Vers un traitement automatique de la néosémie : approche textuelle et statistique

Coralie Reutenauer

Nancy-Université
Université Nancy 2

Vers un traitement automatique de la néosémie : approche textuelle et statistique.

THÈSE

présentée et soutenue publiquement par

Coralie REUTENAUER

le 20 janvier 2012

en vue de l'obtention du titre de

DOCTEUR DE l'Université de Lorraine

Spécialité : **Sciences de langage**

préparée au laboratoire **ATILF (UMR 7118)**

dans le cadre de l'École Doctorale **Langages, Temps et Société**

Directeur de thèse : **Jean-Marie PIERREL**
Co-directeurs de thèse : **Evelyne JACQUEY, Mathieu VALETTE**

Jury :

Rapporteurs :

Anne CONDAMINES, Directrice de recherche CNRS, CLLE-ERSS, Toulouse
Jean-François SABLAYROLLES, Professeur de l'Université Paris XIII, LDI, Paris

Examinateurs :

Ludovic LEBART, Directeur de recherche CNRS, Telecom-Paristech, Paris
Alain POLGUÈRE, Professeur des universités, Université de Lorraine, ATILF, Nancy
Jean-Marie PIERREL, Professeur des universités, Université de Lorraine, ATILF, Nancy
Evelyne JACQUEY, Chargée de recherche, CNRS, ATILF, Nancy
Mathieu VALETTE, Professeur des universités, Inalco, Paris

atilf
ANALYSE ET TRAITEMENT INFORMATIQUE DE LA LANGUE FRANÇAISE

Au fond de l'inconnu pour trouver du nouveau !

Baudelaire, Les Fleurs du Mal, « Le Voyage »

Remerciements

Pendant ces années où j'ai mené ma thèse, j'ai cherché à construire du sens. Pas en solitaire, car une thèse est une co-construction. Le sens, au fond, je l'ai trouvé dans ceux qui m'ont entourée et accompagnée.

Mes remerciements vont en priorité à mes encadrants, sans lesquels je ne me serais pas engagée et grâce auxquels je ne regrette pas de m'être engagée dans ce parcours du doctorant.

Merci à Jean-Marie Pierrel, pour son soutien indéfectible. Conditions matérielles en or, disponibilité, confiance, c'est sur tous les plans que ce soutien s'est traduit. À force de *pourquoi ?*, il m'a poussée à aller toujours plus haut, vers une vue surplombante, vers la défense d'une thèse. Et aussi vers la recherche, à laquelle il m'a fait prendre goût plus que je ne l'aurais voulu. Un protecteur, de ceux qui mènent vers les sommets.

Merci à Evelyne Jacquey, coéquipière du quotidien. Présente sur tous les fronts, scientifique, psychologique, matériel, elle est quelqu'un sur qui j'ai toujours pu compter et qui m'a fait évoluer vers toujours plus de qualité et de maturité. Elle est de ceux avec qui le travail est un plaisir et les obstacles des défis qu'on surmonte. Cette thèse lui doit beaucoup, je lui dois plus encore. Un pilier, une force qui va et qui m'a emmenée de l'avant.

Merci à Mathieu Valette, un guide sur les sentiers de la sémantique. Je garde en tête sa gentillesse et son humour récurrents, notamment face à mes questions existentielles. À défaut de me redomanialiser parfaitement en linguistique, il m'a fait vivre la thèse en néosémie, où l'on retrouve, transposés, les mécanismes des parcours interprétatifs et de l'enrichissement. Un initiateur et un fond sémantique.

Je remercie les membres du jury pour leur implication dans l'évaluation de mon travail et l'intérêt qu'ils y ont porté. Je tiens à témoigner ma reconnaissance à mes rapporteurs pour leur lecture attentive et leurs remarques détaillées. Je remercie Anne Condamines pour ses questions d'ouverture, portes sur d'autres horizons, et Jean-François Sablayrolles pour son regard précis, qui me permettra de combler certaines lacunes et de consolider mon positionnement linguistique. Je remercie Ludovic Lebart pour l'interaction constructive en amont sur le chapitre 5, pour sa façon de rectifier et d'aiguiller en douceur et pour sa curiosité à l'égard de l'ensemble de ma recherche. Merci à Alain Polguère pour ses remarques structurantes, qui devraient m'amener vers un produit plus clair et encore plus solide. Le regard particulier de chacun m'a amenée à changer de perspective, à voir certains aspects sous un autre angle. Les échanges m'ont fait entrevoir les failles à consolider et surtout les rayonnements possibles. Ces échanges ont aussi été menés de telle sorte que j'ai eu le sentiment d'interagir avec des pairs, qui avaient une réelle considération pour mon travail. J'en ai été très touchée.

J'aimerais adresser un merci particulier à certains collègues, artisans actifs de l'édifice. Je dois beaucoup à Michelle Lecolle et à sa générosité. Générosité pour le corpus qu'elle m'a transmis, mais surtout pour tout le reste : beaucoup de temps, de disponibilité, d'attention, des échanges constructifs et des regards d'une grande acuité. Une collègue précieuse, qui elle aussi m'a formée, et une amie plus que fiable. Un grand merci également à Sandrine Ollinger,

une adjuvante du quotidien, pour ses réponses à mes SOS scripts, sa présence attentive à bon nombre de mes présentations blanches, ses questions et remarques intelligentes. Un cocktail réussi de gentillesse, de finesse d'esprit et d'efficacité. Merci aussi à Mick Grzesitchak, dont la plateforme Semy a été déterminante pour mes expériences et qui, en quelques lignes de code, a réussi à lever mes blocages à plusieurs reprises.

Ma reconnaissance va également à des aînés attentifs, qui ont porté un intérêt à mes travaux et avec qui j'ai eu l'occasion d'interagir à diverses reprises. Merci à François Rastier et à Bénédicte Pincemin pour leur soutien, leur présence, leurs conseils, et dont les pensées scientifiques ont été des guides.

Si je me suis sentie portée, intégrée, soutenue, c'est aussi grâce à l'ATILF, mon laboratoire d'accueil. Un tout aux multiples facettes, un chez-moi où beaucoup de visages ont donné du sens à mon quotidien. Je pense en particulier aux doctorants, avec une mention spéciale pour Aurore et Lolita, un duo de choc qui m'a apporté une belle dose de bonne humeur, de rire et de solidarité. Merci aussi à Inga et Cécile qui, à travers des échanges constructifs et des pensées amicales, ont été des éléments porteurs, ainsi qu'à Dorota, toujours avec un mot gentil et prête à m'apporter de l'aide. Un autre personnage-clé de mon quotidien est Gilles, un voisin de bureau idéal humainement et matériellement, à qui je dois des coups de pouce divers et variés, une présence agréable ainsi qu'un magnifique paysage de cactus. Quelqu'un qui a beaucoup compté est Simone, un rayon de soleil à chaque entrée par la grande porte et une navigatrice de bibliothèque hors pair. Merci aussi à Pascale, mine de renseignements sur le TLF toujours disponible, à Véronika, Gérard, Dominique, Isabelle, Véronique, Eva, Jean-Yves, William, Laurent, Olivier, Michèle, Christiane, Cyril et tant d'autres qui ont été là, avec une parole gentille, un coup de main efficace, une pensée encourageante.

J'aimerais adresser une pensée particulière à Pierre Chauvet. Le premier lien avec l'ATILF, c'est à lui que je le dois, lui qui m'a accompagnée dans mes premiers pas, lui qui est toujours resté attentif et bienveillant lorsque j'ai poursuivi ma route.

Je tiens enfin à remercier tous ceux qui ont partagé des moments de vie avec moi, autour de la thèse. Dans les moments de rébellion contre l'intellectuel, c'est eux qui m'ont soutenue, qui m'ont permis d'aller de l'avant et qui m'ont donné d'autres centres, d'autres terrains d'envol. Merci à ma famille, qui m'a toujours suivie quelles que soient mes décisions : mon père qui a toujours aimé les chemins de la linguistique, curieux de mon parcours ; ma mère à l'écoute s'il y avait des creux de vague ou besoin de voir clair ; Elo, sa présence pétillante et son affection protectrice, qui a veillé de loin, toujours prête à m'épauler ; Jean qui, par ses petits signes et ses messages drôles, a été un magicien des sourires, capable d'ancrer dans la vie, apaisant et en même temps vigilant. Merci à ceux qui m'ont donné du rire et qui ont fait sauter mes carapaces, Seb, Pierre, Yves, les catalyseurs de l'hors-thèse. Merci à mes amis, que j'ai toujours sentis à mes côtés malgré mes absences, Colette, Jean-Paul, Micheline, Michel Paradis, Ruben, Carole, Camille, Guillaume, Romain, Yohann, Louise, Nadine et Eben. Ces années n'ont pas été qu'une construction intellectuelle, mais une ascension à bien d'autres niveaux, et c'est à eux tous que je le dois.

Je remercie enfin les organismes qui m'ont permis de mener à bien cette thèse. Je suis reconnaissante envers le CNRS, l'Université de Lorraine et en particulier l'École Doctorale Langage, Temps, Société pour leur soutien financier. Je suis reconnaissante envers l'IUT Nancy Charlemagne qui m'a accueillie comme monitrice, et plus particulièrement à l'équipe d'enseignants en mathématiques. Merci enfin au CIES pour la formation qu'il m'a délivrée.

À vous tous qui, d'une manière ou d'une autre, avez été à mes côtés, merci.

Résumé

Résumé.

L'enjeu de cette thèse est l'acquisition automatique de nouveaux sens lexicaux.

Nous définissons un modèle théorique sur l'émergence d'un nouveau sens pour une unité lexicale ayant déjà un sens codé. Le phénomène ciblé est la néologie sémantique, ou néosémie, définie comme une variation sémantique marquée en cours de diffusion. Nous la modélisons à partir d'indices quantitatifs articulés à des principes issus de la sémantique textuelle. Le sens codé est représenté comme un ensemble structuré de traits sémantiques. Il est modulé en discours sous l'effet de récurrences d'autres traits. La dynamique du sens est représentée à l'aide de descripteurs de granularité sémantique variable.

Ensuite, nous proposons des ressources et outils adaptés, relevant de la linguistique de corpus. Les ressources sont de deux types, lexicographiques pour le sens codé et textuelles pour le sens en discours. En pratique, le *Trésor de la Langue Française informatisé* fournit les sens codés. Une plateforme transforme ses définitions en ensembles de traits sémantiques. Trois corpus journalistiques des années 2000 servent de ressources textuelles. Les outils mathématiques, essentiellement statistiques, permettent de jouer sur la structure des ressources, d'extraire des unités saillantes et d'organiser l'information.

Enfin, nous établissons les grandes lignes d'une procédure pour allouer de façon semi-automatique un nouveau sens. Elles sont étayées par des expériences illustratives. Le déroulement de la procédure repose sur des niveaux de description de plus en plus fins (domaines, unités lexicales puis traits sémantiques). Il s'appuie sur des jeux de contrastes multiples, permettant de nuancer l'information sémantique.

Mots-clés : néologie sémantique, nouveau sens, lexique, dictionnaire, textométrie, corpus, acquisition automatique, traits sémantiques, indices statistiques, spécificités, description sémantique multiniveaux.

Title: Automating meaning acquisition: a textual and statistical approach

Abstract:

The issue at stake is the automated meaning allocation.

In a first time, a theoretical scheme is elaborated to describe meaning change for a lexical unit already defined in a lexical resource. We focus on semantic neology, considered as a significant repeted change. Our model relies on quantitative evidence and it is inspired from text semantics. The preexisting meaning is represented as a structured set of semantic features. The context modifies it due to salient semantic featuresin texts. These dynamic change is comprehended through description strata ranging from coarse-grained to fine-grained semantic units.

In a second time, we dwell on relevant resources and tools from corpus linguistics. The resources are dictionaries and text corpus. Concretely, we use the *Trésor de la Langue Française informatisé* as a dictionary. Its entries are automatically converted into bags of semantic features. The textual data consists in three recent journalistic corpus. The resources are considered are mathematic spaces and statistical tools are used to extract significant units and to structure information.

In a last time, we give an outline of a process to allocate automatically a new meaning. Experiments illustrate each step. This process relies on multiple levels of description, getting finer and finer. Through this approach, it is possible to qualify the new meaning in a precise and structured way.

Keywords: neology, new meaning, meaning aquisition, dictionary, corpus, statistical score, lexical processing, multilevel sense description

Table des matières

Table des figures

Introduction

Le sens se construit en contexte, de façon variable d'un contexte à l'autre ; du moment où elle est employée, toute unité lexicale est soumise à des variations sémantiques.

Pour un individu, cette variabilité n'est pas un obstacle pour accéder au sens. En effet, pour comprendre un mot dans un contexte donné, tout locuteur effectue un parcours interprétatif. Ce parcours interprétatif ne relève pas seulement d'une capacité, mais d'un impératif car on ne peut s'empêcher d'interpréter ce qu'on entend ou lit, autrement dit l'interprétation est compulsive. De ce fait, en contexte, chacun construira dans la majorité des cas une interprétation, sans que le sens soit ambigu.

Il y a donc nécessairement accès au sens en contexte, mais de façon plus ou moins immédiate. Certains emplois paraîtront normaux, le sens semblera facilement accessible ; d'autres emplois, en revanche, susciteront un sentiment de nouveauté ou de singularité chez l'interprétant. Le sens d'une lexie s'adaptera donc de façon plus ou moins consensuelle ou conflictuelle aux contraintes sémantiques imposées par le contexte.

En traitement automatique, la variation sémantique est un problème non résolu. À l'heure actuelle, on ne sait pas construire automatiquement le sens précis d'un mot dans un contexte donné. En particulier, les modèles formels sur l'apport précis du contexte au contenu sémantique d'un mot sont encore peu élaborés.

La modélisation du sens contextuel n'est résolue ni pour des variations fines, où l'accès au sens est relativement immédiat mais se nuance d'un emploi à l'autre (dans les propositions "électronique moléculaire ou l'art de suivre les électrons au sein des molécules" et "cancérologie : voir et soigner avec l'imagerie moléculaire", le sens de *moléculaire* est proche : il renvoie à la dimension microscopique, à l'idée de particules et à des domaines scientifiques, même si, dans le premier cas, les propriétés actives et la dynamique des particules ressortent plus, tandis que l'idée d'infiniment petit est plus saillante dans le second cas), ni pour des variations brutales où l'accès au sens génère un sentiment de nouveauté (le sens de *moléculaire* dans "Dîner moléculaire pour deux à Lille" contraste fortement avec les précédents : il renvoie à l'idée de réaction chimique et de fondement scientifique, mais aussi, par assimilation de la discipline scientifique qu'est la gastronomie moléculaire à ses applications, à quelque chose à la fois de non naturel (sans connotation négative), de déconcertant et d'esthétique).

Le présent travail se concentre sur les variations sémantiques marquées. On entend par "variations marquées" des variations suffisamment importantes pour générer un sentiment de nouveauté et pour rétroagir en profondeur sur le sens associé à un mot. Elles se caractérisent

par un caractère brutal et massif : elles sont le reflet d'une rupture avec les sens connus et elles connaissent une certaine diffusion. L'objectif sera de construire automatiquement le sens d'un mot lorsqu'on est confronté à un emploi qu'on ne connaît pas, c'est-à-dire à un sens nouveau.

Le choix de ce type de variations est motivé par deux raisons. La première raison est pratique : il semble plus abordable d'élaborer un traitement sur des variations marquées que sur des variations fines. On peut supposer que les indices d'une rupture seront plus saillants que ceux d'une variation fine. La deuxième raison est rattachée au degré d'intérêt que présentent les différents cas de variation sémantique. Les outils d'aide à l'interprétation seront plus utiles lorsque l'interprétation se heurte à une difficulté que lorsqu'elle est immédiate : les sens existants constituent une bonne approximation d'un nouveau sens qui ne s'en distingue que faiblement, l'essentiel de l'information nécessaire à l'interprétation est présente dans les sens existants ; en revanche, l'approximation ne suffit pas lorsque l'écart sémantique est important, objectiver l'interprétation nécessite d'expliciter de nouveaux éléments de définition. Ajoutons qu'une variation qui suscite un sentiment de nouveauté et qui se reproduit peut rétroagir sur le système de connaissances, afin qu'il s'ajuste au nouveau type d'emploi et intègre le nouveau sens. Il n'y a donc pas que l'interprétation en contexte qui est affectée, la représentation cognitive du sens associé au mot peut aussi être modifiée.

Ces cas limites de rupture ne concernent, certes, qu'une petite partie de la variation sémantique, mais ils ne sont pas pour autant négligeables. Ils participent d'un principe essentiel d'une langue vivante : une langue vivante est une langue qui évolue ; cette évolution se traduit sur le plan sémantique par l'émergence de nouveaux sens, reflets de variations sémantiques brutales.

De tels phénomènes ne sont pas des créations de l'esprit. Un regard vers le passé témoigne de leur existence. Nombre de mots connaissaient des emplois dont le sens paraîtrait discordant, voire incompatible avec le sens actuel. Par exemple, l'emploi de *décevoir* pour *tromper*, *duper,* paraît aujourd'hui décalé et il est par exemple considéré comme vieilli dans le *Trésor de la Langue Française*. Plus récemment, vers la fin des années 80, le *surbooking* désignait une surréservation relativement à des hôtels, avions, ou autres structures d'hébergement et de transport. Depuis, ce mot a connu une évolution morphologique et sémantique pour s'appliquer aujourd'hui à des agendas, des emplois du temps et, par métonymie, à des personnes (*"je suis surbooké"*). Les changements sémantiques ont jalonné le passé. Par analogie, il est légitime de supposer que des changements sémantiques ont lieu actuellement et qu'ils contribuent à introduire, diffuser et stabiliser de nouveaux sens dans la langue.

Un autre témoin de l'existence de tels changements est la perception qu'on peut en avoir. Ainsi, de façon individuelle et subjective, je perçois certains emplois comme nouveaux. Ce sentiment de nouveauté peut être dû à une ignorance de ma part, donc à des lacunes personnelles que je peux combler grâce au savoir partagé de la communauté à laquelle j'appartiens, ou il peut être dû à quelque chose de nouveau aussi bien pour ma communauté que pour moi-même. Or des indices montrent qu'un sentiment de nouveauté peut être partagé : existence de guillemets dans des textes, demande de précision ou d'éclaircissement sur le sens d'un mot en dialogue, etc. La perception de ruptures ou nouveautés sémantiques, à titre individuel aussi bien que collectif, témoigne donc de leur existence.

Les arguments d'analogie avec le passé et de perception de ruptures sémantiques soulignent qu'il y a bien émergence de nouveaux sens. Toutefois, ce phénomène ne s'inscrit pas

simplement dans la mouvance de passé, il connaît une nouvelle dynamique avec les changements de la société. Ainsi, les nouveaux modes de communication accélèrent et amplifient la circulation du dire, de nouvelles pratiques langagières, moins contraintes donc plus perméables à de l'innovation sémantique, se mettent en place, par exemple à travers les blogs, les SMS, etc. Cette nouvelle dynamique contribue à renforcer les enjeux de l'étude des néologies sémantiques.

Les enjeux du présent travail sont à la fois scientifiques, sociétaux et économiques.

Sur le plan scientifique, l'enjeu est celui de la place de la linguistique relativement à d'autres disciplines. L'évolution de la société est à l'origine d'une nouvelle dynamique, à laquelle les différentes disciplines scientifiques doivent s'adapter. Cette adaptation est particulièrement nécessaire dans le champ de l'information et de la communication. Les questions de veille, notamment de veille technologique, mais aussi de veille documentaire ou informationnelle, prennent de plus en plus de poids et font l'objet d'investigations dans divers champs disciplinaires. La linguistique se doit de suivre le mouvement global en accompagnant cette évolution de la société.

