/</sup>, lui prodiguer des soins adaptés (sorties, jeux, câlins) et lui apporter une présence rassurante.<sup>7</sup>

L'isotopie /domestique/ est actualisée dans *domicile, chien, chat, canari, hamster*. Cette isotopie taxémique est générique parce que le sème récurrent est un de ceux établissant la cohésion du taxème (il s'agit d'ailleurs d'un faisceau d'isotopies génériques car l'isotopie /domestique/ est accompagnée d'une isotopie /animal/ – autre sème générique de la classe des //animaux de compagnie//).

En bref, l'isotopie constitue le socle du parcours interprétatif. Elle permet d'une part, de zoner les textes et d'autre part, de les inscrire dans l'intertexte. Le *zonage isotopique* est la manifestation spatiale d'un faisceau d'isotopies. Le zonage participe à l'identification de passages, lesquels constituent l'unité textuelle d'accueil des formes sémantiques. Quant aux *isotopies intertextuelles*, elles organisent *a minima* la cohésion et l'homogénéité d'un corpus (par exemple, l'ensemble des textes où il est question d'animaux domestiques).

### 2.2.2. Les formes sémantiques

Plusieurs sèmes distincts peuvent être instanciés ensemble dans des textes différents avec une certaine régularité. Ces groupements s'appellent des thèmes, ou des *formes sémantiques*. Nous les représenterons ainsi :

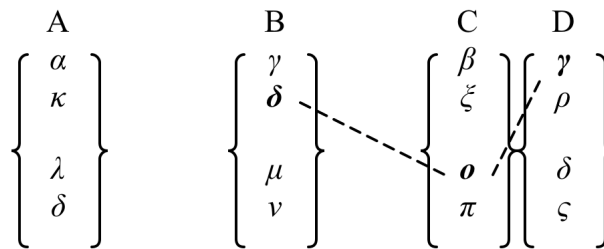


Figure 5 : une forme sémantique (groupement stabilisé des sèmes  $\delta$ ,  $o$  et  $\gamma$  dans différents textes)

Par exemple dans les textes suivants :

7 <http://www.linternaute.com/acheter/depart-vacances/animaux/animaux.shtml>.

Nicolas Sarkozy et les émeutiers<sup>/émeute/</sup> de banlieues<sup>/ville/</sup> sont le produit d'un même terreau.<sup>8</sup>

Prenez un patron de CRS [...] qui gère une échauffourée<sup>/émeute/</sup> dans un quartier<sup>/banlieue/</sup>, il sait qu'il doit laisser les jeunes gredins s'escrimer sur les vitrines.<sup>9</sup>

Les sèmes /émeute/ et /ville/ sont en cooccurrence rapprochée. Ils constituent une petite forme sémantique, c'est-à-dire une unité sémantique dont la lexicalisation n'est ni stable ni synthétique, autrement dit, dont le signifiant est discontinu et variable. Cette unité sémantique correspond, par exemple, à l'unité lexicale « émeute urbaine » :

Les émeutes urbaines<sup>{/émeute/ville/}</sup> sont récurrentes en France depuis le début des années 80. Les premières émeutes<sup>/émeute/</sup> ont lieu en 1979, à Vaulx-en-Velin<sup>/ville/</sup>.<sup>10</sup>

Les formes sémantiques ne sont pas assimilables à de simples périphrases, elles importent en premier lieu pour les modalités qu'elle opèrent sur les signifiés des unités lexicales cooccurrentes. Prenons un exemple de forme sémantique de plus en plus souvent lexicalisée en « *liberté de fumer* » :

Le citoyen est libre<sup>/liberté/</sup> de fumer<sup>/fumer/</sup> ou de ne pas fumer, de manger de la salade si ça lui chante et des rillettes s'il en a envie.<sup>11</sup>

Ils ont la liberté de fumer<sup>{/liberté/fumer/}</sup>, soit, je serai le dernier à la leur retirer, sauf dans les endroits publics.<sup>12</sup>

Cette forme sémantique s'est particulièrement développée et tend à se stabiliser depuis la mise en application de la loi dite antitabac en janvier 2008. Elle constitue un des arguments privilégiés de ses détracteurs (ce serait une loi liberticide). Les industriels du tabac l'exploitent bien évidemment, mais ils ont soin de restreindre cette liberté de fumer aux seuls adultes, la loi leur interdisant de faire la promotion du tabac auprès des enfants :

Notre métier ne consiste pas à inciter des gens à fumer<sup>/fumer/</sup>. Il consiste à offrir des marques de qualité à des adultes qui ont déjà pris la décision<sup>/liberté/</sup> de fumer<sup>/fumer/</sup> [...]. C'est pourquoi nous sommes convaincus que fumer<sup>/fumer/</sup> devrait être le seul fait d'adultes conscients des risques de fumer<sup>/fumer/</sup>. (JTI)

Fumer<sup>/fumer/</sup> repose sur une décision<sup>/liberté/</sup> individuelle qui ne peut être que le fait d'adultes informés des risques liés au tabagisme. (ALTADIS)

Nous sommes convaincus que le choix<sup>/liberté/</sup> de fumer<sup>/fumer/</sup> doit être le choix d'adultes avertis et conscients des risques, un choix qui exclut de fait les jeunes, non adultes. (BAT)

Nous nous engageons ainsi à communiquer de manière responsable avec les adultes qui ont délibérément choisi<sup>/liberté/</sup> de fumer<sup>/fumer/</sup>. (ALTADIS)

8 <http://www.voltairenet.org/article130994.html>.

9 <http://www.forumdesforums.com/modules/news/article.php?storyid=14613>.

10 <http://www.babnet.net/cadredetail-3296.asp>.

11 <http://www.le-tigre.net/>.

12 <http://www.philo5.com/>.

JTI s'engage à fabriquer des cigarettes<sup>/fumer/</sup> de qualité pour les adultes qui choisissent<sup>/liberté/</sup> de fumer<sup>/fumer/</sup> par plaisir. (JTI)<sup>13</sup>

De la sorte, les industriels du tabac construisent une nouvelle forme sémantique {/liberté//fumer//adulte/}. Toutefois, *adultes* subit un certain nombre de modalités valorisantes, le plus souvent sous la forme d'adjectifs qualitatifs (*conscients*, *informés*, *avertis*, etc.) de sorte que se construit un parcours interprétatif < liberté de fumer + adulte<sup>/valorisant/</sup> >. Ce discours de valorisation de l'adulte libre de fumer, autrement dit, de l'adulte fumeur, induit une lecture spéculaire telle que le jeune, non adulte, est dévalorisé. C'est explicite dans le troisième exemple (« un choix qui exclut de fait les jeunes, non adultes »). On peut donc construire les deux formes sémantiques ci-dessous :

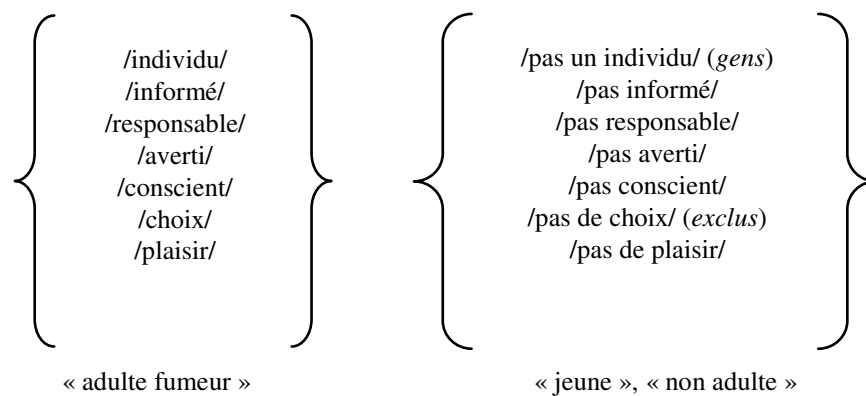


Figure 6 : formes sémantiques d'*adulte* et de *non adulte* dans les textes de l'industrie du tabac

### 2.3. Le sens, un ajustement entre signifiés et formes sémantiques

Le sens d'une unité lexicale dépend autant du contexte dans lequel il apparaît que de sa définition première. On distingue la signification (du dictionnaire) et le sens (dans le texte). En complément de l'opposition sèmes génériques vs sèmes spécifiques (qui permet d'introduire la notion de < classe sémantique >), la sémantique interprétative opère une distinction entre des sèmes inhérents et des sèmes afférents. Les sèmes inhérents peuvent être considérés comme plus « définitoires » (par exemple, pour *chien*, on aura : /canidé/, /quadrupède/, etc.) que les sèmes afférents qui relèvent de l'usage qui est fait du mot dans les textes. Les sèmes afférents sont donc issus des contextes d'énonciation. C'est le cas dans l'exemple de la figure 6 où le signifié d'*adulte* n'est représenté ici qu'avec des sèmes afférents. On retiendra qu'il s'agit de sèmes hérités d'un ensemble de contextes et suscep-

13 Ces exemples et les analyses sont issus des recherches collectives de l'équipe linguistique ERTIM-INaLCO du projet C-MANTIC (ANR-07-MDCO-002).