Dans une perspective non pas interdisciplinaire mais interne à une discipline, un enjeu complémentaire face à cette nouvelle dynamique est celui pour un corps de métier, celui des lexicographes. L'accélération des échanges et la multiplication des données à travers le foisonnement de la Toile amènent les lexicographes à gérer de plus et plus rapidement des données de plus en plus nombreuses. Pour faire face à cette nouvelle configuration, il semble indispensable de mettre en place de nouveaux outils et techniques capables de guider le lexicographe.

L'enjeu pour la société est de mettre en place un accompagnement de qualité parallèle à l'évolution de la langue. Plus précisément, la langue évolue, elle fait l'objet de ruptures sémantiques qui transgressent les usages décrits par des ressources de référence. Ces changements sémantiques peuvent se diffuser dans une communauté plus ou moins large. Une analyse suivie et outillée de ces changements permet d'atténuer l'imprécision de leur statut et de leur définition. Elle offre aussi un regard critique sur leur pertinence. Deux positions sont possibles pour cadrer l'évolution de la langue : la prescription ou l'intégration. Le choix de la prescription est triplement contestable :

- la prescription privilégie une aristocratie plutôt qu'une démocratie du savoir : un petit nombre d'experts impose plutôt que de s'adapter à un savoir présent dans une communauté ;
- en imposant une norme, elle impose une rupture avec d'autres ressources plus à même d'intégrer la réalité du terrain. Celles-ci se développeront en parallèle, sans dialogue avec les ressources officielles, au risque de ne pas s'appuyer sur des fondements scientifiques solides ;
- une science qui cherche à s'imposer par la prescription fragilise aussi sa place dans la société en tant que valeur. Une science qui ne sait pas s'adapter et reconnaître l'existant s'exclut elle-même des valeurs de la société.

À l'inverse, le choix de l'intégration reflète à la fois une considération pour la dynamique impulsée par les usages et un respect de l'existant. En effet, ce qui est partagé mérite d'être décrit. Mettre à la disposition d'une communauté plus vaste ce qui est déjà partagé par une partie de cette communauté revient à démocratiser un savoir. De plus, obtenir une

représentation de nouveaux sens permet de dépasser le flou de l'intuition à travers une description unifiée et rigoureuse. Une approche qui intègre les usages donnera donc une stabilité dans le respect des mouvements amorcés, voire déjà implantés dans la société. Elle contribuera à la cohésion sémantique plus qu'une approche prescriptive, parfois trop éloignée de la réalité du terrain pour être adoptée.

Sur le plan économique, l'enjeu est celui de la maîtrise de l'information. La société actuelle est une société de l'information et de la communication, où la fouille de données et la recherche d'information sont des problématiques d'intérêt croissant pour les entreprises. Étudier des variations sémantiques, brutales puis, à terme, plus nuancées, c'est chercher à construire une information de qualité, où le sens est décrit de façon pertinente, à la fois précise et à jour. Cela revient à participer à l'efficacité du traitement de l'information.

Si les enjeux sont multiples et invitent à attendre beaucoup d'une construction automatique du sens, il convient de préciser la portée à laquelle prétend un traitement automatique.

Une première mise en garde s'impose : un traitement ne fournira pas d'interprétation automatique, mais il génèrera une représentation du sens qui restera soumise à l'interprétation. Cette représentation ne permettra pas d'accéder à de l'extralinguistique et ne se voudra pas référentielle, mais différentielle, autrement dit, elle cherchera à mettre à jour des éléments susceptibles de distinguer le nouveau sens d'un mot de ses anciens sens et de le positionner relativement à d'autres mots.

Par ailleurs, on cherchera à obtenir un équilibre entre validité et robustesse. Les représentations ciblées se voudront d'une part en accord avec des principes théoriques modélisables et concordants avec des phénomènes constatés, d'autre part elles seront ancrées dans la réalité du terrain, c'est-à-dire qu'elles reposeront sur des applications à des données discursives ou textuelles.

Enfin, les représentations obtenues seront tributaires des définitions adoptées pour caractériser aussi bien les idées de rupture que de nouveauté ou de connaissances. Pour construire un nouveau sens, il est impératif de définir au préalable ce qu'est l'ancien sens, donc de s'appuyer sur des définitions de référence, puis il sera nécessaire de préciser ce qu'est une rupture relativement à cet espace de départ. À défaut d'obtenir des représentations adaptées aux connaissances et ressentis subjectifs de chacun, on cherchera à établir des représentations susceptibles de correspondre à une certaine communauté.

Partie I.

Un modèle théorique pour l'allocation de signifié

Chapitre I.1
Variations sémantiques : des interactions entre sens et contexte

Lorsqu'une unité lexicale est employée dans un discours, son sens se construit par interaction entre son sens littéral et les contraintes sémantiques exercées par le contexte : l'unité lexicale apporte un certain contenu sémantique, celui-ci est élagué, modulé ou éventuellement enrichi en fonction du contexte dans lequel il s'insère. Avant d'étudier la façon dont s'articulent les deux sources de construction du sens, la source littérale et la source contextuelle, il convient de préciser ce qu'on entend par 'sens littéral' et 'contexte'.

1. Les deux sources de construction du sens en contexte

1.1 Sens littéral : préalable à un emploi donné, mais issu d'un ensemble d'emplois antérieurs

Le sens littéral correspond approximativement au contenu sémantique associé à l'unité lexicale préalablement à son occurrence en discours, qui constitue le contexte de cette occurrence. Cette description représente le sens littéral comme un amont de la réalisation discursive, c'est-à-dire comme s'il précédait le sens contextuel. Certes, par rapport à une actualisation donnée, un sens littéral préexiste, mais celui-ci n'est pas un point de départ absolu. Il s'est lui-même constitué à partir d'une série d'usages, comme l'ont souligné de nombreux auteurs depuis Saussure :

> « The Aristotelian idea that words correspond to specific objects and concepts was displaced in the 20th century by the ideas of Saussure and others (Meillet, 1926; Hjemslev, 1953; Martinet, 1966; etc.). For Antoine Meillet: *The sense of a word is defined only by the average of its linguistic uses »* (Véronis, 1998:23).

Le sens littéral n'est pas la somme de tous les emplois rencontrés, mais le fruit d'une synthèse d'emplois, dont sont extraites, ou plus exactement abstraites, des tendances principales.

Le sens littéral n'est pas défini de façon intemporelle : il est propre à un moment, il peut rester inchangé un certain temps, mais il est aussi susceptible de se mettre à jour si les emplois évoluent et si un nouvel emploi s'impose massivement.

Le sens littéral peut se concevoir différemment selon qu'on se place du point de vue d'un individu, du point de vue d'une communauté ou du point de vue des ressources qui représentent le savoir partagé. Pour un individu ou un groupe d'individus, le sens littéral est un sens mémorisé, relevant de processus et de structures cognitives que nous ne détaillerons pas. Nous aborderons le sens littéral en tant que représentation d'un savoir partagé codé dans

une ressource, qui correspond au sens décrit par les lexicographes, c'est-à-dire aux entrées de dictionnaire.

Le sens codé dans un dictionnaire peut avoir une portée descriptive plus ou moins large, selon qu'il s'agit d'un dictionnaire de langue ou d'un dictionnaire encyclopédique. Il est le reflet d'un esprit propre à la ressource considérée (Imbs, 1971:XLVI). En particulier, un certain nombre d'approches en lexicographie moderne sont basées sur l'exploitation de corpus et cherchent à décrire les usages. De ce fait, elles sont destinées à expliquer les emplois principaux, fortement représentés. Un dictionnaire ne décrit pas tous les emplois passés :

> « Le dictionnaire ne peut échapper à un certain arbitraire : l'objectif est souvent incertain, par hésitation entre l'impossible exhaustivité et les limites matérielles et pratiques » (Dubois *et al*., 1994:146)
>
> « Ce que les dictionnaires ont en commun est d'abord un trait négatif : ils réduisent au minimum l'information sur les contenus » (Imbs, 1971:XI), à propos des dictionnaires de langue.

Ainsi, les descriptions lexicographiques donnent une représentation épurée des usages, simplificatrice mais aussi clarificatrice par rapport à la complexité et au flou du terrain. Par exemple, la subdivision en paragraphes pour chaque définition ou sous-définition reflète une hiérarchie marquée, mais qui n'a pas de caractère d'évidence. Dans une étude menée par (Kilgarriff, 1997)[1], la répartition en acceptions apparaît comme une tâche jugée parmi les plus difficiles par les lexicographes. Cet aspect doit être pris en compte lors d'un retour au discours : il faut savoir relativiser le sens littéral tel qu'il est décrit et assouplir sa structure pour s'adapter à un emploi donné.

1.2 Le contexte, source de variation

L'influence du contexte sur le sens peut être de nature linguistique ou non linguistique. Une influence linguistique est observable, puisqu'elle se manifeste directement au niveau des données. Une influence non linguistique est plus difficile à saisir, mais elle reste partiellement observable car elle est susceptible de laisser des traces dans les données linguistiques. Trois types d'influence du contexte sur le sens seront abordés, qui nous paraissent fondamentaux dans le cadre de ce travail : une influence provenant d'un environnement proprement linguistique, le *cotexte* ; une influence de nature extralinguistique, qui concerne le niveau référentiel ; une influence à la croisée du linguistique et de l'extralinguistique, qui englobe des phénomènes liés aux pratiques énonciatives ou à la structuration des discours en genres, et qui est conditionnée par des facteurs sociolinguistiques.

1.2.1 Le cotexte, ou contexte linguistique

Le sens d'une unité lexicale est déterminé en partie par son environnement linguistique, c'est-à-dire par le cotexte. Cet environnement linguistique peut être défini de façon plus ou moins large, c'est-à-dire selon des portions de discours plus ou moins étendues, et s'analyser à travers différents paliers. Tous les paliers de la textualité influent sur le sens d'une unité lexicale, cependant, ils ne constituent pas un tout homogène. Comme nous le détaillerons dans les paragraphes qui suivent, le cotexte affecte le sens d'une unité lexicale à travers des mécanismes distincts selon que sa portée est globale ou locale. En effet, les unités qui influencent le sens de l'unité lexicale considérée relèvent de différents niveaux d'observation.

[1] Dans cette étude, les lexicographes affectés à la rédaction de la troisième édition du *Longman Dictionary of Contemporary English* devaient trier par ordre de difficulté une liste répertoriant 13 tâches lexicographiques. Parmi les 11 réponses obtenues, la subdivision en sens apparaissait comme la deuxième tâche la plus difficile, après le fait de trouver une formulation juste.

Par exemple, dans un article de *Libération* du 5 novembre 2010 intitulé "Café grand-père", différents niveaux permettent d'accéder au sens du substantif *café*, selon l'occurrence considérée :

- Dans le syntagme

  « Jean-Pierre Blanc, directeur des ***cafés*** Malongo »

  les restrictions de sélection imposées par *directeur* permettent d'interpréter *cafés* comme chaîne industrielle ou entreprise.

- Dans la phrase

  « La coopérative, au fonctionnement démocratique et participatif, centralise le ***café***, organise la vente directe, utilise une partie des bénéfices pour des programmes sociaux ou éducatifs. »

  la dépendance syntaxique au syntagme verbal *la coopérative centralise* et l'état implicite de complément du nom de *vente directe* permettent d'affecter à *café* le sens de produit alimentaire.

- Dans le paragraphe suivant, la phrase décontextualisée

  « Et les géants du ***café*** peu soucieux du social exploiter le filon de l'équitable »

  ne permet pas d'interpréter *café* ; en revanche, le paragraphe où *café* apparaît permet de l'interpréter par métonymie comme industrie du café :

  « Vingt ans après, Francisco Van der Hoff refuse de parler de succès, malgré une présence de Max Havelaar dans 80 pays, 3,4 milliards de chiffre d'affaires en 2009 et 1,5 million de familles de petits producteurs qui en bénéficient. D'abord parce que les pauvres sont toujours pauvres, que 3 dollars par jour pour vivre, c'est mieux mais encore insuffisant, que *«la misère reste la misère »*. Mais aussi parce qu'il s'inquiète que le commerce équitable, en se diffusant trop, ne perde son âme. Il voit les *«multinationales de l'alimentation »* tourner autour des petits producteurs, en tirant les standards vers le bas. **Et les géants du café peu soucieux du social exploiter le filon de l'équitable**. (…) *»*

- A fortiori, le palier du texte apporte un éclairage supplémentaire au sens de *café*, en particulier pour le titre « Café grand-père » : l'article est centré sur le fondateur de Max Havelaar, la référence du commerce équitable, et il joue en permanence sur l'activation de l'unité de sens /équitable/.

De plus, la linéarité textuelle n'intervient pas de la même façon selon les paliers d'observation. Par-delà les différences selon le degré de localité, il existe des interactions et des dépendances entre les différents niveaux.

a- Le cotexte local

a1) Plusieurs unités pour délimiter le local

L'environnement local correspond au voisinage proche de l'unité lexicale ciblée. La définition précise de ce qu'on entend par "voisinage proche" varie selon la perspective adoptée. Le voisinage proche peut se délimiter sur les plans lexicographique, syntaxique, énonciatif ou typographique, dont chaque unité propre est susceptible de correspondre à une unité sémantique.

L'unité lexicographique correspond au palier minimal de localité. Elle correspond à la lexie, qui peut être simple ou composée. La définition de la lexie composée repose sur le modèle syntagmatique. Une lexie complexe, ou phrasème, se définira comme un syntagme contraint, dont le sens peut être compositionnel, comme dans *sac à main* ou *pompe à essence*, ou, dans

le cas des locutions, non compositionnel (Mel'čuk 2008), par exemple pour *bec de lièvre*, malformation labiale chez l'être humain, ou pour la *rose des sables*, qui renvoie à deux entités qui n'ont rien d'une plante, à la fois un minéral et un dessert chocolaté dont la forme évoque celle du minéral mais sans lien avec le sable, pas même au niveau de la texture ou de la consistance.

Les unités syntaxiques définissent un palier de localité supérieure. Elles correspondent aux propositions ou, à un palier quelque peu élargi, aux phrases. Elles sont définies par des règles syntaxiques de dépendance entre unités. Ces unités, largement utilisées en TAL, présentent plusieurs limites. En effet, les phrases ne sont des unités sémantiques que si elles sont complètes, c'est-à-dire grammaticalement correctes et sémantiquement interprétables. Par exemple, l'extrait de dialogue « Pour la sixième fois. » n'est pas interprétable sans la question qui le précède :

> « On vous a refusé un manuscrit, n'est-ce pas ?
>
> - Pour la sixième fois. »
>
> (Daniel Pennac, *La petite marchande de prose*, 1989:20)

De plus, un traitement automatique se heurtera souvent à des difficultés pour peu que la phrase ne soit pas bien construite, c'est-à-dire pour peu qu'elle ne soit pas grammaticalement correcte. Ce type de problème est particulièrement marqué à l'oral et n'aura pas la même ampleur dans notre cadre d'étude, celui de l'écrit. D'autres difficultés demeurent cependant, telles que le problème des anaphores ou encore l'ambiguïté de la phrase hors contexte énonciatif (Dubois, 1969)[2].

Les unités énonciatives favorisent une cohérence discursive, elles assouplissent les contraintes imposées par la limite de la phrase. Ainsi, l'unité sémantique peut être un regroupement de propositions ou de phrases, par exemple lorsqu'il y a présence d'anaphores. Les unités énonciatives peuvent notamment se définir par la présence d'unités de sens récurrentes localement, les isotopies locales (Rastier, 2006), ou par des groupements stables d'unités de sens dans une zone textuelle donnée, qu'on qualifie de formes sémantiques. Elles peuvent aussi être appréhendées comme des zones d'homogénéité thématique et donner lieu à des regroupements de phrases ou propositions thématiquement proches (Misra et Yvon, 2010). De telles unités se situent à cheval entre globalité et localité.

Les unités typographiques sont définies par des marqueurs tels que les espaces ou les marqueurs de fin de phrase (points, points d'exclamation, d'interrogation, de suspension). Pour le palier supérieur, les retours chariot ou les marqueurs de fin de paragraphe servent de délimiteurs, et au-delà, retours chariot, fin de paragraphe. Les critères typographiques sont souvent insuffisants pour garantir des unités sémantiquement pertinentes, et ils sont régulièrement couplés à des règles syntaxiques, lexicales, etc.

D'autres unités sémantiques se dessinent actuellement, qui remettent en question l'hégémonie de la phrase. (Mellet et Barthélémy, 2009) proposent de voir les textes comme une structure ordonnée, linéairement mais aussi par-delà cette linéarité, de telle sorte qu'elle définit une topologie textuelle. L'unité proposée est le voisinage, dont la délimitation reste soumise à questions, mais qui se caractérise par une très forte cohérence sémantique. Ces unités, objets d'études actuels, présentent un réel intérêt, même si elles ne seront pas utilisées dans le cadre de ce travail.

[2] L'auteur affirme que "toute phrase est nécessairement ambiguë" : le contexte énonciatif est nécessaire pour lever l'ambiguïté sémantique ressentie par le récepteur ; la désambiguïsation se produit en chaîne : l'ambiguïté d'une phrase est levée par la phrase suivante, qui elle-même introduit une nouvelle ambiguïté.

En bref, la délimitation précise du cotexte local dépend de ce qu'on privilégie, elle peut être plus ou moins étendue et privilégier différentes formes de cohérence sémantique.

a2) Descripteurs utilisés : recours à la syntaxe, localisation des effets sémantiques sur les unités lexicales

Le cotexte local met en jeu des liens relativement directs entre unités lexicales. Le sens d'un élément dépend de la séquence dans laquelle il s'insère, autrement dit, il est régi par ses relations avec l'enchaînement d'unités lexicales, ou encore par ses liens avec la linéarité textuelle. De par la localité du cotexte, la portion de texte considérée est réduite, donc il est plus facile de repérer où se situent les unités lexicales qui vont influer sur le sens de l'élément considéré. L'organisation des liens entre cet élément et ceux du cotexte est régie notamment par la structure syntaxique, en particulier par des structures prédicatives. Les contraintes exercées par ces structures proviennent de dépendances entre l'unité lexicale ciblée et d'autres unités lexicales, autrement dit, les contraintes cotextuelles sont localisables en surface car elles ont une position déterminée et caractérisée dans la chaîne syntagmatique.

Par ailleurs, l'existence de dépendances syntaxiques favorise l'étude de la variation sémantique à travers des problématiques qui mettent en œuvre des liens sémantico-syntaxiques. C'est en particulier le cas pour des variations sémantiques dues à des phénomènes de sous-catégorisation. Par exemple, dans le titre d'article

« Pékin ne veut pas d'un Nobel de la paix chinois » (Libération, 4 octobre 2010)

le nom propre Pékin est soumis à une variation sémantique par rapport à un sens toponymique : il est employé par métonymie pour désigner le gouvernement chinois. Les dépendances syntagmatiques permettent d'identifier de quelle unité lexicale provient la variation du nom *Pékin* : le verbe *vouloir* impose à son sujet la restriction d'être animé.

D'autres descripteurs, issus du plan infra-lexical, c'est-à-dire en deçà de l'unité lexicale, peuvent être utilisés pour observer et représenter les effets du cotexte. Un certain nombre d'entre eux restent liés au plan lexical. Ainsi, dans le *corbeau blanc* de (Rastier et *al.*, 1994), la variation sémantique du nom *corbeau* est décrite en termes de traits sémantiques, par inhibition du trait inhérent /noir/ ; sur le plan lexical, l'origine de la variation s'identifie clairement : la contrainte provient de l'épithète *blanc*. De même, la variation sémantique de *lune* dans *pierre de lune*, nom d'une pierre gemme, se caractérise par une inhibition d'un certain nombre de traits sémantiques, dont le trait /astre/, pour ne garder que le trait /irisé/. Pour une localité non pas réduite, mais plus étendue, il est possible de se détacher de l'ancrage dans le plan lexical et des dépendances syntaxiques, pour ne rester que sur des interactions avec le plan infra-lexical. Ainsi, la présence d'unités de sens récurrentes, qui correspond au phénomène d'isotopie locale, peut conditionner le sens d'une unité lexicale. Cette source d'influence se manifeste de façon plus diffuse, plus latente, elle n'intervient ni à travers une localisation précise sur le plan lexical, ni à travers une dépendance syntaxique spécifique, mais à travers des cooccurrences infra-lexicales qui ne reposent pas nécessairement sur des dépendances telles que les relations régies par la syntaxe. Ce type d'interaction se manifeste dès lors qu'on progresse vers un cotexte global.