tibles d'être recontextualisés à l'identique. De la même manière, le sème /pauvre/ est fréquemment actualisé dans *population* lorsqu'il est cooccurrent du lexème « *banlieue* ». Mais il ne s'agit pas d'un sème inhérent : rien dans la banlieue ne la prédispose à accueillir de façon privilégiée une population pauvre. De même, l'expression *jeunes des banlieues*, dans le discours journalistique (qui est aujourd'hui largement prescriptif) ou dans le discours politique (qui lui ressemble) ne signifie pas tous les jeunes résidant en banlieue, mais certains jeunes, défavorisés, résidant en banlieue, et même, fréquemment, des jeunes gens *issus de l'immigration*. Lorsque le quotidien *Le Monde* titre le 30 septembre 2001 : « Les jeunes des banlieues craignent l'amalgame entre musulmans et terroristes », il enrichit implicitement le signifié de l'unité *jeune de banlieue* d'un sème /musulman/. Ces sèmes, /pauvre/ ou /musulman/ sont des sèmes afférents, < subjectifs > ou < socialement normés >, c'est-à-dire circonscrits d'un point de vue historique, géographique et socioculturel. Aux États-Unis, les pauvres vivent en centre-ville. En revanche, lorsque *banlieue* s'intègre dans certaines lexies composées telles que *banlieue de l'ouest parisien*, le sème /pauvreté/ est inhibé par le sème /bourgeois/, afférent à *ouest parisien*, quand même la banlieue ouest de Paris est tout aussi hétérogène que la banlieue dans son ensemble.

Les sèmes contenus dans un signifié ne sont pas tous égaux. Leur nature et leur qualité varient en synchronie (tous les sèmes n'ont pas la même valeur dans le signifié) comme en diachronie (un même sème peut évoluer dans le temps, disparaître, etc.). Les sèmes oscillent entre stabilité et instabilité. Le signifié est constitué de sèmes résistants à la variation, sinon permanents (peut-être les sèmes contenus dans l'intersection du signe associatif de Saussure) et de sèmes instables. Cette variabilité résulte de l'enrichissement ou de l'appauvrissement du signifié à mesure que le mot est actualisé dans les textes. D'une certaine façon, chaque actualisation d'un mot l'enrichit de son contexte d'actualisation, la fréquence de sa participation à un groupement transversal modifie son signifié. Nous dirons, pastichant ainsi une formule archimédienne de Rastier (2001 : 92)<sup>14</sup>, que tout mot placé dans un texte en reçoit des déterminations sémantiques, et modifie potentiellement le signifié de chacun des mots qui le composent.

Ainsi, le sens résulte de négociations entre des signifiés et des formes sémantiques. De la même façon que le cortex visuel traite moins d'informations issues du nerf optique que d'information stockée en mémoire, l'interprétation résulte autant – sinon davantage – d'une activité sémique intense (reconfiguration du signifié, ajustement, convocation des afférences possibles en mémoire, etc.) que du texte lui-même. Rastier et Valette (2009) en présentent quelques exemples dans le cadre de l'évolution sémantique d'une unité lexicale existante (ou *néosémie*).

14 « Tout texte placé dans un corpus en reçoit des déterminations sémantiques, et modifie potentiellement le sens de chacun des textes qui le composent ».

### 2.3.1. Un exemple d'ajustement : la forme sémantique comme *protosémie*

L'élaboration d'un signifié subit des contraintes textuelles et intertextuelles. Les contraintes intertextuelles sont notamment liées aux discours et aux genres, les contraintes (intra)textuelles sont liées à l'économie ou l'organisation sémique du texte<sup>15</sup>. Dans ce contexte, et compte tenu de ce que nous avons présenté dans le paragraphe précédent, nous proposons de considérer la forme sémantique comme le signifié potentiel d'un signe sans signifiant synthétique attitré. Soit l'équivalence hypothétique présentée dans la figure 7.

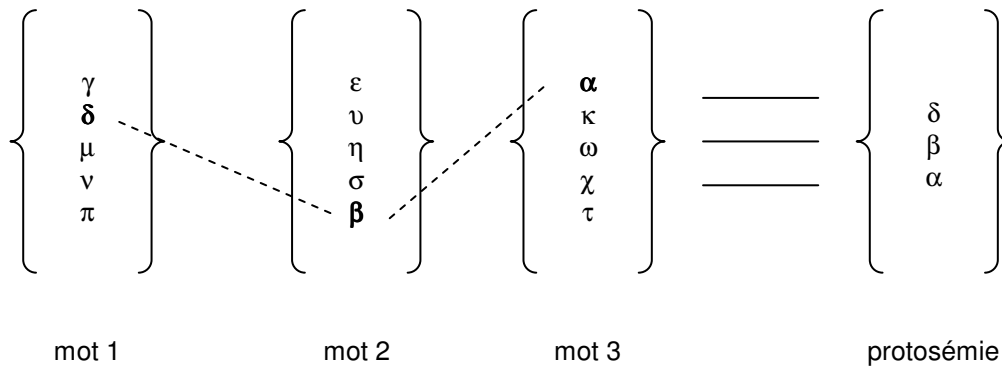


Figure 7 : hypothèse de l'équivalence de la forme sémantique et du signifié

En effet, signifié et forme sémantique sont tous deux identifiables à un groupement sémique de compacité variable, associé à un signifiant stable et synthétique dans le cas du signifié, discontinu et sans lexicalisation privilégiée dans le cas de la forme sémantique. En bref, nous faisons l'hypothèse que certaines formes sémantiques sont potentiellement des signifiés en devenir, ou les signifiés de proto-signes sans signifiant stabilisé ni synthétique attitré. En d'autres termes, certaines formes sémantiques sont des *protosémies*.

Dans le contexte d'une recherche appliquée, en veille lexicale par exemple, le concept de protosémie a pour incidence la possibilité d'identifier le signifié en cours d'élaboration d'un signe en analysant le complexe sémique dans lequel il s'insère. Pour cela, nous empruntons à Rastier (2006) les concepts de < diffusion > et < sommation > qui rendent compte des échanges sémiques entre les fonds et les formes sémantiques. La diffusion relève de propagation des sèmes des formes sémantiques vers le fond sémantique. La sommation ressortit à la propagation du fond sémantique vers des formes sémantiques. Si le signifié peut être considéré à la manière d'une forme sémantique, il est susceptible de se diffuser dans le fond sémantique. Dans ce cas, il nous est possible de restituer une protosémie, c'est-à-dire un signifié-forme sémantique en postulant la sommation de ladite forme à

15 On lira Valette (à paraître) pour un développement.

partir du fond sémantique que nous identifions par le biais d'« isotopies locales ». C'est ce que nous étudierons, analyse à l'appui, dans le paragraphe 3.2.2.

### 3. SÉMIMÉTRIE POUR L'ANALYSE TEXTUELLE DU LEXIQUE

Véronis (2004) et Habert (2005) ont bien mis en évidence l'importance de l'instrumentation TAL pour la validation des propositions théoriques. La mise en œuvre d'un dispositif expérimental destiné à l'étude et à l'automatisation des observations faites précédemment constitue l'objet du paragraphe suivant. Nous y détaillons la réalisation d'une ressource sémique pour l'annotation de corpus et diverses études sur les unités sémantiques et les formes sémantiques destinées à étudier les relations entre le texte et le lexique au moyen des sèmes. Nous évoquons ici la réalisation en cours d'un lexique sémantique alternatif et sans doute complémentaire aux approches lexicographiques telles que WordNet (Miller *et al.* 1990), FrameNet (Fillmore *et al.* 2002), le DiCo (Polguère 2000), ou les classes d'objets du Lexique-Grammaire (Gross 2005), que nous avons décrites ailleurs (Valette 2008). Du point de vue paradigmatique, nous proposons d'envisager les relations entre unités lexicales non pas en termes de construction hiérarchique (comme pour WordNet par exemple) mais en termes de classes sémantiques dont la cohésion est assurée, comme nous l'avons vu, par des sèmes. Du point de vue syntagmatique, l'instanciation des unités lexicales dans les textes ne se fait pas au niveau de la phrase ou de l'énoncé (comme pour FrameNet) mais au niveau des réseaux sémiques. La ressource lexico-sémantique élaborée s'apparente donc à un dictionnaire sémique<sup>16</sup>. À terme, il s'agira d'une collection de dictionnaires sémiques relevant au minimum de discours particuliers et dédiée à des tâches restreintes. Pour l'heure, la ressource est conçue à des fins analytiques et prospectives : notre objectif est d'étudier l'économie générale du contenu sémantique des unités lexicales dans des corpus de textes structurés, dans une perspective tant diachronique que synchronique. De nature prospective et théorique, cette recherche se situe en retrait d'une problématique applicative. Il s'agit d'une application possible de la *sémimétrie*, c'est-à-dire l'application de traitements statistiques et documentaires empruntés ou inspirés de la textométrie et appliqués à des corpus annotés en sèmes.

#### 3.1. Élément de structuration du dictionnaire sémique

Nous avons transformé le *Trésor de la Langue Française informatisé* (Dendien / Pierrel 2003, désormais *TLF*) pour la réalisation du dictionnaire sémique. Le *TLF* est doté de 100 000 mots et de 270 000 définitions. L'extraction fut réalisée sui-

16 Cf. Pincemin (1999) sur l'exploitation du concept de sème en TAL.

vant une hypothèse minimaliste et robuste : une définition est un signifié mis en texte.<sup>17</sup> Ainsi, des définitions, nous retenons les lemmes des substantifs, adjectifs, verbes et de certains adverbes. Ils constituent les *candidats-sèmes* qui forment le signifié d'une unité lexicale en attente d'actualisation. Ce que nous appelons ici et ultérieurement candidat-sème est l'étiquette sémantique résultant d'un traitement automatique. En validant le candidat-sème, le linguiste peut lui octroyer un statut de sème. Mais « comprendre un énoncé, c'est oublier le plus vite possible une grande partie de l'information sémique » nous enseigne Pottier (1992 : 82). C'est à ce titre qu'une part non négligeable des efforts produits dans le cadre de ces recherches consiste à éliminer des candidats-sèmes, tant au niveau paradigmatique de la ressource (il s'agit alors d'éliminer du bruit) que syntagmatique de l'annotation de corpus (il s'agit de sélectionner les bons candidats pour leur allouer le statut de sèmes – et on souhaite cette sélection la plus automatique possible). Beaucoup des recherches rapportées ici visent à ouvrir des pistes pour l'organisation et l'élagage massif des signifiés dans le dictionnaire et dans les textes.