En résumé, le cotexte local favorise principalement le plan lexical et les dépendances syntaxiques, même si des descripteurs infra-lexicaux peuvent s'ajouter aux descripteurs lexicaux et syntaxiques. Une description complète de l'impact du cotexte local sur le sens d'une unité lexicale nécessiterait de modéliser en détail les relations qui structurent la linéarité textuelle, en exploitant notamment la syntaxe. Dans notre cadre, une telle précision ne sera pas nécessaire : on cherchera à observer un phénomène récurrent et à en dégager des

tendances distinctives, non à obtenir une représentation fine du sens pour chaque occurrence rencontrée. On fera l'hypothèse que l'existence de redondance entre différents niveaux de description (Brunet, 2006) sera à même de fournir les caractéristiques principales recherchées.

b- Le cotexte global

b1) Paliers de la globalité

Les informations linguistiques globales relèvent de paliers supérieurs, tels que le texte ou le corpus de texte.

Chaque palier peut se concevoir comme un tout pour le palier qui lui est immédiatement inférieur : le corpus de textes est un tout pour tout texte qui le compose, le texte est un tout pour la période (l'approximation de la période sera le paragraphe dans notre approche), qui elle-même sera un tout pour l'unité lexicale. L'influence d'un palier sur un autre pourra se transmettre par paliers décroissants successifs : le corpus influe sur le texte, qui lui-même influe sur la période, qui influe sur l'unité lexicale, donc le corpus influe sur le sens de l'unité lexicale[3]. Dans le paragraphe suivant, tiré d'un corpus sur la crise financière, l'idée de crise et de contexte économique conditionne l'interprétation du paragraphe suivant et renforce ainsi l'interprétation de *produits toxiques* comme des instruments financiers tels que des titres et créances, non comme des substances chimiques :

> « La troisième période, celle du passage par les hedges funds et les banques d'affaires, voit la crise du subprime s'amplifier et éclater lorsque les produits toxiques sont diffusés mondialement. » (extrait d'un article intitulé *Ne pas occulter la dimension criminelle de la crise financière*, journal *Le Figaro*, 16/12/2008)

Les paliers du global influent sur ceux du local avec une portée variable, qui s'étend de l'élément isolé à quelque chose de plus systémique : ils agissent sur un élément donné, c'est-à-dire que le global affecte le local pour une occurrence donnée, mais aussi sur l'ensemble des occurrences de l'unité représentative d'un palier inférieur, et, plus largement, sur les systèmes constitués par l'ensemble des unités caractéristiques des paliers inférieurs. Par exemple, le genre ou la thématique agissent sur le lexique. Dans le prolongement de l'exemple précédent, les études sur corpus que nous avons menées (*cf.* chapitre III.1) montrent que le thème de la crise financière influe sur l'interprétation d'un nombre non négligeable d'occurrences de *toxique*, de façon répétée dans la durée, ce qui nous invitera à voir une évolution lexicalisable du sens de *toxique*.

b2) Descripteurs utilisés : granularité large ; recours au niveau infra-lexical

La détermination du sens d'une unité lexicale ne repose plus sur la linéarité textuelle, mais elle dépend d'informations caractéristiques de l'unité textuelle considérée, c'est-à-dire du texte ou du corpus de textes. En effet, au-delà de la période, l'unité lexicale ciblée n'est dépendante des autres unités lexicales ni directement, puisque celles-ci n'appartiennent plus à la même unité syntaxique, ni indirectement, puisque les éléments en relation avec la lexie par anaphore ou coréférence sont inclus dans la période (qui est le palier supérieur de cotexte local). Cette rupture avec la linéarité textuelle est au demeurant le critère sur lequel repose implicitement la définition de la séparation entre cotexte global et cotexte local : le global commence là où s'efface la dépendance syntaxique, directe ou indirecte.

Les relations en jeu ne reposent donc plus sur la syntaxe. Or c'est à travers ces dépendances que peut s'effectuer la localisation des contraintes sémantiques sur le plan lexical. Les

[3] Notons que l'influence d'un palier sur un palier inférieur peut être plus directe : il n'y a pas systématiquement une transition par les paliers intermédiaires, le corpus peut directement influer sur le sens de l'unité lexicale.

contraintes exercées sur le sens d'une lexie proviennent donc d'ailleurs, soit d'un autre type de relation, soit d'un autre niveau.

Sur le plan lexical, le lien avec le cotexte global peut être envisagé à travers certaines unités lexicales privilégiées. Ces unités peuvent se distinguer par leur fréquence, leur distribution ou encore par une position particulière dans les textes, comme dans un titre d'article en discours journalistique.

Le plan lexical peut fournir des éléments éclairants, mais il semble plus pertinent d'aborder l'impact du cotexte global à partir d'un niveau infra-lexical, en s'appuyant sur des informations de type mésosémantique ou macrosémantique, c'est-à-dire des caractéristiques sémantiques de paliers supérieurs au syntagme. Il s'agit ici d'unités de sens qui caractérisent un paragraphe, un texte ou un corpus de texte. En effet, un contenu sémantique majeur peut être présent de façon sous-jacente tout au long de ce texte sans apparaître explicitement sur le plan lexical. C'est par exemple le cas de l'idée d'inceste, qui sous-tend la pièce de théâtre *Agatha* de Marguerite Duras mais n'est jamais explicitée lexicalement, ou, de même, celle de l'ennui caractéristique de *Madame Bovary* de Flaubert. Ce contenu sémantique sous-jacent peut correspondre à des unités de sens récurrentes à l'échelle du texte, c'est-à-dire à des isotopies globales, et relever d'un niveau de granularité grossière, comme c'est le cas des domaines. Ce type d'information sémantique est une marque de l'identité du texte, dont la présence n'est pas nécessairement directe, mais toujours sous-jacente et propre à influer sur le sens de toute unité lexicale.

c- Le global détermine le local…

Le cotexte global et le cotexte local ne sont pas indépendants. L'approche dominante dans la lignée logico-grammaticale privilégie la compositionnalité du sens, autrement dit, une détermination du global par le local. À l'inverse, la tradition rhétorique-herméneutique a renversé la perspective et elle a mis en avant le principe de détermination du local par le global. Nous proposons une position intermédiaire, qui privilégie le global sans pour autant négliger l'apport du local.

Le global joue un rôle fondamental dans la détermination du local. Cette position se retrouve dans plusieurs courants théoriques, notamment la Gestalttheorie ou encore la sémantique interprétative de (Rastier, 1987). Dans une perspective gestaltiste, le niveau global se caractérisera par un tout signifiant, construit à partir d'une interprétation qui ne s'explique pas par compositionnalité. Ce tout se distingue par des saillances qui peuvent correspondre à des lignes structurantes ou à des motifs plus locaux, dont l'effet est similaire à celui, visuel, des doubles images telles que le vase de Rubin, qui sont interprétées par identification de fonds et de formes graphiques. Par analogie, la sémantique interprétative propose de voir des fonds et des formes sémantiques dans ces saillances textuelles, qui correspondent respectivement à des récurrences de traits sémantiques et à des regroupements de traits sémantiques. Si des éléments singuliers s'ajoutent localement ou sont observés dans le détail, ils s'interprèteront relativement à la forme d'ensemble qui se dessine. Par analogie, considérons un tableau impressionniste. L'observation d'un détail indépendamment de l'image d'ensemble ne sera rien d'autre qu'un aplat de couleur. Son interprétation ne sera possible et pertinente que si elle est effectuée non pas isolément, mais relativement à l'image d'ensemble. On verra ainsi émerger la feuille d'un arbre, une vague, le reflet du soleil sur l'eau, etc. Le principe est le même en sémantique : les informations sémantiques caractéristiques du niveau global détermineront celles du niveau local, palier par palier : le texte détermine la période, qui elle-même détermine le syntagme, qui lui-même influe sur la lexie. À titre d'exemple, considérons la phrase suivante :

> « Sauf qu'entre-temps les banques américaines ont largement contaminé en "produits toxiques" l'ensemble du système économique mondial. » (article de *L'Humanité* intitulé "Une crise peut en cacher une autre", 8 novembre 2008)

Le cotexte global ancre la phrase dans un contexte de crise financière, ce qui conditionne l'interprétation de "produits toxiques" comme des produits financiers néfastes pour le système. Décontextualisé, le syntagme "produits toxiques" aurait eu tendance à orienter l'interprétation vers le domaine de la chimie et des questions de pollution ; sa cooccurrence avec "contaminé" aurait pu créer l'ambiguïté. Or, de fait, l'interprétation ne se heurte pas à une ambiguïté, car les informations globales priment, elles situent l'interprétation de la phrase dans le domaine de la finance, et au sein de la phrase, l'interprétation de "produits toxiques" s'ajuste pour rétablir l'ellipse de l'adjectif "financier" dans le syntagme.

d- … mais les interactions entre global et local sont plus complexes

Le global a un impact sur le local, mais il se construit lui-même à partir du local. (Pincemin, 1999) propose une analogie éclairante. Partant de leur parenté étymologique, elle compare le texte à un tissu :

> « le tissu joue des effets de motifs et de texture : ce sont des effets globaux bien que créés par des contributions locales (insignifiantes à elles seules) (…) [de même] l'interprétation du texte se nourrit d'informations locales et globales, simultanément : on ne peut avoir une compréhension juste des unes sans une connaissance des autres, et réciproquement » (Pincemin, 1999:164)

Les contraintes sémantiques exercées par le global conditionnent donc l'interprétation locale, mais l'information locale contribue aussi à réajuster ces contraintes sémantiques globales. Par exemple, dans l'*Albatros* de Baudelaire *Les Fleurs du Mal, 1861*), l'analogie du dernier verset entre l'oiseau de mer et le poète entraîne une réinterprétation des versets précédents, pour y voir un tableau métaphorique de la condition du poète, incompris parmi les Hommes.

Les interactions entre global et local se manifestent également de façon dynamique, par processus d'amorçage et de diffusion sémantiques (Cordier, 1996). Lorsqu'une unité lexicale occurre, elle joue un rôle d'amorce et préactive un réseau d'autres unités lexicales, qui lui sont sémantiquement associées. L'accès au sens de la cible, c'est-à-dire l'unité lexicale qui lui succèdera, sera d'autant plus rapide si cette unité appartient au réseau préactivé du lexique et, le cas échéant, renforcera la navigation dans le réseau en question. Ainsi, un ensemble d'amorces sémantiquement associées favorisera une contrainte dynamique globale. L'occurrence d'une cible hors réseau préactivé s'accompagnera d'un accès au sens plus difficile, mais elle contribuera à la préactivation d'un nouveau réseau, qui pourra éventuellement s'imposer si les unités consécutives s'y inscrivent, entraînant ainsi une réorientation du réseau préactivé, c'est-à-dire des contraintes sémantiques globales.

1.2.2 Le contexte non linguistique

Le contexte non linguistique recouvre un ensemble hétérogène au sein duquel on s'attardera sur la référence et sur le contexte sociolinguistique.

a- La référence : un positionnement par rapport au monde… non nécessaire ?

a1) Le sens est étroitement lié à la référence

La question du sens est indissociable de celle de la référence, c'est-à-dire du rapport du sens au monde. La sémantique peut donc se positionner selon une approche soit référentielle, soit non référentielle, dont relève la sémantique différentielle.

Selon (Kleiber, 1999), la sémantique ne peut faire l'économie de la référence et se cantonner à un terrain exclusivement linguistique. Il précise que le monde n'est ni une réalité qui correspond à un donné objectif, ni, à l'inverse, un construit intra-linguistique généré par le langage mais qu'il s'agit de quelque chose d'externe qui relève d'une réalité expériencée. Cette réalité expériencée fait l'objet d'une stabilité inter-subjective, elle peut donc être considérée, par approximation, comme objective.

Une stricte position différentielle serait contestable pour deux raisons : d'une part, parce qu'une des fonctions du langage est de parler du monde, donc, en sémantique, on ne peut nier l'existence d'un rapport au monde et l'exclure de la question du sens ; d'autre part, parce que la sémantique différentielle se construirait de façon négative, à travers des oppositions, et non de façon positive. Or ceci permet de savoir en quoi un sens se distingue d'un autre, mais cette description du sens reste incomplète.

L'approche différentielle défendue par (Rastier, 1987) invite à considérer les choses sous un autre angle : elle n'exclut pas la référence, mais elle conteste le rapport direct entre les expressions et le monde. Cette approche rétablit une médiation entre expression et monde extérieur, à travers l'existence de représentations. Le rapport entre expressions et représentations permet de traiter le sens en restant sur le plan linguistique, sans pour autant nier l'existence d'une référence.

> « Pour ce qui concerne à présent la référence, nous ne pouvons prendre en considération la référence directe qui relie sans médiation des expressions et des objets, car elle dénie de fait l'existence d'un niveau sémantique propre aux langues. […] La sémantique différentielle traite en premier lieu de la référence en décrivant les contraintes sémantiques sur les représentations. » (Rastier, 2001:110-111) [cité par (Duteil-Mougel, 2004)]

Les positions de Kleiber et de Rastier ne remettent donc pas en cause l'existence d'une réalité externe, avec laquelle le langage est en relation. La question est plutôt celle d'un accès au sens qui exige une transition par ce réel, ou qui peut se borner à l'espace des représentations du réel.

a2) Variations sémantiques, variations référentielles et traces du changement

Une variation sémantique peut correspondre à une variation de la référence, c'est-à-dire du lien entre un signe et la réalité extralinguistique à laquelle il renvoie. Or si le lien entre une expression et son référent change, ce changement affecte également le niveau intermédiaire, c'est-à-dire celui de la représentation sémantique.

Si une expression pointe vers quelque chose de différent dans la réalité, elle laisse donc des traces au niveau de la représentation sémantique correspondante. La question est alors de savoir quelle part d'information est perdue, et quelle part d'information subsiste au niveau des traces laissées.

Prenons l'exemple de *souris* en informatique : sur le plan linguistique, les contextes d'apparition de *souris,* de même que ses définitions lexicographiques, ne donnent généralement pas d'information sur les matériaux et les composants qui la constituent, sur son apparence, et en particulier sur sa forme qui évoque celle de l'animal de même nom. Ce qui apparaît à travers les données linguistiques, c'est ce qui est important en termes d'usage : le lien avec l'informatique, la fonction d'outil pour l'ordinateur, la mobilité ou encore la propriété de sélection d'objets virtuels. C'est à travers ces caractéristiques que la souris informatique se distinguera de la souris zoologique. Les informations saillantes sur le plan linguistique

permettent d'accéder à une partie de la réalité correspondante : elles donnent une approche fonctionnelle plus qu'ontologique, mais qui correspond aussi à une partie fondamentale des pratiques.

On peut donc faire l'hypothèse qu'une variation sémantique donnera des informations linguistiques partielles mais suffisantes par rapport aux variations associées au niveau de la réalité externe. Autrement dit, la communication reflètera ce qui est important dans la variation de la référence.

b- Sociolinguistique et variation linguistique : accessibilité partielle et indirecte

Les facteurs sociolinguistiques ont un impact fort sur la langue : ils sont à l'origine de variations linguistiques marquées entre communautés.

De telles variations linguistiques affectent tous les niveaux de la langue : la phonétique, la syntaxe, le lexique, ou encore la sémantique, qui est le niveau sur lequel nous nous focaliserons. L'étude des variations sémantiques ne peut se passer d'un positionnement par rapport à la sociolinguistique, pour laquelle la notion même de variations linguistiques est centrale.

Les variations sociolinguistiques peuvent se répartir selon quatre axes (Gadet, 2003:13-17) : variations diachroniques, diatopiques, diastratiques et diaphasiques, qui sont respectivement selon le temps, le lieu, la classe sociale et les situations discursives. En jouant sur les différents axes et en les combinant, on constate qu'il existe une très grande diversité de cas de figure, ce qui donne à penser que le sens d'un mot est susceptible d'être soumis à une infinité de variations.

Les variations sociolinguistiques ne sont pas anarchiques, elles sont structurées par l'existence de communautés, et de même, la variation sémantique est cadrée par cette structure. Certes, au sein de ces communautés, il peut exister des variations idiolectales, c'est-à-dire propre à des individus, mais le fait que la communication soit possible exige qu'il existe au sein de la communauté une stabilité sémantique globale, c'est-à-dire une stabilité du sens de l'essentiel des unités lexicales pour les différents individus composant la communauté.

L'étude des variations sémantiques nécessite de se positionner par rapport à l'existence de communautés définies socialement. La notion de variation sémantique ne sera pas la même selon qu'on se place au sein d'une communauté ou en dehors de cette communauté. Observer les variations sémantiques d'une communauté à une autre revient à mettre l'accent sur les différences sociales entre ces communautés et à privilégier une perspective sociolinguistique. Les observer au sein d'une communauté donnée revient à adopter une position linguistique, dans la mesure où les paramètres sociolinguistiques constituent le cadre, c'est-à-dire un invariant.

La perspective retenue dans le cadre de ce travail est linguistique, non sociolinguistique, donc centrée sur une communauté donnée. Il est nécessaire de préciser cette communauté. En effet, toute approche TAL ancrée dans la textualité se positionne, implicitement ou explicitement, relativement à une communauté dont les données sont représentatives. Elle se fait le reflet de spécificités discursives et se destine à un individu lambda dont il convient de préciser le profil socialement défini à travers les ressources utilisées.

Une parfaite adaptation des traitements à toute variation linguistique, ou même une parfaite coïncidence d'un traitement avec une variation linguistique donnée est utopique. Pour un certain nombre de partisans de la linguistique variationnelle, la langue standard est considérée comme un idéal jamais actualisé et approximatif par rapport aux différentes variations linguistiques. La même objection pourrait être formulée à l'égard de chaque variation

linguistique considérée séparément, car elle connaît des variations internes, ne serait-ce que d'un individu à l'autre de la communauté.

De ces réflexions, il faut retenir que tout traitement automatique doit être accompagné d'une réflexion sur le cadre sociolinguistique auquel il s'applique, c'est-à-dire sur la communauté dont il reflète les particularités et celle à laquelle il s'adresse. L'interprétation de résultats doit s'effectuer en ayant conscience du positionnement sociolinguistique choisi.

Dans le cadre de cette thèse, la langue ciblée est assez proche de ce qu'on pourrait qualifier de français standard. Plutôt que de considérer le français standard comme une langue idéale mais sans ancrage dans la réalité textuelle ou comme une norme, nous préférons adopter le point de vue de (Guerin, 2008), qui propose de voir le français standard comme une variété située. Celle-ci se caractérise par une distance physique ou conventionnelle entre l'émetteur et le récepteur, elle tend à neutraliser les spécificités saillantes d'autres variétés linguistiques.

En ce qui concerne le cadre sociolinguistique induit par les ressources utilisées, les sens littéraux sont issus de ressources lexicographiques élaborées pour des intellectuels littéraires, une communauté de personnes ayant une maîtrise solide de la langue française (Henry, 1996)[4]. Les ressources textuelles qui constitueront le cotexte seront issues d'un corpus de presse et définissent donc une communauté dont le profil est celui du lecteur de presse, qui pratique le français courant. On fera l'approximation que les variations linguistiques pratiquées par les deux communautés décrites sont suffisamment proches l'une de l'autre et on considèrera qu'elles peuvent être mises en relation sans qu'il y ait de biais majeur dû à leurs particularités sociales respectives. Les spécificités sociolinguistiques qui distinguent le corpus de presse de la ressource lexicographique ne pourront être complètement éliminées. Cependant, les efforts de ce travail seront orientés vers des variations sémantiques marquées (*cf.* chapitre I.2). On supposera que ces variations seront suffisamment tranchées pour ne pas être dues aux spécificités discursives du corpus textuel par rapport à la ressource lexicographique, mais qu'elles relèveront d'une évolution qui concerne à part égale les deux communautés représentées.

1.3 Rétroaction des emplois sur le sens littéral

Une description de sens littéral n'est pas permanente, elle est propre à un moment donné et peut évoluer dans le temps. Certains emplois récurrents, où l'apport contextuel est trop important par rapport au sens littéral, peuvent modifier celui-ci, autrement dit, il peut y avoir rétroaction. On reviendra sur ce point dans le chapitre suivant.

2. Le sens in vivo : variation par interaction entre sens littéral et contexte

On peut accorder une importance variable à l'apport respectif du sens littéral et des contraintes discursives dans la détermination du sens *in vivo,* c'est-à-dire du sens tel qu'il apparaît dans les emplois discursifs. Il est possible d'adopter des positions marquées, qui valorisent de façon préférentielle l'une ou l'autre des deux sources du sens en contexte. Ainsi, certaines approches donneront le primat au sens littéral, d'autres au contexte.

[4] « …dans le cas du TLF, [le] public est constitué par "l'homme cultivé moderne" [Imbs, 1971:XVII], tourné vers une culture de type humaniste dans un environnement matériel et intellectuel élargi, culture actualisée par l'intérêt nouveau porté aux sciences et aux techniques, sans pour autant que cet homme cultivé soit un spécialiste d'aucun domaine » (Henry, 1996).

2.1 Deux traditions en conflit : logico-grammaticale et rhétorique-herméneutique

Deux traditions de pensée reflètent ces deux tendances : la ***tradition logico-grammaticale*** et la ***tradition rhétorique-herméneutique***. Ces traditions sont anciennes. (Anscombre, 1998) les fait remonter à Aristote, dont les *Analytiques* seraient l'ancêtre du logico-grammatical, et les *Topiques* celui du rhétorique-herméneutique. Par la suite, la linguistique s'est construite en s'appuyant de façon privilégiée sur la tradition logico-grammaticale. Depuis quelques années, la tradition rhétorique-herméneutique est réapparue sur le devant de la scène linguistique, par exemple avec la sémantique textuelle.

La ***tradition logico-grammaticale*** privilégie le sens littéral et accorde un moindre rôle à l'apport contextuel. Elle favorise une perspective ontologique. Autrement dit, le sens est étroitement lié à l'existant, à quelque chose de donné. Cet existant est le support du sens littéral. Ce dernier est donc le reflet du réel, ou au moins d'un réel expériencé. Le sens littéral définit donc un ensemble fixe, qui recouvre les valeurs possibles associées à un sens, donc qui délimite un espace au sein duquel le sens discursif peut, et même doit se réaliser.

L'emploi d'une unité lexicale correspond à une navigation dans l'ensemble des possibles que constitue le sens littéral, donc à un positionnement dans un espace sémantique préexistant. Ce positionnement est progressif : au fur et à mesure que la syntaxe définit des contraintes linguistiques, certaines parties du sens littéral sont exclues, autrement dit l'espace des possibles se réduit et le sens se stabilise sur un état compatible avec les restrictions imposées par la syntaxe. Il y a donc une évolution couplée entre syntaxe et sémantique : en parallèle de la saturation syntaxique, le sens se précise et, donc, définit une zone de plus en plus localisée dans l'espace des possibles, jusqu'à définition d'une zone précise et clairement délimitée.

Ainsi, le sens littéral peut se voir comme un espace prédéterminé, fixe, sur lequel le contexte exerce des contraintes par le biais de la syntaxe. Le palier de la textualité par excellence est ici celui propre à la syntaxe, à savoir la proposition ou la phrase. Le sens se précise grâce à des règles logiques, donc particulièrement adaptées à des approches formelles.

Le contexte n'enrichit pas ou ne déforme pas le sens littéral, mais il le réduit à un sens précis. La variation sémantique peut prendre deux formes :

1) Cas de désambiguïsation sémantique : la variation sémantique peut être considérée comme une variation interne entre les différentes acceptions qu'autorise le sens littéral.
2) Le sens en contexte n'est pas compatible avec la division en acceptions du sens littéral. Cette situation est le plus souvent exclue de la plupart des approches en TAL, où seules sont considérées les phrases bien formées. En revanche, dans les modèles prédicatifs, qui décrivent le sens à partir de restrictions de sélection et à travers les phénomènes de sous-catégorisation, la variation sémantique se traduira par une violation des règles syntaxico-sémantiques.

À l'inverse, la ***tradition rhétorique-herméneutique*** met le contexte à l'honneur. Dans cette approche, le sens n'est pas considéré comme naturel, mais comme culturel (Rastier, 2008) : le sens est déterminé par les usages, il est fonction d'effets individuels ou sociaux. Par exemple, le sens culturel d'ADN n'est pas celui de son homologue ontologique, *l'acide* désoxyribonucléique : tandis *qu'acide désoxyribonucléique* reste ancré dans le domaine de la biologie et se rattache à des aspects moléculaires, *ADN* est chargé d'un sens culturel, rattaché à l'empreinte génétique, à l'identité et à l'identification, ainsi qu'au judiciaire et à l'idée d'expertise. Le seul sens ontologique ne suffit pas à caractériser le sens d'*ADN*. La dynamique

du sens prime sur la fixité, si bien que le sens littéral s'inscrit dans un processus : il est le produit de l'histoire d'une unité lexicale à un temps t.

L'emploi d'une unité lexicale correspond à une intégration dans un environnement discursif. Ce n'est donc plus l'ensemble des possibles défini par le sens littéral qui prime, mais le discours. Une unité lexicale s'insère dans un tout, celui du tissu textuel (Pincemin, 1999) mentionné précédemment (*cf.* 1.2.1), qui est constitué de fonds et de formes sémantiques, et elle participe d'un rythme sémantique (Missire, 2005)[5], c'est-à-dire d'un mouvement qui accompagne le déroulement du discours, qui n'est pas fragmenté mais structuré, cohérent et témoin d'une unité textuelle. La phrase n'est plus le palier de textualité par excellence, celui du texte la supplante, sans pour autant être un palier exclusif : son effet se conjugue à celui de tous les autres paliers de textualité, notamment les paliers supérieurs que constituent le paragraphe, le texte et le corpus de textes.

Le contexte conditionne fortement le sens des unités lexicales : ce sens est modulé, voire modifié en fonction de l'environnement discursif. Le contexte est considéré comme partie intégrante du sens, autrement dit, il ne se contente pas de préciser le sens, mais il contribue aussi à l'enrichir. À titre d'illustration, le sens de *violon* reçoit l'unité de sens /sentiment/ ou /émotion/ aussi bien dans le poème « Harmonie du Soir » des *Fleurs du Mal* de Baudelaire ("Le violon frémit comme un cœur qu'on afflige") que dans la chanson de Jean-Jacques Goldman *Tournent les violons*, mais il y a opposition dans la connotation émotionnelle plus spécifique dont le contexte enrichit le sens de *violon* : /tristesse/ prime chez Baudelaire, tandis que /joie/ et /fête/ s'imposent chez Goldman. La variation sémantique est donc omniprésente et la hiérarchie est inversée : le global détermine le local, le contexte joue un rôle de premier plan dans la détermination du sens d'une unité lexicale et ainsi, le sens est plus pragmatique, c'est-à-dire déterminé par les conditions d'usage et le contexte d'emploi, que spéculatif, c'est-à-dire prédéterminé par un sens intrinsèque, existant en amont de la production discursive.

Pour résumer, la tradition logico-grammaticale et la tradition rhétorique-herméneutique adoptent des positions distinctes, en particulier, concernant la construction du sens, sur les points suivants :

- La base sur laquelle le sens se construit est le sens littéral pour la tradition logico-grammaticale et le contexte pour la tradition rhétorique-herméneutique. Cependant, le fait de privilégier un pôle n'exclut pas totalement l'autre : dans les deux approches, sens littéral et contexte ont tous deux leur rôle à jouer dans la construction du sens.
- Le sens littéral semble relever d'origines opposées, naturelle en logico-grammatical et culturelle en rhétorique-herméneutique.
- La fonction du discours sur laquelle l'accent est mis n'est pas la même : la cognition prime en logico-grammatical, puisque le sens est ramené à de l'ontologique ; la communication prime en rhétorique-herméneutique, puisqu'il s'agit de considérer le sens d'une unité lexicale relativement au sens textuel, donc au message véhiculé par le discours.
- La variation sémantique émerge avec un profil différent selon l'approche : elle est plus cadrée dans l'approche logico-grammaticale car elle reste soumise à une combinatoire formelle (dont certaines versions sont représentées dans le cadre de la logique) et parce qu'elle est restreinte aux possibles prédéfinis par le sens littéral ; elle est plus fondamentale dans l'approche rhétorique-herméneutique, de par la variabilité des pratiques. Elle y est abordée plus souplement, car elle bénéficie du rattachement de la

[5] « P. Sauvanet propose de considérer comme *rythmique* "tout phénomène perçu, subi ou agi, auquel on peut attribuer un des trois critères suivants : structure, périodicité, mouvement" (Sauvanet, 2000:195) » (Missire, 2005)

sémantique des textes à de l'artistique et parce que le sens lexical peut s'enrichir du sens qui émerge globalement du contexte, par-delà l'enveloppe définie par le sens littéral.

- La qualification de la variation n'atteint pas le même stade. Dans l'approche logico-grammaticale, elle correspond à un déplacement dans l'ensemble des possibles, mais pour des variations marquées que l'ensemble des possibles ne suffit plus à expliquer, elle peut être qualifiée soit comme négation de l'existant, soit comme mise en relation de deux parties incompatibles, soit syntaxiquement s'il s'agit de préciser ce qui est nouveau. Dans l'approche rhétorique-herméneutique, la qualification de la variation est sémantique, même pour des variations marquées qui exigent de chercher de nouveaux éléments que ne contenait pas le sens littéral.

Un positionnement relativement à ces deux approches s'impose, positionnement que nous préciserons après analyse de quelques exemples.

2.2 Compromis entre sens littéral et apport contextuel

Comme le souligne (Recanati, 2007), les positions radicales qui accordent une influence quasi-exclusive sur le sens des mots à un des deux pôles, soit le sens littéral, soit le contexte, sont à exclure. Nous verrons que, selon les unités lexicales considérées et les emplois observés, le poids relatif à accorder au sens littéral et au contexte ainsi que leur mode d'interaction changent. Un type de modèle n'est pas systématiquement le plus adapté, il fait le choix d'une approche modérée et qui combine différents niveaux. Cela semble un compromis nécessaire pour faire face à la diversité des cas de figure.

2.2.1 Variabilité dans la perméabilité du sens littéral au contexte

Le sens littéral peut être plus ou moins perméable à l'influence du contexte. Autrement dit, il existe des contenus sémantiques faibles, suffisamment lâches pour absorber les contraintes sémantiques du contexte, et à l'inverse des contenus sémantiques forts, pleins, qui sont moins susceptibles de varier sous l'effet du contexte et qui, eux-mêmes, marquent le contexte et participent fortement du sens global.

Un cas extrême de perméabilité, certes artificiel mais très éclairant, est celui de l'unité lexicale *schtroumpf*. Création linguistique de l'auteur de bande dessinée Peyo et récurrente dans les discours de ses petits lutins bleus, elle est une unité passe-partout capable de se substituer à n'importe quel mot du vocabulaire. Cette unité présente une perméabilité totale au contexte. Dotée d'un sens littéral vierge, elle constitue son sens en emmagasinant l'ensemble des contraintes contextuelles. Son sens littéral peut se concevoir comme un cas de polysémie infinie, où aucune unité sémantique n'est présente pour délimiter l'espace des possibles.

Des cas similaires, moins artificiels, sont des unités lexicales telles que les substantifs *truc* ou *chose*. Ils sont pour ainsi dire sémantiquement vides, et bien qu'ils n'appartiennent pas à des catégories grammaticales caractéristiques qui servent généralement à définir les mots-outils, ils se rapprochent de tels mots. Ils contribuent peu au contenu sémantique global et leur sens se construit essentiellement à partir de l'apport du contexte.

À l'inverse, certaines unités sont sémantiquement pleines et leur sens littéral définit un champ de possibles fortement contraint. Les termes présentent un tel profil : ils renvoient à un contenu sémantique précis, spécialisé, que le contexte adaptera ou transformera plus difficilement, tandis que ce contenu lui-même se répercutera sur le sens global du cotexte d'occurrence. Ainsi, pour ce type d'unités lexicales, l'apport du sens littéral ne peut être négligé, il marque le contexte.

De façon générale, la perméabilité dépend de l'ensemble initial des possibles, elle est rattachée au degré de polysémie. Les unités lexicales que nous étudierons seront des unités sémantiquement pleines, c'est-à-dire dotées d'un contenu sémantique propre, qui ne soient pas totalement perméables au contexte sans pour autant être complètement fermées à son influence.

2.2.2 Type d'action du contexte : exemple d'apport multiniveau

La syntaxe et le cotexte global peuvent avoir une contribution relative variable dans la détermination du sens, ce qu'illustrent les exemples suivants. On s'appuiera sur deux artefacts pour illustrer notre propos. Le premier est l'emploi du mot *schtroumpf*. Loin d'être représentatif des unités lexicales, il a un caractère très particulier, à savoir une perméabilité au contexte poussée artificiellement à l'extrême, ce qui en fait un outil intéressant pour appréhender les mécanismes en jeu. Le deuxième artefact est un texte à trous.

Dans la phrase

> « Il boit quelques gorgées de schtroumpf après l'avoir laissé infuser dans sa tasse »

les contraintes syntaxiques imposées par *infuser* et *boire* suffisent à déterminer le sens de *schtroumpf*, indépendamment d'un contexte plus large : les restrictions de sélection imposent à *schtroumpf* d'être une boisson susceptible d'infuser, dont les représentants les plus prototypiques sont le thé ou la tisane.

À l'inverse, considérons le texte à trous suivant, tiré du site Internet de l'école élémentaire d'Orgeoise :

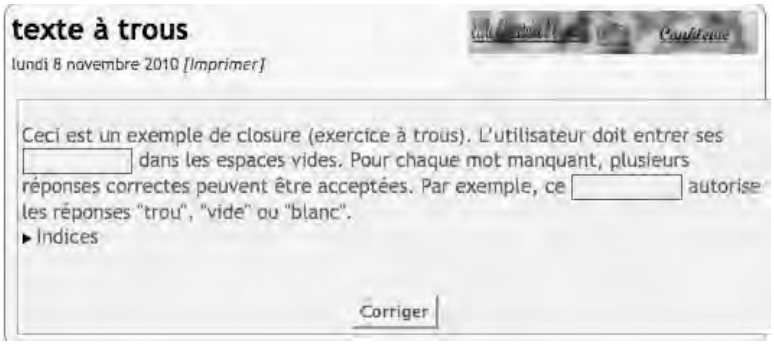


Figure I.1.1 Texte à trous et influence du contexte

Le palier de la phrase et les contraintes syntaxiques ne suffisent pas à préciser le sens du mot inconnu dans la phrase « L'utilisateur doit entrer ses (…) dans les espaces vides », car la phrase ne permet de déterminer si les *espaces vides* sont des espaces physiques ou des espaces virtuels. Il pourrait tout à fait s'agir d'une entité matérielle (jetons d'un jeu de Puissance 4, dosettes d'une machine à café, piles d'un appareil électronique). Le cotexte plus large permet de préciser qu'il s'agit d'une entité non matérielle, porteuse d'un contenu informatif.

Ces illustrations sur le degré de perméabilité et les niveaux d'impacts du contexte invitent à ne pas choisir de position extrême, où ne primerait que le sens littéral ou que l'apport du contexte, mais à chercher une position intermédiaire, qui laisse sa place à chacune des deux sources, un sens préexistant qui va au moins partiellement se réaliser et une transformation sous l'effet du contexte, et qui soit capable d'intégrer une certaine variabilité en fonction du cas considéré.

Pour définir un positionnement plus précis relativement aux deux traditions évoquées, on se place dans une perspective d'automatisation des processus avec deux objectifs :

1. faire face à la multiplicité des situations de variation sémantique par le simple fait de l'augmentation de la quantité de données numérisées et échangées ;
2. observer « in vivo » la tension entre ces deux extrêmes.

2.3 Décrire l'actualisation : représentations de précision variable de la variation sémantique

Plusieurs modèles peuvent servir à aborder la variation sémantique. Ils reposent sur des degrés de finesse variables. Trois grandes façons de procéder peuvent être dégagées, qui apparemment s'excluent, mais qu'on proposera d'articuler et d'exploiter successivement pour parvenir à une représentation de plus en plus fine.

2.3.1 Direction générale et basculement entre acceptions

La première façon de modéliser l'actualisation repose sur une représentation du sens littéral comme un ensemble discret de sens, les acceptions, qui correspondent à des états distincts et en nombre limité. Déterminer le sens en contexte revient à choisir une acception, c'est-à-dire à désambiguïser le sens par basculement entre différentes acceptions. Le contexte a donc, dans ce cadre, un rôle de sélection entre un petit nombre d'états définis par la structuration en acceptions.

La mise en œuvre d'un tel modèle nécessite de disposer d'une représentation appropriée. Les ressources lexicographiques présentent un profil adéquat : une entrée est divisée en définitions, chaque définition est supposée correspondre à une acception.

Cependant, plusieurs critiques peuvent être émises à l'égard de ce modèle et des répartitions en sens proposées par les ressources lexicographiques. Ainsi, l'existence et, partant de là, le bien-fondé d'une division en *des* sens, a été remise en question notamment par (Hanks, 2000). En effet, la superposition du sens en contexte avec un sens précis codé dans l'entrée correspondante est loin d'être systématiquement vérifiée : dans certains cas, seule une partie de la définition retenue est effectivement pertinente, dans d'autres, elle doit être complétée par des unités de sens issues d'autres définitions. De plus, les subdivisions des ressources lexicales sont imparfaites et elles ont un caractère non spontané. Ces imperfections sont perçues aussi bien au niveau de l'élaboration des ressources que lors de leur exploitation. Comme nous l'avons déjà signalé, (Kilgarriff, 1997) a fait ressortir que, lors de l'élaboration des ressources lexicographiques, la subdivision d'une entrée en définitions et sous-définitions fait partie des tâches les plus difficiles pour les lexicographes. Ces difficultés semblent témoigner d'une certaine artificialité et d'incertitudes sur le découpage de définitions. De même, au niveau de l'exploitation des ressources, l'appariement entre le sens dans un contexte donné et une définition issue d'une entrée de ressource lexicographique est loin d'être immédiat et consensuel. Une expérience de (Véronis, 2004) en témoigne : un ensemble d'annotateurs devait sélectionner le sens approprié à partir de définitions tirées du *Petit Larousse* ; les résultats obtenus présentaient des valeurs d'accord inter-annotateurs médiocres.

Or tout individu interprète en contexte et, pour permettre la communication, il semble légitime de supposer qu'il existe une stabilité intersubjective au niveau de l'interprétation. Les difficultés aussi bien du lexicographe que de l'interprétant pour associer les représentations en subdivisions de sens aux emplois mettent en évidence l'imperfection des représentations du sens littéral comme des sens.

Certes, le découpage en sens a ses limites, cependant, il n'est pas à exclure complètement. L'alternative qui consisterait à déstructurer totalement une entrée fait perdre des éléments d'information importants et nécessitent des efforts de restructuration qui ne sont pas nécessaires. Ainsi, dans le cas d'homonymes, ou pour des définitions associées à des domaines d'emplois particuliers, il semble pertinent de s'appuyer sur les divisions préexistantes dans les représentations lexicographiques. Autrement dit, pour des divisions marquées, qui présupposent un découpage de haut niveau, qui ne soit pas trop fin, il semble judicieux d'exploiter un basculement entre sens possibles, pour déterminer la direction principale de sens, et surtout pour effectuer un élagage grossier. Cette forme de désambiguïsation reste grossière, mais elle peut jouer un rôle indicatif, pour proposer une orientation générale.