Les premiers travaux (Valette, *e. a.* 2006) ont porté sur la constitution de petites classes sémantiques (de l'ordre du taxème ; cf. *supra*, note 5) au sein d'un domaine donné et sur la structuration des signifiés en fonction des classes obtenues. Ils ont montré que la réalisation locale de classes sémantiques à partir des définitions dictionnairiques était possible, même si elle ne pouvait se substituer à un apprentissage sur corpus. Une étude sur le sème /arbre/ nous a en effet permis de distinguer, par Classification Ascendante Hiérarchique (CAH), des sous-classes pertinentes, tant d'un point de vue gnosique (Essences d'arbre, Parasites) que pratique (Plantation, Bûcheronnage, Arboriculture) sans que de telles classes n'aient été dessinées *a priori*, ni par nous, ni par les lexicographes du *TLF*. Puis, nous avons procédé à une pondération interne des signifiés à l'aide d'un calcul d'écart réduit à l'intérieur de la classe la plus importante, celle des < essences d'arbres >. Pour ce faire, nous avons pondéré tous les signifiés de façon à en dégager l'organisation interne relative à la classe considérée. L'objectif est de dégager les sèmes génériques de chaque classe et les sèmes spécifiques des éléments des différentes classes. Dans ce contexte, l'écart réduit se calcule ainsi :

$$z = \frac{f_E - f_c * p}{\sqrt{f_c * p * q}}$$

– où  $f_E$  est la fréquence du sème observée dans l'entrée,  $f_c$  la fréquence du sème observée dans la classe,  $p$  est la proportion de l'entrée dans la classe et  $q$  le complément de  $p$ . Pour l'entrée  $i$  de la classe  $C_K$ , cela donne :<sup>18</sup>

17 Cf. Rastier 1987 : 41 ; Martin 2001.

18 N. B. : dans cette étude, les travaux statistiques ont été réalisés par A. Estacio-Moreno.

$$p_{i,K} = \frac{\sum_{i=1}^{|V|} E_i}{N_K} \quad \text{et} \quad q_{i,K} = 1 - p_{i,K}$$

La pondération des signifiés de chaque entrée à l'aide de l'écart réduit de la sous-classe des Essences d'arbres donne des résultats satisfaisants. Pour chaque signifié, les candidats-sèmes génériques (c'est-à-dire ceux dont l'écart réduit est le plus bas) permettent de distinguer des sous-classes potentielles (par exemple : arbres valant pour leur fruit, arbre valant pour leur bois). L'organisation du signifié a une incidence sur les candidats-sèmes spécifiques. Ainsi par exemple, le manglier est caractérisé par les endroits où il croît (sèmes spécifiques /plage/, /lagune/, /maritime/) (figure 8).

Candidats-sèmes	Valeur de l'écart réduit
/plage/	18,04
/lagune/	18,04
/maritime/	18,04
/intertropical/	12,72
/rhizophoracées/	12,72
/croître/	4,03
/région/	2,88
/famille/	1,46

Figure 8 : Signifié pondéré de *manglier*. Définition : « Arbre de la famille des Rhizophoracées, qui croît dans les lagunes et les plages maritimes des régions intertropicales » (extrait de Valette, *e. a.* 2006 : 364).

Au vu des résultats obtenus, la réalisation préliminaire de classes sémantiques à partir de définitions dictionnairiques semble possible. Les items retenus qui composent une définition peuvent être légitimement considérés comme des sèmes minimaux et ce, malgré une segmentation sommaire et la perte sensible d'information que l'éclatement syntaxique induit. Admettons toutefois que le domaine choisi, la sylviculture, est exceptionnellement bien structuré et correspond à un savoir davantage encyclopédique que praxéologique. Mais la caractérisation des signifiés à l'intérieur d'une classe donne à voir une organisation sémantiquement pertinente, où les sèmes spécifiques sont susceptibles d'être opérationnels, notamment dans la perspective textuelle qui est la nôtre. En effet, les sèmes /plage/, /lagune/ caractérisent mieux le manglier que son appartenance à la famille des rhizophoracées. En d'autres termes, le sens apparaît privilégié par rapport à la référence.

L'étude présentée ici avait, rappelons-le, une visée exploratoire. Il s'agissait de déterminer la pertinence du recours à un dictionnaire de langue dans l'élaboration d'un dictionnaire de sèmes initial destiné à l'annotation de corpus. Les résultats se sont avérés positifs mais des apprentissages sur corpus ciblés, en faisant apparaître la régularité des réseaux sémiqques, permettront de stabiliser les signifiés. Cette seconde phase aura notamment pour effet de valider ou mettre à jour les résultats de la classification première.

### 3.2. Des objectivations sémantiques

Le dictionnaire sémique donne ainsi l'opportunité d'effectuer un certain nombre d'expérimentations d'annotation. On abordera dans ce paragraphe différents travaux exploratoires visant à évaluer son utilisation dans l'identification des fonds et des formes sémantiques.