2.3.2 Fluctuations autour de la direction générale

La deuxième façon d'aborder l'actualisation consiste à considérer le sens littéral comme un ensemble de potentiels de sens. Ceux-ci peuvent s'organiser selon une structure indicative, mais ils ne définissent pas un compartimentage rigide. En contexte, certains potentiels de sens se réalisent, ils sont actualisés de façon souple et modulée par rapport au champ de potentiels initial. Considérons l'exemple de *panier* dans les deux contextes qui suivent :

> « Emballé dans un torchon, posé avec un verre dans le **panier** noir à anses […], le précieux litre [de vin] faisait le tour des fermes proches », *Composition française, retour sur une enfance bretonne*, 2009:50, *La Bretagne incarnée,* Ozouf Mona.
>
> « Ajouter au panier », formule récurrente renvoyant aux paniers électroniques des sites commerciaux.

Les potentiels de sens de *panier* connaissent des actualisations différentes. Dans les deux cas, l'idée de contenant est activée, mais dans le premier exemple, le caractère matériel, tangible, ressort tandis que dans le second cas, le *panier* est dématérialisé et constitue une entité virtuelle. De même, la fonction actualisée de *panier* diffère : la fonction de transport s'ajoute à celle de contenant dans le premier cas ; elle est absente dans le cas du panier électronique, où seule compte la fonction de stockage. Le rôle du contexte est d'activer ou d'inhiber certains potentiels de sens, il joue sur les composantes du sens littéral de façon plus nuancée et moins contrainte que précédemment.

Le jeu sur des potentiels de sens n'exige pas de déterminer au préalable une direction principale, comme dans le cas précédent. Cependant, cette approche peut être exploitée de façon complémentaire à l'autre approche. En effet, elle peut se voir comme un affinement du sens obtenu précédemment par sélection d'acception, ou comme l'introduction de fluctuations autour de la direction principale, car elle module l'acception retenue en faisant ressortir certaines unités sémantiques parmi celles qui sont présentes, et elle y adjoint des nuances sémantiques provenant d'autres acceptions.

2.3.3 Précision de l'apport contextuel pour constituer un nouveau sens

Les deux modélisations de l'actualisation, à savoir la désambiguïsation grossière et l'actualisation modulée de potentiels de sens, reposent sur de la sélection et de la modulation appliquées au sens littéral, autrement dit, elles restructurent le sens littéral. Cependant, elles restent sur la même base qualitative et n'enrichissent pas le sens littéral en nuances de sens. La troisième façon d'aborder l'actualisation repose donc sur l'introduction de nouvelles unités de sens apportées par le contexte.

Cet apport sémantique du contexte est nécessaire. (Fuchs, 2008) l'illustre à travers l'exemple du verbe polysémique *passer*, dont l'interprétation dessine des sens divers et entrelacés selon

les emplois, qui ne peuvent se traduire sous forme d'une liste exhaustive et consensuelle. Elle invite alors à "récuser l'idée intuitive selon laquelle le contexte environnant (c'est-à-dire les autres éléments de l'énoncé) fonctionnerait à la manière d'un tamis qui 'filtrerait' le sens effectif du polysème, parmi tous ses sens potentiels [et à ne pas] voir dans le contexte une sorte de projecteur qui éclairerait certains traits pré-existants, ou 'activerait' certaines facettes de sens, aux dépens d'autres". Elle suggère au contraire de privilégier une approche gestaltiste, dans l'esprit de ce qui a été décrit en 1.2.1.c.

Partir d'une représentation du sens permet donc d'obtenir certaines unités pertinentes pour un emploi donné, mais cela reste insuffisant pour garantir d'avoir la description sémantique complète de cet emploi, quel que soit le jeu réalisé sur les unités de sens disponibles. À titre d'exemple, considérons le nom *généalogie* employé dans un article concernant l'existence de fichiers sur les minorités ethniques, dans le contexte de mesures politiques visant l'expulsion de Roms (« Un fichier bien caché stigmatise les Roms », *Libération*, 8 octobre 2010). Dans ce contexte, le sens de *généalogie* semble se charger des unités de sens /ethnie/, voire /politique raciale/ ou /déportation/, dont témoignent des indices textuels mais aussi intertextuels, en particulier l'existence d'un article paru une semaine plus tard dans le même quotidien, intitulé « Roms : Reding "regrette" la comparaison avec la déportation, Paris "prend acte" ».

Le modèle de la sémantique interprétative oriente vers une solution, puisqu'il propose une description du sens en unités sémantiques préexistantes, les sèmes inhérents, et en unités de sens provenant du cotexte, les sèmes afférents. Il propose donc une représentation nuancée, qui intègre l'influence du sens global. Il accorde une place centrale au contexte dans la construction du sens et lui donne ainsi un vrai statut en précisant son apport.

L'intérêt d'un modèle qui intègre l'apport contextuel est double : il revient à accepter les limites des ressources, en termes d'adéquation avec les processus interprétatifs ou de caractère insuffisamment à jour, et à essayer de pallier ces limites. De plus, il prend en compte la perspective rhétorique-herméneutique, notamment la détermination du local par le global.

Ce modèle peut intervenir en complément des précédents : une fois le sens littéral grossièrement élagué, puis modulé plus en finesse, il est enrichi et remodelé par le contexte.

3. Gradualité dans les variations : pourquoi privilégier les variations marquées ?

La variation sémantique n'est pas absolue, elle s'observe toujours relativement à un point de référence. N'importe quel point de référence peut théoriquement être utilisé, même s'il correspond à un sens très marginal en langue. Dans notre cadre, il sera défini par des ressources lexicographiques. Il sera considéré comme propre à refléter des usages conventionnels. Cet état n'est pas à proprement parler une norme, mais, pour fluidifier le propos, on se permettra d'adopter une telle approximation et de qualifier de normaux de tels états.

3.1 Existence d'une échelle allant de l'emploi normal à la rupture

Les variations sémantiques s'échelonnement progressivement de variations faibles à des variations marquées. L'existence d'une telle gradualité transparaît à plusieurs niveaux qui sont en interrelation : celui du ressenti, celui de l'interprétation et celui des données discursives, soit en termes de cohérence, soit en termes d'indices de surface.

3.1.1 Gradualité du ressenti

L'emploi d'une unité lexicale peut être ressenti comme évident, ou au contraire produire un effet particulier. Selon les emplois, la variation sémantique peut être plus ou moins marquée, elle est perçue selon une échelle qui s'étend de l'emploi normal à la rupture sémantique.

Un emploi dit normal ne suscite pas de difficulté interprétative. Le sens se construit à partir du sens littéral de façon relativement aisée, l'emploi associé fait partie des emplois types que la ressource cherche à décrire, et la structure codée du sens littéral permet d'aboutir facilement au sens ad hoc. Cet emploi ne crée pas d'effet particulier.

En revanche, les variations plus marquées s'accompagnent d'effets : ludiques, rhétoriques ou encore sensations d'erreur ou de décalage. Typiquement, les figures de style participent de ces effets. On est alors face à une transgression de la normalité qui exige un parcours interprétatif plus complexe. Parmi les figures de style, le degré de normalité ou de transgression est perçu de façon variable : une métonymie, qui correspond à un glissement de sens interne, sera généralement moins transgressive qu'une métaphore, qui s'accompagne d'un glissement entre champs sémantiques ; les métaphores elles-mêmes connaissent des classifications selon leur impact, qu'on peut considérer comme le degré de variation sémantique ressenti, avec par exemple l'échelle suivante : « poétique ou créative (= unique) - vivante ou vive - courante - codée - cliché - usée - morte, (…) (*cf.* Leech 1974, Newmark 1981) » (Landheer, 2001), si bien qu'une métaphore cliché telle que le *manteau blanc* de la neige (« *l'herbe avait revêtu son* ***manteau blanc*** *et nous claquions des dents* », MRÉJEN Valérie, *L'Agrume*, 2001, p.41) sera moins percutante qu'une métaphore vive telle que la *citrouille* de l'exemple suivant, réalité désagréable et brutale qui se rappelle soudainement au bon souvenir du narrateur :

> « L'instant est propice; je prétexte une plaidoirie à revoir à l'aube pour m'éclipser. Que de référés me sont tombés dessus soudainement vers minuit ! À chacun sa ***citrouille***... », PIERRAT Emmanuel, *Troublé de l'éveil,* 2008, p.131, *Dîner en ville.*

Autrement dit, la variation sémantique sera perçue comme plus importante dans le second cas.

Les variations les plus marquées s'accompagnent d'un sentiment de rupture. Celui-ci peut se traduire par un sentiment de nouveauté, sur lequel on reviendra (*cf.* chapitre I.3, 2.1.1), ou par un sentiment d'échec interprétatif. Ce dernier cas rejoint la question de l'asémantisme. L'interprétation se heurte à ses limites, elle n'est pas spontanée, l'interprétant rencontre une difficulté dont il a conscience.

3.1.2 Gradualité de la vitesse d'interprétation : immédiateté ou non de l'accès au sens

On fait ici l'hypothèse que le degré de variation sémantique peut être corrélé au degré de difficulté interprétative, qui est lui-même fonction du degré d'immédiateté d'accès au sens.

a- Compulsivité de l'interprétation : il y a accès au sens

Un individu parvient presque systématiquement à accéder à un sens. En effet, l'interprétation est compulsive : on ne peut s'empêcher d'interpréter.

Le caractère compulsif de l'interprétation pourrait donner à entrevoir celle-ci comme quelque chose de direct, de sûr et d'uniforme en termes de rapidité d'accès au sens. Il convient de revenir sur cette idée : l'interprétation n'est pas uniforme en termes d'immédiateté, il peut exister une discordance entre le sens littéral, qu'on abordera ici d'un point de vue cognitif en tant que sens mémorisé, et le sens autorisé par le contexte.

b- Effet Stroop : l'accès au sens est plus ou moins rapide

L'existence de décalages, voire de ruptures, apparaît chez (Le Ny, 1989) lorsqu'il rattache la question de l'accès lexical à l'effet Stroop. L'effet Stroop est un effet d'interférence entre un contenu sémantique et une information sémiotique. Il a été mis en évidence lors d'une expérience en psychologie cognitive de John Ridley Stroop en 1935. Un mot désignant une couleur, par exemple *rouge*, est écrit dans une certaine couleur (*rouge* ou *rouge*), que les sujets de l'expérience doivent dénommer. Le temps de réponse est accru si la couleur désignée ne correspond pas à la couleur appliquée au signifiant. Il y a donc interférence entre la représentation cognitive et une information contextuelle, ici sémiotique. (Le Ny, 1989) montre de plus que le temps de réponse dépend de l'amorçage sémantique, c'est-à-dire d'une préparation à l'accès sémantique impulsée par le mot précédant la désignation de couleur. Par exemple, le temps de réponse sera plus rapide pour identifier le code couleur 'bleu' pour *lac rouge* que pour *sang rouge*. Ces exemples témoignent d'une opposition sur un même axe sémantique, l'axe de la couleur, mais un phénomène similaire est à l'œuvre dans le cas d'incompatibilité sémantique de la deuxième série d'expériences rapportées par (Le Ny, 1989) et réalisées par (Tabossi, 1988). Ces expériences se situent dans une perspective de désambiguïsation et elles mettent en jeu une information contextuelle exclusivement sémantique, et non plus visuelle. Un mot apparaît dans un contexte, des parties de signifié sont proposées à la personne soumise à l'expérience et le temps de réponse est mesuré. Ainsi, pour le mot *or* dans la phrase *l'artisan travaille l'or,* le temps de réponse est mesuré pour les unités de sens /malléable/ et /jaune/. Il est plus court pour la partie de signifié pertinente, /malléable/ dans l'exemple relaté. Pour /jaune/, le temps de réponse plus long témoigne d'une incompatibilité entre le contexte et une partie du contenu sémantique.

c- Variations sémantiques marquées : un accès au sens plus lent ?

On peut supposer qu'une incompatibilité sur un ensemble d'unités clés du signifié, qu'on cherchera à préciser ultérieurement, correspond à une absence de désambiguïsation satisfaisante et donc à un amorçage sémantique non immédiat.

Ce pourrait être le cas d'*or* dans ce titre d'article du 16 août 2010 tiré du *Figaro* : « Le caoutchouc naturel, un « or vert » très fragile ». Le cotexte du titre n'amorce pas plus le sens de /métal/ que de /malléable/, /jaune/, /éclat/, /précieux/ ou /monnaie/. De plus, il génère des interférences qui interdisent certaines parties du signifié, comme /métal/ ou /jaune/. L'analogie entre pétrole et caoutchouc, c'est-à-dire l'or noir et l'or vert, est facilement accessible à tout lecteur, mais elle joue sur un processus analogue à celui de l'effet Stroop, donc, malgré son caractère d'évidence a posteriori, au moment de la lecture, elle ne s'impose probablement qu'après un certain temps de latence, correspondant à une interprétation non immédiate.

Une variation sémantique marquée pourrait s'expliquer par une difficulté d'amorçage sémantique et par une réaction au temps nécessaire au cerveau pour accéder au sens pertinent. Elle résulterait d'une incohérence ou d'une incompatibilité entre contexte et sens mémorisé. Ce dernier autorise une actualisation dans un certain espace sémantique, le contexte contraint à respecter un autre espace sémantique, et ces deux espaces, celui autorisé par les connaissances et celui autorisé par le contexte, ne sont pas compatibles. Le sens de la variation sémantique résultera d'un compromis entre l'espace sémantique autorisé par le contexte et celui défini par le sens littéral.

3.1.3 Gradualité au niveau des données discursives

a- Gradualité de la cohérence textuelle

Pour un emploi normal, les contraintes aussi bien sémantico-syntaxiques que globales (par exemple, par conformité avec le domaine présent ou par renforcement d'une isotopie) sont respectées. L'apport contextuel est faible, ou, plus précisément, il correspond à un renforcement maximal de l'information contenue dans le sens littéral, autrement dit, il ne crée par de tiraillement.

Les variations sémantiques plus marquées s'accompagnent d'un enrichissement, d'une déformation ou d'une reconfiguration atypique du sens littéral.

Au niveau du contexte, deux cas de figure peuvent se présenter :

1. Les contraintes textuelles ne permettent pas de naviguer de façon claire et univoque au niveau du sens littéral. C'est en particulier le cas des quiproquos, où le contexte génère une reconfiguration particulière du sens littéral, avec l'activation de deux acceptions à la fois. Ce cas de figure se présente également lorsqu'il y a incertitude interprétative, notamment lorsque le contexte n'est pas assez précis pour lever l'ambiguïté, cas que (Fuchs, 2008) qualifie de défaut de sens ou de sous-détermination du contexte.
2. Le sens littéral ne peut respecter les contraintes contextuelles. Cela peut se traduire par une non-coïncidence entre le domaine associé à l'unité lexicale et ceux induits par le contexte, ou par une difficulté à s'accorder avec des isotopies, voire une rupture avec celles-ci. De tels cas de figure peuvent résulter de stratégies palliatives, telles que décrites par (Duvignau *et al.*, 2004) : des enfants ou des personnes atteintes d'aphasie ne disposent pas du mot de vocabulaire approprié, ils recourent à des unités lexicales qui se rapprochent de celle qui leur manque, par analogie ou par hyperonymie ; l'emploi de telles unités crée une distorsion avec le contexte, par rupture de domaine ou d'isotopie.

Pour distinguer les variations, il est également nécessaire de considérer la cohérence textuelle de deux points de vue : celle d'une unité lexicale donnée relativement au cotexte et celle interne au cotexte, relativement à l'ensemble des unités présentes. Les variations les plus marquées sont identifiées par un ressenti (le sentiment de nouveauté) et par un échec interprétatif : elles ne se fondent donc pas dans le cotexte et, par définition, vont à l'encontre d'une certaine cohérence. On ne peut donc d'appuyer sur le critère de cohérence de l'unité relativement au cotexte pour faire la différence entre les deux cas de figure précédemment évoqués. En revanche, le critère de cohérence interne du cotexte permet d'établir une distinction entre variations marquées : celles pour lesquelles le cotexte oriente assez clairement vers une interprétation, bien que cette interprétation paraisse en contradiction avec le sens littéral (1) ; celles pour lesquelles le cotexte maintient un flou interprétatif (2).

(1) Les variations qui génèrent un sentiment de nouveauté se caractérisent par un contraste de type noir/blanc entre unité lexicale et cotexte : le cotexte présente une cohérence sémantique interne et l'unité lexicale présente une incompatibilité sémantique avec cette cohérence.

(2) Dans le cas de l'échec interprétatif, il y a un manque trop important de cohérence sémantique au niveau du cotexte, les informations contextuelles ne sont pas suffisamment convergentes ou claires pour faire émerger un sens local : le sens global ne se dessine pas assez clairement pour définir des contraintes sémantiques précises sur l'unité lexicale ciblée. Pour l'asémantisme, cas d'échec interprétatif maximal, la variation est plus de type rouge/{bleu, jaune, gris, vert}, c'est-à-dire qu'elle correspond à une nouvelle hétérogénéité par

rapport à un tout déjà hétérogène : l'incohérence ne se limite pas à l'incompatibilité d'une unité lexicale, mais le manque de cohérence sémantique s'étend plus largement sur l'ensemble du cotexte. Les cadavres exquis en sont une illustration percutante, comme dans cet exemple[6] "Au bout d'une nuit pénible, François premier des cabinets de gauche imite un globule ayant perdu le nord derrière le lit" : l'interprétation de *globule* est problématique à plusieurs égards, par rupture domaniale car le domaine de la biologie associé à *globule* n'est pas présent ailleurs dans le cotexte, par rupture sémantico-syntaxique, où la restriction de sélection du sujet de *perdre le nord* à un humain n'est pas respectée ; de plus, aucune isotopie ne s'impose ; l'effet produit est comique, il correspond à un sentiment de décalage.

b- Présence variable de traces discursives

La variation sémantique peut s'accompagner d'épiphénomènes dans les discours, témoins d'un sentiment de décalage, d'une annonce d'effets, etc.

Les emplois normaux ne génèrent pas de traces particulières. En revanche, pour des variations sémantiques plus marquées, il peut y avoir présence d'indices discursifs. À l'oral, ces indices peuvent prendre la forme de précautions oratoires ou, en dialogue, donner lieu à une négociation du sens, par exemple à travers une demande de précision, de rectification, etc. À l'écrit, auquel nous nous cantonnerons par la suite, des indices textuels variés peuvent signaler une variation sémantique : indices typographiques tels que des guillemets, gloses, explicitations, formules introductives, etc. On reviendra ultérieurement sur les indices de variations marquées (*cf.* chapitre I.3, 2.1).

3.2 Première réduction de l'échelle : seuil de sensibilité interprétative et seuil d'asémanticité

Parmi l'ensemble des variations sémantiques évoquées jusqu'ici, il semble judicieux d'écarter d'emblée deux extrêmes dans une approche axée sur du traitement automatique : les variations trop faibles et les variations donnant lieu à de l'asémantisme.

3.2.1 Variations faibles : sous le seuil de sensibilité interprétative

L'interprétation est certes compulsive, mais elle n'est pas toujours sûre. Il peut notamment y avoir une incertitude interprétative :

> « le sens […] résulte, dans un ajustement toujours fragile et provisoire, d'une co-construction qui comporte une marge indépassable d'ajustements intersubjectifs plus ou moins réussis » (Fuchs, 2008)

Le sens est certes le fruit de parcours interprétatifs individuels, mais, d'un individu à l'autre, les interprétations possèdent un noyau commun, pour permettre la communication. Cependant, ce noyau commun ne correspond pas à une coïncidence parfaite entre les différentes interprétations. Ainsi, l'interprétation intersubjective d'une unité lexicale n'est pas d'une précision chirurgicale, mais elle possède une certaine souplesse. De ce fait, cibler des variations sémantiques très faibles peut s'avérer problématique, notamment pour le traitement automatique : la validation serait alors périlleuse, puisque la sensibilité interprétative d'un ensemble d'individus ne peut dépasser un certain seuil de précision.