#### 3.2.1. Le fond sémantique

(Grzesitchak, *e. a.* 2007) ont montré qu'il était possible d'observer des récurrences de sèmes dans un texte isolé. Par exemple, dans un article du *Monde diplomatique* qui traite de l'administration de la ville de Toulon par le Front National (« Toulon, ville amirale du front national », juillet 96), le candidat-sème /harceler/ apparaît dans 90 % des paragraphes alors que ni le mot *harceler*, ni ses dérivés morphologiques ne sont actualisés dans le texte. De même, dans l'article « Replis communautaires à Sarcelles » (février 96) où il est essentiellement question des jeunes de la ville de Sarcelles, le sème /enfant/ apparaît parmi les 10 candidats-sèmes les plus saillants. Or, le mot *enfant* est toujours absent du texte. Mais si, précédemment, on pouvait interpréter le candidat singularisé en termes d'impression ou d'idée générale (l'idée de harcèlement), on pourrait avancer que l'auteur ici, pratique l'euphémisme en employant des mots tels que *jeune* ou *adolescent* pour qualifier les protagonistes. Le sème isotopique isolé donnerait accès plus crûment à la réalité des faits. En bref, l'article parle des jeunes de Sarcelles ; le fond sémantique présumerait qu'il s'agit d'enfants et rappellerait incidemment que les adolescents sont des enfants. Encore expérimentale, cette recherche sur les régions peu explorées de l'infralexicalité devrait bénéficier d'une typologie des sèmes et des isotopies.

#### 3.2.2. Les formes sémantiques

(Reutenauer, *e. a.* 2010) évaluent les déformations subies par une forme sémantique dont l'élément stable est constitué de l'unité lexicale *économie réelle* dans un corpus de presse constitué de 1587 articles, tirés du *Figaro* et de l'*Humanité*

entre septembre 2008 et février 2009. Le thème du corpus est la crise économique et financière. Il se présente sous forme de deux versions parallèles : la version lexicale, d'un million d'occurrences de formes, et une image sémique de ce même corpus, composé de 23 millions d'occurrences de candidats-sèmes. À partir d'un calcul de spécificités (implémentation Lexico3, Salem, *e. a.* 2003), les auteurs mesurent la sensibilité des informations au contexte éditorial des deux quotidiens du corpus. Ainsi, au voisinage d'*économie réelle*, *l'Humanité* active les sèmes /bien/ (substantif), /revenu/, /consommation/ et /dépense/, tandis que /ressource/ et /économie/ sont actualisés par *le Figaro*. Les auteurs avancent l'hypothèse aisément corroborable d'une perception plus macroéconomique de l'économie dans *le Figaro* et plus localisante dans *l'Humanité*. En dépit d'un bruit persistant, l'étude de Reutenauer, *e. a.* montre une convergence des résultats sur les plans sémique et lexical.

S'appuyant sur une méthodologie similaire, (Reutenauer, *e. a.* à paraître) effectuent une analyse de l'environnement sémique d'un signifié en cours d'élaboration, celui d'*Outreau*. Le corpus porte donc sur l'affaire judiciaire dont la ville d'Outreau est l'éponyme. Divisé en cinq périodes, il est constitué d'articles de presse publiés entre novembre 2001 et avril 2006 comprenant au moins une occurrence du mot étudié. Lecolle (2007 : 101), à laquelle est emprunté le corpus, observe que le sens d'*Outreau* évolue du toponyme à « l'erreur judiciaire par excellence ». Comme dans l'étude précédente, le corpus se présente sous deux versions parallèles : la version lexicale de 400 000 occurrences de formes (empruntée à Lecolle 2007) et une image sémique de 10 millions de candidats dont a été extraite une sous-image constituée des seuls candidats correspondants à des formes rendues saillantes par un calcul de spécificité effectué là encore avec Lexico3. Ainsi, les auteurs ont à manipuler une sélection de quelques dizaines de candidats-sèmes seulement.

L'une des analyses effectuées par les auteurs nous intéresse particulièrement. À partir de listes de candidats-sèmes spécifiques à chaque période, des regroupements sémantiques sont réalisés manuellement. Selon nous, ces regroupements peuvent être assimilés à des isotopies. Par exemple, le regroupement de candidats tels que /police/, /procureur/, /écrouer/ s'apparente à une isotopie domaniale //judiciaire// particulièrement présente dans les périodes précoces du corpus. En son sein, on peut observer les traces de différents champs génériques : l'//arrestation// apparaît à travers l'ensemble de candidats-sèmes {/écrouer/, /police/, /arrestation/, /incarcération/, /incarcérer/, /prévenu/}. L'isotopie domaniale //politique// préfigurant le sens d'« erreur judiciaire par excellence », apparaît dans la quatrième période et se renforce ensuite. Inversement, l'isotopie taxémique des //dénominations de crimes// (comportant des candidats tels que /pédophilie/, /meurtre/, /viol/) décroît en importance en quatrième période puis disparaît. Les isotopies taxémiques locatives (//lieu d'habitation//, //lieu géographique//) ne sont représentées significativement qu'à la première période. L'isotopie //fiasco// (constituée de /nauffrage/, /drame/, /faillite/, /faute/) est représentative de la cinquième période.

Si nous considérons que les faisceaux d'isotopies locales observables ici sont le résultat d'une diffusion de la forme sémantique d'*Outreau* vers le fond (cf. *supra*, 2.3.1.), nous pourrions, par *simulation* sommatrice inverse, restituer la forme sémantique, c'est-à-dire le signifié en cours d'élaboration d'*Outreau*, lequel est minimalement définissable par un changement de sème générique : le sème domanial //judiciaire// se substitue au sème inhérent toponymique /ville/, avant d'être inhibé pour céder la place au sème domanial //politique//.<sup>19</sup>

#### 4. PUISSANCE ET HUMILITÉ DU SÈME

En estimant qu'« on pourrait déterminer les différents âges d'une science par la technique de ses instruments de mesure », Bachelard (1938 : 216) suggérait que la maturation des sciences s'accompagne d'une dotation d'instruments (Herbert 2005). Pourtant, longtemps dominant, le paradigme introspectif du générativisme posait que la compétence du locuteur (et notamment du locuteur linguiste) était une sanction suffisante à l'établissement et l'analyse des données langagières. Ainsi, le linguiste pouvait légitimement les produire pourvu qu'elles lui paraissent bien construites (critères de bonne formation et de grammaticalité). Dans cette perspective, les épistémologues générativistes observèrent que la linguistique était une science sans observatoire, sans instrument ou dont l'instrument était « mental » (Milner 1989).

La linguistique de corpus propose de collecter des données attestées (c'est-à-dire non produites par le linguiste pour les besoins de l'analyse) et repose sur deux conditions d'observation conjointes : d'une part, l'usage d'outils d'analyse, d'autre part, la nécessité d'un certain volume de données. Ainsi, d'un côté le corpus impose l'usage d'instruments d'observation, de l'autre, l'instrumentation permet d'obtenir des mesures objectives. C'est ainsi que la linguistique de corpus s'est développée en même temps que les statistiques lexicales (Guiraud 1960, Muller 1964). À l'outillage voué à la lecture (indexation, concordancier), on a associé des outils dédiés au comptage et à la pondération (calcul de l'écart réduit, loi hypergéométrique, etc.).

Que la science soit instrumentée ou non, et à moins d'adopter une position empiriste radicale, les données observables sont toujours des constructions. Doter la linguistique d'une instrumentation implique aussi la production des données, c'est-à-dire la constitution des corpus de textes et de plus en plus souvent celle de données métalinguistiques, d'observables qui ne sont plus des données empiriques.

Les sèmes sont ces observables. Ils ne sont pas des universaux, ils sont des traits sémantiques, des valeurs modestes élaborées, construites par le linguiste ou

19 Pour une discussion sur les changements de domaines dans l'économie sémique d'une néologie sémantique, cf. Rastier / Valette (2009).

par quelque artifice mécanique pour les besoins d'un texte ou d'un corpus. Face aux universaux, aux invariants et aux primitives, qui sont des géants, puissants, translingues et peu nombreux, les sèmes opposent leur indénumbrabilité (l'inventaire en est inachevé, il s'en crée chaque fois que nécessaire de nouveau) et surtout leur dépendance déterminante à la langue, la parole, la pratique et la culture. En définitive, s'ils n'ont guère de valeur intrinsèque, leurs associations – dans les textes en particulier – sont leur puissance.