Les travaux en désambiguïsation font état d'un problème similaire (Ide et Véronis, 1998 ; Agirre et Edmonds, 2006) : l'évaluation humaine de la désambiguïsation se heurte à ses limites, notamment lorsque les variations sémantiques sont trop fines, et les taux d'accord

[6] Source : site Internet *Mots de tête*, dédié aux jeux de mots, dont une rubrique est consacrée à la génération de cadavres exquis : http://www.mots-de-tete.com/cadavre/accueil.php3?nbreaffich=511.

inter-annotateurs peuvent atteindre des valeurs parfois peu encourageantes (Véronis, 2004). De plus, un degré de granularité trop fin dans la représentation de la polysémie est triplement problématique : d'une part, il présente une précision excessive par rapport à la plupart des objectifs en traitement automatique des langues, donc il n'est pas adapté ; d'autre part, il génère des difficultés lors des traitements, notamment en termes de combinatoire (Ide et Véronis, 1998) ; enfin, il n'est pas conforme à une caractéristique de la communication humaine, à savoir la propriété de flou qui est partie intégrante de celle-ci. L'étude des variations sémantiques fines entre le sens littéral codé dans une ressource et la nuance apportée par le contexte peut susciter des objections similaires, c'est-à-dire en matière d'intérêt que présente une telle étude et en matière de complexité de la combinatoire.

Les variations sémantiques trop fines, c'est-à-dire inférieures au seuil de sensibilité interprétative, seront donc exclues de notre propos, seules seront considérées les variations sensibles.

3.2.2 Asémantisme : marginal, difficile d'accès et sans rétroaction sur le sens littéral

On a évoqué précédemment le rôle du manque de cohérence textuelle dans l'asémantisme. Il convient de s'interroger sur la représentativité d'un degré de variation en termes de densité textuelle, c'est-à-dire, sur la quantité de variations sémantiques marquées pour une taille de cotexte donnée : une forte concentration des variations transgressives pour une fenêtre textuelle donnée est-elle un phénomène fréquent ?

Les discours qui ne comportent que des emplois normaux, c'est-à-dire tels que le sens soit facilement déductible du sens littéral et faiblement reconfiguré ou enrichi par le cotexte, sont marginaux. (Duvignau *et al.*, 2004) montre qu'un individu normal dispose d'une compétence de "flexibilité sémantique fondamentale", c'est-à-dire qu'il a la capacité d'émettre et d'interpréter des approximations sémantiques. À l'inverse, l'absence d'une telle flexibilité relèverait de la pathologie. Chez les personnes atteintes du syndrome d'Asperger, la production discursive est extrêmement précise, avec un respect systématique de ce que nous qualifions de sens littéral ; à l'inverse, au niveau de la compréhension, ces personnes se heurtent à des "problèmes d'interprétation du non-littéral dont les énoncés métaphoriques", c'est-à-dire que les variations sémantiques marquées leur sont difficiles d'accès. La présence de variations sémantiques conséquentes n'est donc pas anecdotique.

Certains genres ou pratiques discursives favorisent une faible densité de distorsions ou de transgressions sémantiques, d'autres les encouragent. Des documents techniques privilégieront une rigueur de l'expression conforme au sens codé dans une ressource terminologique, de même que des essais chercheront une précision du vocabulaire employé, c'est-à-dire une certaine conformité des emplois avec le sens littéral et sa structure. Cette rigueur n'exclut cependant pas les variations sémantiques. À l'inverse, certains genres favorisent des variations sémantiques marquées par rapport aux usages conventionnels, en particulier la poésie (notamment, la poésie symboliste, la poésie surréaliste) où prolifèrent les effets stylistiques, donc le jeu sur des variations sémantiques.

Il semble légitime de supposer que les variations sémantiques, qu'il s'agisse d'approximations ou de transgressions, sont nombreuses mais peu denses. En effet, selon Grice (Moeschler et Reboul, 1994:204), les échanges discursifs sont régis par un principe de coopération, qui s'appréhende à travers quatre maximes, selon lesquelles l'information transmise doit être nécessaire et suffisante en quantité, respecter une certaine véridicité (maxime de qualité), être pertinente et être claire. Or une accumulation de transgressions ne respecte pas la maxime de pertinence si elles ne s'éclairent pas mutuellement, ni la maxime de quantité puisque les différentes unités lexicales ne fournissent pas suffisamment de matière

pour permettre l'accès au sens, ni la maxime de clarté, par dispersion du sens et absence de cohérence sémantique. De même, (Sperber et Wilson, 1989) estiment que la communication est régie par un principe de pertinence, selon lequel le locuteur cherche la réussite de la communication, donc le succès interprétatif. Or la pertinence d'un énoncé est moindre lorsqu'il y a peu d'effets contextuels, c'est-à-dire lorsque la construction du sens est peu étayée par le contexte et qu'il y a peu d'ancrage et de liens interprétatifs avec les informations contextuelles. Ce point de vue impose donc au contexte d'être suffisamment cohérent, ce qui invite à voir comme marginal les cas d'importantes dispersions sémantiques dues à une forte densité de variations sémantiques indépendantes les unes des autres. Autrement dit, il semble légitime de supposer que la densité de transgressions reste limitée : elles peuvent être fréquentes, mais doivent rester disperses. Les cas d'asémantisme existent, mais ils restent anecdotiques, ils sont plus le fruit d'une production artificielle, issue de jeux littéraires avec la langue pour en tester les limites, que le résultat d'une production discursive spontanée et véritablement ancrée dans les pratiques. L'asémantisme constitue donc un cas marginal, non représentatif des usages, c'est pourquoi, par la suite, il ne fera pas l'objet de nos investigations.

3.3 Mécanismes des variations sensibles : recensement et hiérarchisation

(Fuchs, 2008) décrit les états du signifié correspondant à de l'incertitude interprétative :

- Le ***défaut ou l'excès de sens***, respectivement dus à une sous-détermination, où les informations contextuelles ne sont pas suffisantes pour permettre une interprétation sûre, ou à une surdétermination, qui impose une double lecture par cumul de sens, mise en œuvre de sous-entendus ou de présupposés, en bref où interviennent « plusieurs strates signifiantes », comme dans le slogan de l'eau minérale Cristaline, « Cristaline, ça coule de source », qui comporte la double interprétation du choix de la marque comme un choix qui s'impose et de l'origine naturelle du produit, ou encore comme dans le jeu de mots « Il pleut des cordes / J'ai envie de me pendre » (*Le Secret du Temps Plié*, de l'humoriste Gauthier Fourcade), où le sens intensif de *cordes* dans le phrasème *pleuvoir des cordes* se superpose à celui d'instrument de suicide.
- La ***concurrence de sens***, où il y a existence d'une ambiguïté correspondant à une « univocité dédoublée, qui induit une instabilité interprétative momentanée » :il y a hésitation entre plusieurs sens qui émergent, mais qui ont peu de chance d'être justes simultanément, par exemple dans "*Il a rapporté un vase de Chine"*, où il y a hésitation entre le sens de vase quelconque que le porteur a fait voyager depuis la Chine et un vase dans le style chinois, mais acquis ailleurs qu'en Chine, par exemple dans une échoppe française.
- Le ***glissement de sens*** : « dans un grand nombre de cas, l'interprétation de l'énoncé est constitutionnellement instable : elle semble osciller et couvrir toute une zone de sens variable, sans pouvoir se fixer en un point déterminé. L'incertitude tient alors, non pas à l'impossibilité d'un choix entre plusieurs sens mutuellement exclusifs, mais à la difficulté à pointer un sens précis », autrement dit plusieurs sens sont possibles, l'interprétation fait l'objet d'hésitations et elle varie chez un même sujet ou d'un sujet à l'autre, sans qu'un sens s'impose de manière stable et durable.

(Fuchs, 2008) aborde ces états dans la perspective de l'incertitude interprétative, ce qui rattache son propos à la question du seuil de sensibilité interprétative. Cependant, en partant de ces états et en les développant, il est possible de passer de la question de l'incertitude interprétative à celle du basculement interprétatif, donc d'y voir des descripteurs de variations sémantiques marquées. Le défaut et la concurrence de sens sont d'un intérêt moindre dans cette perspective : ils résultent non pas de contraintes qui imposent un sens différent, mais

d'un manque de contraintes qui ne permet pas de préciser suffisamment le sens. Par contre, l'excès de sens et le glissement de sens évoquent des mécanismes plus pertinents pour les variations sémantiques. L'excès de sens est dû à une activation particulière du signifié, avec une double focalisation, autrement dit, il met en jeu une configuration particulière dans l'activation du signifié. Quant au glissement de sens, plutôt que d'y voir une dynamique oscillatoire, on peut adapter cette représentation à celle d'un déplacement entre deux états sémantiques. On retiendra donc deux bases de mécanismes : la reconfiguration du signifié par des activations particulières et le glissement sémantique entre deux états.

Une approche plus éclairante consiste à analyser les mécanismes à l'origine de la polysémie. La polysémisation est le fruit de variations sémantiques, dont les tropes sont souvent à l'origine (Lecolle, 2007). Des indices sur les mécanismes à l'origine de changements de sens sont généralement décrits dans des ressources lexicographiques. Ainsi, le TLFi propose notamment les indicateurs suivants : par métaphore ; par métonymie ; par comparaison ; par analogie ; au figuré ; par extension ; en particulier ; par plaisanterie ; par antiphrase ; par euphémisme ; par hyperbole ; par exagération ; par ironie. Les indicateurs utilisés répondent à des impératifs littéraires et linguistiques. (Lecolle, 2007) propose une classification plus sélective, reposant sur cinq modalités de la polysémie qui sont sources du changement de sens : la métaphore ; le passage du concret à l'abstrait et inversement ; la métonymie ; la restriction de sens ; l'extension de sens.

À partir des mécanismes de polysémisation décrits ci-dessus, il est possible de dégager la liste suivante de mécanismes qui sous-tendent les variations sémantiques :

- des jeux sur les degrés d'intensité qui peuvent caractériser le sens de l'unité lexicale (cas de l'exagération, de l'euphémisme et de l'hyperbole) ;
- des polarités inversées : un trait sémantique ne se contente pas d'être inhibé, il est transformé en son opposé (cas de l'ironie, de la plaisanterie, de l'antiphrase ; les acceptions 'au figuré' se rapprochent de cette catégorie, par transformation du concret en abstrait et vice-versa) ;
- des élargissements ou restrictions du sens (extension) ;
- des glissements de sens : ce sont les cas de métonymie, de comparaison, d'analogie et de métaphore ; ces glissements de sens peuvent être internes à un champ sémantique donné ou encore à un domaine donné, ou ils peuvent être externes, c'est-à-dire avec passage d'un champ ou d'un domaine à un autre.

À ces mécanismes s'ajoutent des ruptures plus brutales, qui se traduisent en pratique par l'apparition d'homonymes dans des ressources lexicographiques.

Nous proposons la hiérarchie suivante, en fonction du mécanisme sous-jacent et du degré de variation sémantique qui semble y correspondre. L'ordre correspond à un degré décroissant de variation :

1) ***Rupture complète :*** le sens contextuel ne peut s'expliquer par la structure ou les composantes du sens littéral préexistant ; le lien peut exister, mais il n'émerge pas linguistiquement.
2) ***Changement de topos :*** ce changement est typique des métaphores et repose sur de l'analogie ; il se caractérise par un glissement externe correspondant à un changement allotopique, c'est-à-dire d'un champ sémantique à un autre, ou encore d'un domaine à un autre.
3) ***Changement cotopique :*** il correspond à un glissement de sens interne, par contiguïté, caractéristique des métonymies ; le champ sémantique ou domaine se conserve, mais

la focalisation sémantique change, avec une inhibition de parties du sens littéral caractéristiques du concept et adjonction d'un contenu sémantique correspondant au concept contigu.

4) ***Reconfiguration particulière :*** elle correspond à un profil particulier d'activation ou d'inhibition de parties du sens littéral, avec en particulier une activation de deux parties qui ne sont généralement pas activées simultanément.
5) ***Changement de polarité :*** la variation repose sur la transformation d'une partie du sens littéral en son opposé, par exemple par passage du concret à l'abstrait, de l'animé au non-animé, d'une connotation positive à une connotation négative.
6) ***Changement de degré :*** il correspond notamment à la perte d'intensificateurs constitutifs du sens littéral, ou à l'évolution de parties de sens littéral à caractère graduel vers un degré plus fort ou plus faible.

Il semble légitime de supposer que les changements les plus marqués s'accompagnent de variations du cotexte à de multiples niveaux, en particulier à des niveaux qui dépassent le cadre des restrictions syntaxiques. On reviendra sur cette question (*cf.* chapitre I.3).

Le degré de variation sémantique semble donc fonction du mécanisme en jeu. Cependant, ce mécanisme n'est pas le seul à influer sur le degré de variation : celui-ci dépend des informations lexicalisées ou non, ainsi que d'autres facteurs, plus quantitatifs. Ainsi, les variations marquées, dont on précise dans la sous-section suivante l'intérêt tout particulier, sont susceptibles de s'imposer et d'être candidates à la lexicalisation si elles s'accompagnent d'une répétition ou d'une diffusion.

3.4 Exemple prototypique des variations marquées : le cas des nouveaux sens

Parmi les variations sensibles, nous privilégierons les variations marquées, et plus précisément celles susceptibles de générer de nouveaux sens. Celles-ci présentent un véritable enjeu, car elles sont un défi ignoré dans un premier temps par les travaux en désambiguïsation sémantique et auquel elle ne sait pas répondre.

À l'heure actuelle, les variations internes au sens littéral codé sont de mieux en mieux traitées (Agirre et Edmonds, 2006). L'homographie est considérée comme un problème résolu, avec plus de 95% de précision. La polysémie générale a fait l'objet de résultats concluants, particulièrement en apprentissage supervisé : les meilleurs systèmes se sont avérés capables de rivaliser avec l'évaluation humaine, en termes d'accord inter-annotateurs lors de la campagne d'évaluation Senseval-3, même s'il faut souligner que celle-ci repose sur une approche *in vitro* et non *in vivo*.

En revanche, la construction de nouveaux sens reste un problème non résolu et peu abordé, d'autant plus délicat que ce problème doit s'accommoder d'un autre problème, à savoir de la non-exhaustivité des descriptions lexicographiques des sens existants :

> « **Sense discovery**. A sense inventory that a priori lists all relevant senses will never be able to cope with borrowed words, new words, new usages, or just rare or spurious usages. In practical terms, this makes it very difficult to move a system into a new domain. » (Agirre et Edmonds, 2006:19)

Les variations sémantiques marquées susceptibles de donner naissance à de nouveaux sens sont donc un objet d'étude pour lequel beaucoup reste à faire et qui, de plus, présentent un enjeu lexicographique de taille, à travers la question des mises à jour de ressources lexicographiques. Le chapitre qui s'ouvre aura pour objectif de préciser ce que sont de telles variations sémantiques marquées, susceptibles de faire l'objet d'une lexicalisation, et qu'on qualifiera désormais de néologie sémantique.

Chapitre I.2
La néosémie : enjeu et profil d'une variation marquée en cours de diffusion

1. Intérêt de la néologie sémantique

1.1 La néologie sémantique, témoin de l'évolution de la langue

Une langue vivante est une langue qui évolue. Cette dynamique, aussi bien phonologique que syntaxique ou sémantique, constitue un principe essentiel de la langue. L'émergence de nouveaux signifiés en est une manifestation élémentaire.

1.1.1 La langue n'est pas seulement un état, mais aussi un processus

La langue peut être appréhendée selon deux éclairages différents, comme état ou comme processus. Selon que la langue est étudiée en synchronie ou en diachronie, l'un ou l'autre des deux éclairages est privilégié. La perspective synchronique revient à considérer la langue comme un système, c'est-à-dire comme un état ou éventuellement une succession d'états, autrement dit, elle invite à considérer la langue prioritairement d'un point de vue statique. La perspective diachronique présente la langue comme un processus, elle invite à la considérer d'un point de vue dynamique, comme un objet sous le signe du changement.

Au cours de l'histoire de la linguistique, le point de vue dominant a été celui de la langue comme état. Cette tendance principale est un héritage du *Cours de Linguistique Générale* (CLG ; Saussure, 1960). Saussure présente la langue comme un système, où prime l'étude des états de langue. (Picton, 2009:9-18) souligne que le CLG a ainsi amené la linguistique du XX^{e} siècle à privilégier un regard synchronique sur la langue, donc moins axé sur son évolution dans le temps que sur la description du système linguistique à un moment donné, et, surtout, moins attentif aux transitions entre états de langue. Cependant, des courants se sont distingués de la tendance dominante, comme celui défendu par (Coseriu, 1973), et ont mis en œuvre des approches privilégiant la diachronie, donc un point de vue sur la langue qui met l'accent sur le processus.

Nous privilégions ce point de vue : nous abordons la langue comme un processus, qui n'est pas seulement un donné mais qui correspond à une évolution permanente. La dynamique ne repose pas seulement sur l'existence d'une succession d'états, mais aussi sur l'existence de passages d'un état de langue à un autre. Ces transitions participent du caractère vivant de la langue, c'est sur elles que nous focaliserons notre attention.

L'évolution de la langue peut être considérée globalement, comme celle d'un tout, mais elle se traduit aussi localement par des évolutions au niveau des unités qui la composent, à savoir les unités lexicales.

Le changement peut intervenir à tous les niveaux : phonologie, morphologie, syntaxe, lexique ou encore sémantique. Dans le cas de la sémantique, centre de notre intérêt, la dynamique de la langue devient une dynamique du sens. Cette dynamique du sens correspond à des variations du sens d'unités lexicales au cours du temps, c'est-à-dire à un changement durable de leur signifié, éventuellement couplé à un changement de signifiant. L'évolution de signifiés participe au processus d'évolution de la langue. L'étude de la néologie sémantique revient donc à celle des processus élémentaires qui sous-tendent la dynamique de la langue.

Loin d'être un phénomène uniforme d'une unité à l'autre, la néologie sémantique s'inscrit dans un système complexe, marqué par des dépendances entre partie et tout (1.1.2), ainsi que par des propriétés de déformation qu'on qualifiera de plastiques (1.1.3).

1.1.2 Analogie avec le jeu d'échecs : le changement lexical dans sa dépendance au système

La dynamique du sens des unités est corrélée à celle du tout sur plusieurs plans, comme l'illustre l'analogie avec le jeu d'échecs de (Saussure, 1960:43-125-127). Dans le CLG, Saussure compare le système de la langue à un jeu d'échecs. Les pièces du jeu correspondent aux composantes de la langue, c'est-à-dire qu'elles sont assimilables, dans notre cadre, aux unités lexicales. Tout comme les pièces du jeu, les composantes de la langue exercent des contraintes mutuelles. Elles sont partie intégrante d'un système qui évolue dans le temps. Chaque état du système s'inscrit dans une progression dynamique et les transitions entre états sont provoquées par des déplacements élémentaires. Ces déplacements élémentaires, qui correspondent à des évolutions d'unités lexicales dans notre cadre, ont chacun leur spécificité, ils définissent une nouvelle configuration pour le système dans sa globalité et ils modifient les relations de l'élément en mouvement avec d'autres éléments du système. On retiendra de la comparaison de la langue à un jeu trois aspects fondamentaux :

- ***Dépendance entre partie et tout :*** unité et tout sont fortement dépendants, ils rétroagissent l'un sur l'autre. L'évolution d'une composante dépend de l'environnement proche, mais elle est aussi conditionnée par des informations de plus haut niveau, qui caractérisent l'environnement au sens large. Une observation locale ou globale est à même de donner un certain nombre de renseignements sur les états ou transitions, mais une étude solide gagne à exploiter les différentes échelles et à les faire interagir.
- ***Évolution progressive globalement, possibilité d'évolution brutale localement*** : l'ensemble du système ne progresse pas de façon brutale, mais il évolue lentement. Ainsi, dans le jeu d'échecs, le déplacement d'une seule pièce permet le passage d'un état à un autre, tandis que toutes les autres pièces restent en place. De façon similaire, dans la langue, les évolutions lexicales n'affectent simultanément qu'une petite fraction de la langue : trop de changements simultanés nuiraient à l'intelligibilité du discours, et affecteraient la communication[7]. Le système dispose donc d'une certaine inertie dans son ensemble : il n'évolue que faiblement dans son ensemble sur un laps de temps restreint. Cependant, cette inertie globale n'exclut pas que les changements élémentaires, c'est-à-dire au niveau des unités, puissent être marqués : un fou peut passer en un coup d'une extrémité à l'autre de l'échiquier ; le *chat* informatique, discussion virtuelle, présente une rupture sémantique avec son homographe, le *chat* animal domestique : bien que les deux

[7] Trop d'évolutions simultanées nuisent à la cohérence du discours et, de ce fait, ne respectent pas la règle de pertinence de Grice (*cf.* chapitre I.1, 3.2.2).

homonymes partagent le même signifiant, le nouvel arrivant qu'est le *chat* informatique a une autre étymologie et un sens sans rapport avec l'animal. Mais ces changements marqués ne se produisent que sur une petite partie du système.