## 5. BIBLIOGRAPHIE

- Bachelard, Gaston, *La formation de l'esprit scientifique*, Paris : Vrin, 1938.
- Blumenthal, Peter / Hausmann, Franz Josef (éds.), *Collocations, corpus, dictionnaires (Langue française, 150/2)*, Paris : Larousse, 2006.
- Duteil-Mougel, Carine, « Introduction à la sémantique interprétative », in : *Texto ! Textes et cultures* (2004), [cf : <http://www.revue-texto.net/>].
- Fillmore, Charles J. / Baker, Collin F. / Sato, Hiroaki, « The FrameNet Database and Software Tools », in : *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC) (Las Palmas 2002)*, p. 1157–1160.
- Gross, Gaston, « Un dictionnaire électronique des adjectifs du français », in : *Cahiers de Lexicologie* 86 (2005), p. 11–33.
- Grzesitchak, Mick / Jacquey, Évelyne / Valette, Mathieu, « Systèmes complexes et analyse textuelle : Traits sémantiques et recherche d'isotopies », in : *ARCo'07 – Cognition, Complexité. Collectif (Acta-Cognitica)*, 2007, p. 227–235.
- Guiraud, Pierre, *Problèmes et méthodes de la statistique linguistique*, Paris : PUF, 1960.
- Habert, Benoît, « Portrait de linguiste(s) à l'instrument », in : *Texto ! Textes et cultures* (2005), [cf. <http://www.revue-texto.net/>].
- Lecolle, Michelle, « Polysignifiante du toponyme, historicité du sens et interprétation en corpus. Le cas de *Outreau* », in : *Corpus* 6 (2007), p. 101–125.
- Martin, Robert, *Sémantique et automate*, Paris : PUF, 2001.
- Mel'čuk, Igor / Clas, André / Polguère, Alain, *Introduction à la lexicologie explicative et combinatoire*, Louvain-la-Neuve : Duculot, 1995.
- Miller, George A., Beckwith, Richard T. / Fellbaum, Christiane D. / Gross, Derek / Miller, Katherine J., « WordNet : An on-line lexical database », in : *International Journal of Lexicography* 3 / 4 (1990), p. 235–244.
- Milner, Jean-Claude, *Introduction à une science du langage*, Paris : Seuil, 1989.
- Missire, Régis, *Sémantique des textes et modèle morphosémantique de l'interprétation*. Thèse de doctorat Université de Toulouse II Le Mirail, in : *Texto ! Textes et cultures* (2006), [cf. : <http://www.revue-texto.net/>].
- Muller, Charles, *Essai de statistique lexicale. L'illusion comique de Pierre Corneille*, Paris : Klincksieck, 1964.
- Pincemin, Bénédicte, « Sémantique interprétative et analyses automatiques de textes : que deviennent les sèmes ? », in : *Sémiotiques* 17 (1999), p. 71–120.
- Polguère, Alain, « Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French », in : *Proceedings of EURALEX'2000*, Stuttgart : Euralex, 2000, p. 517–527.
- Pottier, Bernard, *Théorie et analyse en linguistique*, Paris : Hachette, 1992 [<sup>1</sup>1987].
- Rastier, François, *Sémantique interprétative*, Paris : PUF, 1987.
- Rastier, François, *Arts et sciences du texte*, Paris : PUF, 2001.

- Rastier, François, « Formes sémantiques et textualité », in : *Langages* 163 (2006), p. 99–114.
- Rastier, François, « Passages », in : *Corpus* 6 (2007), p. 125–152.
- Rastier, François / Valette, Mathieu, « De la polysémie à la néosémie », in : *Le français moderne* 77 (2009), p. 97–116.
- Reutenauer, Coralie / Valette, Mathieu / Jacquy, Évelyne, « De l'annotation sémique globale à l'interprétation locale : environnement et image sémiques d'« économie réelle » dans un corpus sur la crise financière », in : *Cognita – Actes du colloque de l'Association pour la Recherche Cognitive ARCo'09 : Interprétation et problématiques du sens (9–11 novembre 2009)*, Rouen : Université de Rouen, 2010, p. 29–39  
[cf. <http://arco09.colloques.univ-rouen.fr/>].
- Reutenauer, Coralie / Lecolle, Michelle / Jacquy, Évelyne / Valette, Mathieu, « Sémème au microscope : genèse et variation sémiques d'une unité lexicale », in : *Actes JADT'2010 (9–11 Juin 2010)*, Rome [à paraître fin 2010].
- Salem, André / Lamalle, Cédric / Martinez, William / Fleury, Serge / Fracchiolla, Béatrice / Kuncova, André / Maisondieu, Aude, « Lexico3 – Outils de statistique textuelle. Manuel d'utilisation », in : *Syled-CLA2T*, Paris : Université de la Sorbonne nouvelle – Paris 3, 2003.
- Saussure, Ferdinand de, *Cours de linguistique générale*, Paris : Payot, 1976 [<sup>1</sup>1916].
- Saussure, Ferdinand de, *Écrits de linguistique générale*, Paris : Gallimard, 2002.
- Valette, Mathieu, « À quoi servent les lexiques sémantiques ? Discussion et proposition », in : *Cahiers du CENTAL* 5 (2008), p. 43–58.
- Valette, Mathieu, « Méthodes pour la veille lexicale », in : *Actes de la journée d'étude « Le dictionnaire électronique. Quelles perspectives pour les sciences humaines et sociales ? »*, éd. par L. Messaoudi, Publication du laboratoire Langage et société, Université Ibn Tofail Kénitra [à paraître, disponible sur <http://hal.archives-ouvertes.fr/>].
- Valette, Mathieu / Estacio-Moreno, Alexander / Petitjean, Étienne / Jacquy, Évelyne, « Éléments pour la génération de classes sémantiques à partir de définitions lexicographiques. Pour une approche sémique du sens », in : *TALN 06 / 1 : Verbum ex machina*, éd. par P. Mertens, *e. a.* (Cahiers du CENTAL), Louvain : Presses Universitaires, 2006, p. 357–366.
- Véronis, Jean, « Annotation automatique de corpus : panorama et état de la technique », in : *Ingénierie des langues*, éd. par J.-M. Pierrel, Paris : Hermès, 2000, p. 111–129.