- ***Comportements distincts des unités :*** toutes les unités n'ont pas le même statut. Chacune a ses caractéristiques propres, en tant qu'entité isolée et aussi à travers ses dépendances avec les autres entités. Un pion n'aura ni le même type, ni la même importance de déplacement potentiel qu'une tour ; de même, les changements sémantiques ne peuvent donc se concevoir de manière uniforme.

L'analogie avec le jeu d'échecs permet d'appréhender les rapports entre changement local et changement global. Néanmoins, elle ne rend pas compte de facettes de la dynamique du sens qui sont fondamentales pour comprendre la complexité et le caractère essentiel de la néologie sémantique.

L'image du jeu met en valeur les états successifs, non les transitions entre états. Les pièces de l'échiquier se déplacent, certes, mais seuls l'état initial et l'état final importent dans le déplacement : entre les deux, les contraintes cessent de s'exercer sur la pièce, qui peut aussi bien glisser sur le plateau qu'en décoller. De ce fait, le jeu d'échecs apparaît comme un processus discret. En revanche, la langue n'évolue pas de façon discontinue, et de plus, les transitions ne sont négligeables ni par leur rôle, ni par leur étendue. Au niveau de leur rôle, elles assurent la continuité de l'évolution du système et permettent d'anticiper et d'expliquer les états successifs. En termes d'étendue, elles se mettent en place progressivement à plusieurs niveaux : elles s'étalent dans le temps, mais elles connaissent aussi une diffusion progressive dans des discours, parmi des communautés de locuteurs, ou encore dans l'espace.

Par ailleurs, le jeu d'échecs donne à voir des positions clairement définies et rigides. Il n'existe pas de demi-mesure dans les variations : elles correspondent à des déformations permanentes, qui participent nécessairement de la progression du système. L'analogie avec le jeu ne rend pas compte du caractère souple et modulable de langue, et notamment du sens.

Pour compléter l'analogie avec le jeu et prendre en compte les aspects qu'elle éclipse, nous proposons une autre analogie, celle de la plasticité, qui met en valeur les transitions et permet d'intégrer l'idée de réversibilité d'une déformation, donc de distinguer les changements dus à la modularité de la langue des changements durables, participant du processus d'évolution de la langue.

1.1.3 Analogie avec la plasticité : de la déformabilité à la déformation permanente du sens lexical

La langue peut être vue comme un système doté d'une propriété fondamentale : la plasticité. Le concept de plasticité inclut deux idées majeures : celle de déformabilité, c'est-à-dire de changement potentiel, et celle d'irréversibilité de la déformation, c'est-à-dire de changement qui n'est pas accidentel mais qui s'inscrit durablement dans le système, même lorsque les contraintes de l'environnement se relâchent – autrement dit, par-delà la contingence d'un emploi contextuel.

Pour préciser ce qu'on entend par ce concept, on s'appuie sur l'analogie avec la plasticité de la matière telle qu'elle est définie en physique. Certains matériaux ont la capacité de se déformer. La déformation peut être réversible ou irréversible. La déformation réversible est possible grâce à la propriété d'élasticité, qui est une propriété complémentaire de la plasticité : le matériau se déforme sous l'effet de sollicitations, puis revient à son état initial une fois que les sollicitations cessent. Sous certaines contraintes de l'environnement, lorsqu'elles deviennent trop importantes, la déformation franchit le seuil d'élasticité et devient irréversible. Elle stabilise la matière dans un nouvel état, où la configuration atomique change et où la

matière peut acquérir de nouvelles propriétés, par exemple, un métal pourra connaître un renforcement. Dans ce cas, c'est la plasticité qui intervient, c'est-à-dire la propriété de la matière à se déformer de façon irréversible.

Nous tenons à insister particulièrement sur deux points saillants : 1) la plasticité peut être abordée comme potentiel, comme processus ou, à travers la notion de déformation plastique, comme résultat du processus ; 2) la notion de plasticité, déformation irréversible, est indissociable de la notion d'élasticité, déformation réversible.

On propose de considérer la plasticité comme une propriété linguistique, et plus précisément sémantique. L'analogie entre la plasticité de la matière et la langue met en relief les aspects suivants :

- ***Déformabilité et potentiel de sens*** : le sens a un potentiel de déformation. Ce potentiel de déformation correspond aux potentiels de sens, c'est-à-dire des facettes sémantiques susceptibles de s'actualiser de façon souple et modulée, comme évoqué au (chapitre I.1, 2.3.2). La variation sémantique est une propriété inhérente au système linguistique.
- ***Transformation irréversible de la matière et nouveau sens littéral*** : si la matière est déformée sous les contraintes de l'environnement, mais pas durablement, cela correspond dans la langue à une déformation temporaire du sens, qui reste tributaire des contraintes de l'environnement, c'est-à-dire du contexte, mais qui n'affecte pas le sens lexical lorsque les contraintes contextuelles se relâchent. Autrement dit, il n'y a pas de déformation durable ou de restructuration du sens littéral. À l'inverse, la matière peut subir une déformation plastique, qui perdure lorsque les contraintes de l'environnement se relâchent et qui fait éventuellement acquérir à la matière de nouvelles propriétés. Cette déformation plastique peut être associée à un changement durable du sens littéral : celui-ci connaît une nouvelle configuration ou même acquérir un nouveau contenu d'un point de vue cognitif, qui pourra se traduire par l'enregistrement du nouveau contenu dans la ressource de référence.
- ***Idée de rupture ou de transgression :*** la déformation plastique résulte d'un franchissement du seuil d'élasticité. Ce franchissement de seuil résulte des sollicitations de l'environnement. La notion même de seuil renvoie à l'idée de transgression. Cette transgression se retrouve pour la néosémie, qu'on a définie comme une variation sémantique marquée, qui présente une forme de rupture entre sens littéral et contexte : les sollicitations du contexte, ou plus exactement les contraintes exercées par le contexte, sont à l'origine de la transgression.
- ***Couples élasticité/plasticité et polysémie/néosémie :*** la plasticité est indissociable de l'élasticité. On propose de mettre en parallèle le couple élasticité/plasticité et le couple polysémie/néosémie. L'élasticité correspond à des propriétés de souplesse, de modularité et d'adaptation aux contraintes de l'environnement ; la plasticité correspond à des propriétés de souplesse limite, d'évolutivité et de transformation par l'environnement. De même, la polysémie est associée à l'adaptation du sens littéral au contexte, par reconfiguration et par extensions autorisées par les potentiels de sens. La néosémie correspond, dans notre définition, à une variation sémantique transgressive, susceptible de faire évoluer le sens littéral sous l'effet de l'environnement contextuel. La polysémie invite à voir le sens comme modulable, la néosémie invite à le voir comme évolutif. Pour la matière, une déformation élastique peut être vue par un observateur extérieur comme une déformation élastique qui dure. Autrement dit, l'élasticité renvoie à la capacité de la matière à se déformer ; la plasticité renvoie à la capacité à se déformer durablement. De même, en sémantique, la néosémie peut être vue comme une polysémie qui dure, ou plus exactement comme une polysémisation qui dure. Lorsqu'on est face à de la néosémie, la

polysémisation devient permanente, elle survit au fait que le sens s'affranchit du contexte (par analogie avec le relâchement des sollicitations exercées sur la matière), autrement dit, elle s'intègre au sens littéral.

Par-delà l'analogie avec la matière, le recours au terme de plasticité est éclairant car il établit un lien entre processus linguistique et processus cognitif. En effet, le terme de plasticité est employé pour qualifier une propriété du cerveau, où on parle de plasticité cérébrale ou de plasticité neuronale. Cette propriété caractérise des mécanismes cognitifs : certaines aires du cerveau associées à des connaissances ou à des capacités sont susceptibles de s'accroître ou de diminuer sous l'effet de pratiques répétées, ou lors de la cessation de ces pratiques. Par exemple, l'aire du cerveau associée au petit doigt tend à s'accroître chez les violonistes, dont les pratiques amènent à solliciter fortement les capacités motrices de leur auriculaire.

Les processus linguistiques sont associés à des processus cognitifs, comme cela apparaissait au (chapitre I.1, 3.1.2) : le sens littéral est susceptible de renvoyer au sens mémorisé ou au sens codé dans une ressource de référence, selon la perspective. Une partie de la plasticité cérébrale se rattache aux processus cognitifs d'allocation d'un espace-mémoire au sens associé à une lexie. La plasticité sémantique peut se concevoir comme le versant linguistique de cette partie de la plasticité cérébrale.

Pour résumer, la plasticité peut être considérée comme une propriété au cœur même du processus de transformation de la langue, plus précisément du sens. Elle prend corps à travers la néosémie. La néosémie apparaît à travers cette association en étroite relation avec la polysémie : la néosémie est une polysémie qui dure, donc qui nécessite de prendre en compte le facteur temps et la question de la répétition dans le temps. La néosémie contribue donc à ce qu'une langue soit vivante et à expliquer le passage d'états en états. Elle est un pilier fondateur, au niveau élémentaire, d'un système en progression. Nous précisons dans la sous-section qui s'ouvre l'approche de la langue que nous adoptons.

1.2 Mise à jour du sens codé : trancher entre une apparente transition plastique et un devenir inconnu

La langue est un processus, elle est soumise à une évolution permanente et progressive. Cependant les outils qui servent à la décrire ne peuvent saisir cette dynamique continue, ils renvoient des images figées qui correspondent à des successions d'états. (Coseriu, 1973) propose ainsi de différencier langue réelle et langue abstraite :

- la *langue réelle* est la langue telle qu'elle est pratiquée, animée d'une dynamique incessante. Elle incarne les propriétés d'élasticité et de plasticité décrites précédemment ;
- la *langue abstraite* est la langue représentée par les dictionnaires ou les grammaires. Elle peut se concevoir comme une série d'arrêts sur image, reflets de la langue réelle à un moment donné.

Entre langue réelle et langue abstraite, un décalage existe, il n'y a jamais identité parfaite : tout état de langue représenté par une ressource de référence est déjà dépassé par la langue réelle. Le décalage est d'autant plus important que les mises à jour de la ressource de référence considérée sont espacées. De plus, l'état de langue reflété par une ressource relève de choix par rapport à certains phénomènes linguistiques en cours de construction. À titre d'illustration, le *Petit Robert 2008* n'a pas intégré le substantif *bravitude*, néologisme introduit involontairement par Ségolène Royal lors de la campagne présidentielle en 2007, puis repris et diffusé notamment par la presse. Les raisons de son rejet ont été précisées :

> « *Bravitude* aurait-il pu entrer dans le *Petit Robert 2008* ? Non. Parce qu'on ne le retrouve qu'en contexte particulier, en référence à Ségolène Royal. Il n'appartient donc pas (encore ?) au langage courant ni ne correspond à une réalité sociale nouvelle. » (Alain Rey, l'édito du blog de la rédaction du Petit Robert, http://www.lerobert.com/le-blog-de-la-redaction.html, consulté le 24 octobre 2011)

Ces propos d'Alain Rey témoignent d'incertitudes sur la situation et le devenir de *bravitude*, ainsi que de la nécessité pour le lexicographe d'effectuer des choix face à des phénomènes en cours.

Il n'est pas nécessairement indispensable d'avoir une connaissance précise de la langue entre deux états de langue, encodés par des ressources : cela dépend de l'objet d'investigation et de l'échelle d'observation. Si on cherche à observer un phénomène qui s'étale sur une grande période de temps, par exemple sur plusieurs décennies ou siècles, les états de langue décrits par les ressources seront suffisamment nombreux et denses pour caractériser l'ensemble de la période considérée et donner une vue d'ensemble du comportement du phénomène linguistique.

En revanche, la maîtrise des états transitoires revêt une importante particulière dans les deux cas suivants :

- ***Anticipation d'un état de langue inconnu :*** un phénomène linguistique est en cours, un changement lexical est amorcé mais le résultat de ce changement n'est pas connu. Ce cas de figure concerne les changements linguistiques actuels, c'est-à-dire les états de langue en cours de structuration et d'implantation, qui n'ont pas encore été intégrés dans des ressources de référence. Cette problématique concerne en particulier la veille lexicographique et la mise à jour des ressources. Les lexicographes incarnent des professionnels chargés de proposer une description d'états successifs de la langue. Le cas de *bravitude* précédemment cité est une illustration des problèmes rencontrés : certains changements lexicaux existent de façon sensible, mais ont une certaine chance de ne pas être pérennes ou ne connaissent pas une diffusion suffisante.
- ***Investigation sur une période réduite :*** on peut observer des changements linguistiques sur une période de quelques mois ou années. La ressource considérée n'en fera pas nécessairement état. La fréquence de mise à jour de la ressource, trop espacée, peut être une explication du décalage.

À l'heure actuelle, il est de plus en plus nécessaire d'accéder à de l'information à jour de façon réactive, c'est-à-dire de saisir l'évolution de la langue de façon anticipative et sur une période réduite. Le problème est central lorsqu'un observateur se situe au temps présent et qu'il doit trancher sur des phénomènes en cours d'évolution, autrement dit lorsque le nouvel arrêt sur image n'a pas encore été défini mais qu'il est en cours d'élaboration. Ce problème est celui de la veille lexicale, d'autant plus sensible que le XXI^{e} siècle s'est accompagné d'une accélération dans l'évolution linguistique, comme nous allons le détailler.

1.3 Un triple accroissement : masse de données, liberté des pratiques discursives, rapidité d'implantation

L'évolution de la société, marquée par le développement des Technologies de l'Information et de la Communication (TIC) et la mise en place d'une société de l'information, favorise l'émergence de nouveaux sens, et plus généralement l'apparition de néologies. Cette évolution se traduit à travers une triple expansion qui confère un intérêt accru à la néologie sémantique et à l'élaboration d'outils pour accéder aux nouveaux sens. Les trois pôles d'évolution sont

celui de la quantité de données, de la souplesse introduite par de nouvelles pratiques discursives et de la vitesse d'évolution de la langue.

1.3.1 Accroissement des données : nouveau terrain et nouveaux potentiels pour la linguistique

Les TIC sont à l'origine d'un foisonnement des données discursives. L'internet connaît une expansion croissante, il contribue à une diffusion toujours plus vaste de l'information. Au sein des entreprises ou des collectivités, les outils informatiques facilitent la production et le stockage de données, ils favorisent le développement de bases documentaires de plus en plus conséquentes qui présentent deux aspects : elles transforment les pratiques et facilitent la circulation de l'information et l'accès à celle-ci ; elles constituent des supports de données exploitables.

Pour la néologie sémantique, ce nouveau terrain présente un intérêt double :

- D'une part, il change le profil du phénomène, car il favorise son amplification. En effet, il permet une diffusion plus large de la néologie et, en favorisant les échanges d'informations, il permet d'y être confronté plus facilement.
- D'autre part, il propose une matière à même de révolutionner les pratiques d'observation, autrement dit, il donne les moyens d'observer la néologie sémantique. De nouvelles pratiques linguistiques ont été mises en place avec ce nouveau champ de données, en particulier, l'accroissement et l'accessibilité informatique des données ont favorisé le développement de la linguistique de corpus, et plus particulièrement de la lexicographie de corpus. L'existence de vastes bases de données se prête à l'élaboration de nouveaux outils capables d'exploiter la masse textuelle pour observer l'évolution du lexique, en particulier à des fins de veille néologique (Rousseau et Depecker, 1999).

1.3.2 Assouplissement de contraintes exercées sur la néologie à travers de nouvelles pratiques discursives

L'apparition de nouveaux modes de communication s'est accompagnée de l'émergence de nouvelles pratiques discursives, à travers les blogs, les chats, ou les SMS, qui sont plus favorables à l'innovation lexicale. Ces pratiques présentent généralement plus de souplesse par rapport aux contraintes d'expression et de rédaction, imposées par une autorité éditoriale ou par une volonté de respect d'une certaine normativité. À titre d'illustration, une étude de (Ollinger et Valette, 2010) témoigne de la fertilité néologique des blogs. La productivité en néologie formelle de deux corpus, un corpus de presse et un corpus de blogs, y est estimée à partir d'une plateforme prototypique détectant des candidats à la néologie formelle et catégorielle, puis à l'aide du calcul d'un indice lexicométrique de richesse néologique. Les résultats témoignent d'une « richesse néologique sensiblement plus importante » dans les blogs que dans le corpus de presse. Bien que les auteurs s'interrogent, à juste titre, sur la légitimité des blogs comme support d'exploration en veille lexicale, ces questionnements ne doivent pas amener à nier l'impact des blogs dans le processus d'innovation lexicale de la langue : les blogs participent de pratiques discursives qui contribuent à la circulation du dire, ils contribuent donc à stimuler la dynamique néologique.

L'époque actuelle se caractérise également par un relâchement au niveau des frontières. Cet assouplissement des frontières est double. Il se manifeste d'abord entre différentes langues, à travers l'internationalisation des échanges. À travers l'étude des internationalismes, c'est-à-dire des néologismes qui émergent conjointement à plusieurs langues sur la scène internationale tels que le verbe provenant de l'anglais *to chat* (*chatter* en français, *chatten* en allemand,

chatear en espagnol, …), (Petralli, 1999) souligne que les nouveaux modes de communication permettent une circulation immédiate et mondiale de l'information, où l'étude des néologismes multilingues est légitime car les échanges internationaux se démocratisent et l'évolution de la langue se construit à travers l'amplification du brassage des usages de cette langue, internes mais surtout externes à une nation donnée.

L'assouplissement des frontières est également sensible entre langue de spécialité et langue générale. Ainsi, (Rousseau, 2010) met en lumière le renforcement de l'innovation lexicale aussi bien dans les langues de spécialité que, par interaction avec celles-ci, dans la langue générale :

> « En effet, non seulement assistons-nous à l'accélération de l'innovation lexicale dans les technolectes, mais nous pouvons constater la "terminologisation" croissante de la langue générale consécutive à l'apparition de nouveaux objets, technologiques ou non (…). Il s'agit là d'un changement majeur depuis les 20 dernières années. Et ce mouvement semble être en accélération. »

Cette terminologisation apparaît par exemple dans le blog de (Martinez, 2011)[8], qui répertorie les mots nouveaux intégrés au Petit Robert et au Petit Larousse chaque année. Martinez signale l'introduction de nombreux termes de la médecine dans le *Petit Robert 2008*, tels que *hémorroïdaire* ou *hypoparathyroïde* ; dans le *Petit Robert 2010*, les termes de la chimie (*protolyse* par exemple) occupent une part importante des nouveautés et l'intégration de recommandations officielles d'organisme de terminologie est notable.

1.3.3 Accélération des échanges et du temps d'implantation

Conséquence du relâchement des contraintes associé aux pratiques discursives précédemment citées et à la mondialisation des échanges, l'évolution de la langue s'accélère (Astrid, 2010). Les nouveaux modes de communication permettent un échange instantané de l'information, dont témoignent les exemples particulièrement frappants des SMS ou de l'outil de micro-blogging Twitter. La circulation du dire est de plus en plus rapide, ce qui accroît notamment la vitesse de diffusion de nouveaux emplois. Cette accélération se manifeste dans le cas particulier de la « technologisation » de la langue générale, qui correspond à un « changement majeur depuis les 20 dernières années. Et ce mouvement semble en accélération » (Rousseau, 2010). De façon plus générale, l'accélération de l'évolution de la langue se traduit, au niveau des unités lexicales, par une diminution du temps d'implantation des néologismes : (Pruvost et Sablayrolles, 2003:36-37) estiment que la durée d'implantation a évolué d'une dizaine d'années dans les années 70 à moins de cinq ans depuis les années 2000. Cette estimation n'est pas un absolu et ne se veut pas comme telle : elle est une donnée approximative, susceptible de varier d'un néologisme à l'autre. Elle reflète cependant une tendance globale, que des experts amenés à observer un grand nombre d'exemples ont perçue.

Pour conclure, la veille lexicale présente un intérêt croissant car l'émergence de néologies, en particulier de néologies sémantiques, participe d'un mouvement :

- qui exige de gérer un foisonnement croissant de données ;
- qui s'amplifie en langue générale par effacement des frontières territoriales et des frontières avec certains champs terminologiques ;
- qui s'accélère.

[8] A l'adresse http://orthogrenoble.net/page-de-camille-club-orthographe-grenoble.html

2. Définition de la néologie sémantique

La néologie sémantique constitue un sous-ensemble de la néologie. Les développements qui suivent cherchent à mettre en évidence des paramètres fondamentaux pour la néologie : le rôle du temps et la diffusion dans les discours.

2.1 Une variation en diachronie : rôle fondamental du temps

2.1.1 Un processus

Deux termes, *néologie* et *néologisme*, ont dominé historiquement pour renvoyer à deux facettes d'un même phénomène. Ils renvoient à une notion évolutive (Sablayrolles, 2000:57-65), mais qui s'est stabilisée à partir de la deuxième moitié du XX^{e} siècle autour de deux acceptions principales, telles que proposées par Guilbert (Guilbert, 1977, cité par Mejri, 1995:41) : la néologie correspond au « processus de création lexicale inhérente au système linguistique et au développement de la société » et le néologisme au « résultat de ce processus ».

L'idée de processus donne à voir la néologie comme un mouvement qui s'inscrit dans le temps. Il ne s'agit donc pas d'un événement ponctuel, même si le processus ponctuel constitue un cas limite, mais acceptable, de ce qu'est un processus. De plus, la présence récurrente de *nouveau* ou *nouvellement* dans les définitions de la néologie, ainsi que l'étymologie (*neos*, grec, "nouveau") positionnent la néologie par rapport à un état antérieur situé dans le temps.

L'étude de la néologie apparaît comme indissociable de la question du temps, donc d'un positionnement en diachronie.

2.1.2 Liens entre néologie, figures de style et changement lexical

La définition de la nouveauté peut recouvrir une échelle de temps plus ou moins large. Elle s'étend de l'événement ponctuel que représente l'hapax à des répétitions qui s'étalent suffisamment dans le temps pour autoriser une lexicalisation. Dans une définition large de la néologie, la durée est donc susceptible de recouvrir une large gamme de phénomènes : en limite inférieure, elle peut correspondre à un accident ou, à l'inverse, pour la limite supérieure, elle peut être le témoin d'une implantation en cours, voire achevée.

Deux pôles se distinguent par la place qu'ils accordent à la durée : l'un relègue en arrière-plan la question de la durée, il peut s'accommoder d'événements occasionnels et renvoie notamment à la problématique des figures de style, dont l'étude n'exige pas une répétition dans le temps sur divers énoncés ; l'autre est indissociable d'une certaine étendue temporelle et renvoie à la problématique du changement lexical. Or ces deux problématiques, celle de la tropologie, ou étude des figures de style, et celle du changement lexical, se doivent d'être dissociées (Sablayrolles, 2003). Certes, le changement lexical et les tropes entretiennent des liens, car les tropes font partie des procédés à l'origine du changement lexical. Cependant, certains changements lexicaux ne s'expliquent par les tropes, qui génèrent un sentiment de rupture ou, du moins, d'écart marqué, mais par des glissements plus progressifs :

> « Les figures de la rhétorique constituent des écarts intentionnels, repérables et significatifs dans un énoncé particulier, alors que les évolutions de sens, qui sont collectives et passent le plus souvent inaperçues des membres de la communauté, n'obéissent à aucune stratégie expressive. On ne les identifie souvent qu'après, par comparaison avec d'autres états de langue. » (Sablayrolles, 2003:117)

Ainsi, dans la presse, le sens d'*éponyme* a évolué d'un lien de filiation (le roman ou personnage éponyme est celui dont provient le nom) à un lien d'analogie (partage d'un même nom, sans lien de filiation)[9].

Notre perspective est celle de la veille lexicale, donc du changement sémantique. De là, nous privilégierons une définition restreinte de la néologie, qui inclut une certaine durée et qui est témoin d'une implantation en cours. Les tropes peuvent constituer des procédés à l'origine de la néologie, mais il convient de se libérer de ce cadre. Une étude en relation étroite avec les tropes serait pertinente pour analyser des hapax sémantiques, où les liens avec les figures de style sont beaucoup plus étroits, mais hors du champ de définition que nous adoptons.

2.2 Diffusion des néologismes

Cibler les néologies en cours d'implantation n'exige pas seulement de s'appuyer sur un critère de répétition dans la durée, mais également de prendre en compte la diffusion du néologisme, que (Sablayrolles, 2000:200) qualifie de « circulation du dire ».

La circulation du dire reste une notion relative et floue. Elle dépend de la communauté relativement à laquelle on se positionne. Ainsi, une lexie pourra être stabilisée et intégrée dans un vocabulaire de spécialité, mais elle aura un caractère de nouveauté en langue générale. Par exemple, avec la crise financière de 2008, les *subprimes* ont fait massivement leur apparition dans le vocabulaire des médias, alors que ce nom était déjà intégré à la terminologie de la finance.

Plusieurs critères sont susceptibles de caractériser la diffusion du dire : le nombre de locuteurs, la variété des sources d'émission, le nombre et la diversité des récepteurs potentiels, etc. Néanmoins, la définition précise de la circulation du dire à travers une sélection de critères dépend de l'objet d'investigation. À titre d'illustration, considérons le cas de *lumière* chez le poète Gaspar (Gérard, 2010). Le sens de lumière a un caractère matériel dans l'idiolecte de Gaspar. Ce sens matériel revient de façon récurrente dans ses œuvres poétiques. Si on prend le critère de diversité des œuvres, il y a bien circulation du dire, où le dire correspond à l'emploi de *lumière* comme entité tangible. En revanche, si on élargit le champ d'observation à un cercle de poètes contemporains, évaluer la circulation du dire nécessiterait de prendre en compte la diversité des individus. Pour étendre encore le champ, par rapport au français courant, la lumière matérielle de Gaspar ne s'est pas imposée, elle reste un micro-phénomène qui reste cantonné à un champ très restreint de la poésie. Autrement dit, au regard du français courant, il n'y a pas de diffusion de la lumière matérielle. En revanche, pour le *spleen* de Baudelaire, on constate un franchissement de la frontière de la poésie pour se diffuser dans la langue générale, comme en témoigne l'emploi de spleen répertorié dans le TLFi, même si le sens de *spleen* reste rattaché à la pratique littéraire.

Le terrain d'étude que nous retenons est celui du français courant. La circulation du dire associé à une néologie en cours d'implantation requiert, dans ce cadre, une multiplicité de locuteurs ou d'interlocuteurs. Cet argument rejoint le point de vue d'Alain Rey (Mejri, 1995:45) qui se positionne d'un point de vue lexicographique et pour qui un néologisme est significatif que s'il n'est pas associé à des créations individuelles, mais à une pratique collective.

Si certains indices sont susceptibles de témoigner d'une diffusion marquée d'un néologisme, ils ne sont en rien garants d'une implantation pérenne de ce néologisme. (Sablayrolles, 2000:200 sqq) souligne en effet que le devenir d'un néologisme est imprévisible et qu'il existe

[9] Source : blog *Langue Sauce Piquante*, 17 août 2008, à l'adresse http://correcteurs.blog.lemonde.fr/2008/08/17/les-puristes-sont-ils-des-losers/.

des destins très variables : un néologisme peut connaître une diffusion extrêmement massive, et pourtant de courte durée ou il sera prédit comme mort-né, mais parviendra pourtant à la lexicalisation, comme c'est le cas d'*inexorabilité*, néologisme annoncé sans avenir au XVII^{e} siècle (Sablayrolles, 2000:201). Si le devenir est incertain, cibler un stade de diffusion de manière relative semble donner plus de chance d'obtenir un candidat à la lexicalisation que de cibler un néologisme naissant : par sa diffusion même, le néologisme acquiert une certaine significativité et n'est plus de l'ordre de l'accident.

La restriction de la définition de la néologie à de la néologie en cours d'implantation revient également à exclure du champ d'étude certains types de néologismes et orientera le choix des données. En effet, certaines sources sont plus à même de refléter une diffusion en français courant ou d'y contribuer. Ainsi, les néologismes littéraires sont souvent le fruit d'un travail poussé du sens d'une lexie, associé à une volonté de créativité et une recherche d'expressivité (Mejri, 1995:119). Or ces néologismes restent généralement des hapax, ils ne sont pas réemployés ultérieurement (Sablayrolles, 2000:166). Par exemple, l'auteur contemporain William Gibson utilise un vocabulaire propre à son livre *Identification des schémas*, avec des lexies telles que *retard d'âme*, *a-sérotonie*, *monde-miroir*, mais ces néologismes restent propres à son ouvrage et à l'univers particulier associé à son personnage principal, qui porte un regard atypique sur le monde. Ces néologismes n'ont pas vocation à se diffuser, et ont une certaine probabilité de rester cantonné à l'ouvrage de l'auteur. À l'inverse, le discours journalistique aura tendance à refléter des néologies en cours de diffusion ou à participer à une diffusion relativement large. On reviendra ultérieurement sur la question du choix de données discursives.

En résumé, on emploiera par la suite le terme de "néologie" pour renvoyer à ce qui n'est, à strictement parler, qu'une restriction de l'ensemble des phénomènes qu'on pourrait qualifier de néologie. Il s'agira des néologies en cours d'implantation, à la fois à travers une diffusion dans le temps et à travers une diffusion dans la communauté ciblée, supposée représentative du français courant. La néologie qui fera l'objet de notre propos se caractérise donc par une significativité dans l'usage, qui se traduit par des réemplois dans le temps et au sein d'une communauté.

La néologie que nous ciblons se caractérise donc par une diffusion dans le temps et dans les discours. Il reste à préciser comment se distingue la néologie sémantique, objet de nos investigations, par rapport à l'ensemble des néologies.

3. Détection et allocation de signifié pour délimiter la néologie sémantique dans le champ de la néologie

Nous aborderons sous deux angles la délimitation de la néologie sémantique dans le champ des néologies : d'une part, sa détection ; d'autre part, sa caractérisation à travers la question de l'allocation de signifié.

3.1 Tripartition classique des néologies : emprunt, néologie de forme et néologie sémantique

Les néologies sont régulièrement classées selon trois classes principales : les emprunts aux langues étrangères (par exemple, le *step*, activité de fitness), les néologies de forme (*désescalade* en alpinisme, *cosmétotextile*), les néologies de sens (les *elliptiques,* appareils de fitness entre vélos et tapis de course). Les néologies de forme recouvrent les signifiants nouveaux, obtenus notamment par procédés morphologiques, mais qui ne proviennent pas

d'une langue étrangère. Les néologies de sens recouvrent les changements sémantiques, c'est-à-dire les nouveaux signifiés affectés à des signifiants déjà existants dans la langue.

Cette tripartition a été remise en question, pour deux raisons principales :

- Elle ne recouvre pas l'ensemble des phénomènes néologiques (Sablayrolles, 2010). La phraséologisation, c'est-à-dire l'apparition de nouveaux syntagmes contraints, ou nouveaux *phrasèmes*, comme par exemple *cuisine moléculaire* (phrasème entré dans le Petit Robert en 2010), n'apparaît pas explicitement dans le classement. Ce phénomène peut se rattacher à la néologie de forme, puisqu'il est fondé sur l'apparition d'une nouvelle lexie complexe, donc d'un nouveau signifiant, ou il peut être associé à la néologie de sens, car les lexies simples qui la composent sont des signifiants connus, ou encore il peut se concevoir comme un type de néologie distinct de l'emprunt, de la néologie de forme et de la néologie de sens. De même, les néologismes syntaxiques, par exemple par acquisition ou d'alternances syntaxiques comme *halluciner* (évolution de l'emploi transitif vers l'emploi absolu) semblent absents de la tripartition classique.
- Elle ne met pas en évidence l'existence de couplages ou d'interactions entre les différentes classes. Par exemple, des mots tels que *biodesign*, *coacher* ou *chatteur*[10] cumulent emprunt et néologie par procédé morphologique. (Gévaudan, 2002) montre ainsi que les néologies de formes obtenues par procédés morphologiques peuvent s'accompagner ou non d'un changement de sens par rapport aux morphèmes qui interviennent. Ainsi, *bravitude* est une néologie de forme en raison de la suffixation en *–itude*, mais présente aussi une néologie de sens en cela qu'il s'accompagne d'une connotation politique, plus précisément rattachée au parti socialiste et à Ségolène Royal, et se distingue en cela de *bravoure* ; en revanche, *judiciariser*[11], défini par le Larousse comme "confier à la justice le contrôle d'une situation, l'exécution d'une procédure", a un sens qui rejoint celui obtenu par décomposition morphologique et par analyse sémantique de la suffixation en *-iser*. *Bravitude* témoigne d'un couplage entre néologie de sens et néologie de forme que (Gévaudan, 2002) qualifierait de néologisme morpho-sémantique, que ne présente pas *judiciariser*, qui, selon sa terminologie, serait simplement un néologisme morphologique. Au contenu sémantique accessible par construction morphologique peut donc s'ajouter de façon plus ou moins marquée un contenu sémantique propre aux emplois discursifs.

La néologie sémantique présente une multiplicité de facettes et elle entretient des interactions avec les autres types de néologies qui définissent des imbrications complexes. Il semble nécessaire de préciser ce que recouvre la néologie sémantique et de la positionner par rapport aux autres types de néologie, afin de déterminer dans quels cas elle permet un accès au sens de façon autonome et dans quels cas elle offre une complémentarité par rapport à d'autres types de néologies pour accéder au sens. Cette analyse sera construite à partir de l'analyse des typologies de néologies de (Sablayrolles, 2000: 71-100).

3.2 Articulation entre néologie sémantique et les typologies de néologies

(Sablayrolles, 2000:71-100) met en évidence la richesse en nombre et en diversité des typologies de néologies, à travers un examen d'une centaine de typologies environ. Certaines d'entre elles sont difficilement comparables et l'ensemble qu'elles constituent n'est pas unifiable. Toutefois, des tendances communes se dégagent.

[10] Ces trois exemples ont été entrés successivement dans le *Petit Robert* et le *Petit Larousse* comme néologismes, respectivement en 2001 et 2005 pour *biodesign*, en 2002 et 2007 pour *coacher*, et en 2003 et 2009 pour *chatteur*.
[11] Introduit en 2002 dans le Petit Larousse.

3.2.1 Dominantes des typologies existantes : perspective lexicologique et structuration selon les procédés

Les typologies varient selon le cadre de travail des différents auteurs, qui privilégient plusieurs perspectives : la lexicologie, l'histoire, la littérature, la terminologie, la sociolinguistique ou la psycholinguistique. La lexicologie est le point de vue dominant, et il s'agit également de celui que nous adoptons, en accord avec la perspective de veille lexicale. La lexicologie est choisie comme point de vue non pas à l'exclusion de la morphologie et la syntaxe, mais dans son articulation avec celles-ci. La conformité entre notre approche et le courant dominant des différentes typologies de néologies nous invite à nous appuyer sur les travaux de (Sablayrolles, 2000).

Le profil des typologies varie en fonction de la perspective des auteurs, mais aussi en fonction des critères privilégiés pour structurer les typologies, partiellement corrélés à la perspective des auteurs. Ces critères constituent un ensemble complexe (Sablayrolles, 2000:77-78), ils sont exposés, en fonction des auteurs, selon des degrés d'explicitation variables. Néanmoins, le critère des procédés (procédé morphologique par affixation par exemple ; procédé sémantique par métaphore ou encore par métonymie ; etc.) domine. Ce critère des procédés sert de base à la typologie de synthèse proposée par (Sablayrolles, 2000:245), qui s'appuie sur les matrices lexicogéniques de (Tournier, 1985, cité par Sablayrolles 2000: 239) et les adapte à son cadre d'étude. Ces matrices sont organisées sour forme de tableaux qui structurent les néologies en fonction des procédés par catégories enchâssées, de degré de généralité croissant. Le tableau de synthèse des matrices lexicogéniques établi par Sablayrolles (Sablayrolles, 2000:245 ; une version adaptée est présentée au (3.2.2, tableau I.2.1)) présente une structure à plusieurs niveaux, qui propose des classes de précision variable ou, inversement, de degré de généralité variable. On s'appuiera principalement sur cette structure pour positionner notre objet d'étude, la néologie sémantique.

Enfin, les typologies se distinguent par leur degré d'approfondissement, c'est-à-dire de niveaux enchâssés, et de finesse, c'est-à-dire de détail de description pour chaque niveau. Le degré de développement peut être plus ou moins poussé sur l'ensemble de la typologie, mais il peut aussi être plus marqué sur seulement certains sous-ensembles de la typologie, lorsque l'auteur se focalise sur un type de néologie particulier. Quelques auteurs ont centré leur approche sur la néologie sémantique. Parmi eux, (Bastuji, 1974) se distingue et a retenu notre attention, car sa démarche adopte le point de vue lexicologique et sa structuration repose sur le critère des procédés (Sablayrolles, 2000:82), conformément à ce que nous privilégions. Sa position, articulée autour de la néologie sémantique, l'a amenée notamment à tester les limites de définition de ce type de néologie et à se pencher sur deux questions pour délimiter le phénomène : celle du rapport entre signifiant et signifié et celle de la place des procédés syntagmatiques.

La typologie de synthèse de (Sablayrolles, 2000) servira de base à notre démarche. Cette typologie est générale, elle ne cherche pas à privilégier un type particulier de néologie. Nous chercherons à la recentrer sur la néologie sémantique, démarche que nous étayerons en particulier avec l'approche de (Bastuji, 1974).

3.2.2 Procédés : orientation de la focalisation vers la néologie sémantique

Dans les typologies, aussi bien la typologie de synthèse de (Sablayrolles, 2000:245 ; *cf.* tableau *infra*) que des typologies à l'origine de la synthèse, observées dans le détail, la néologie sémantique inclut des emplois figurés, par métaphore et par métonymie ; elle se distingue des néologies où un nouveau signifiant est obtenu par procédés morphologiques.

Elle exclut la plupart du temps les emprunts, du moins les emprunts aux langues étrangères. La tripartition classique définit donc trois ensembles distincts, qui apparaissent comme disjoints, et, en cela, elle garde une certaine légitimité. Cependant, ces trois ensembles constituent moins une partition que le noyau de groupes susceptibles de constituer une partition : si l'on vise l'exhaustivité et l'obtention d'une partition, ces trois ensembles doivent être élargis à des néologies dont le type de rattachement n'est pas immédiat. Les frontières entre ensembles sont donc marquées par un certain flou.

La typologie de synthèse de (Sablayrolles, 2000:245) se présente sous forme d'un tableau à plusieurs niveaux, celui des matrices lexicogéniques, présentée ci-dessous (*cf.* figure 1.2.1) avec quelques ajustements mineurs (ajouts d'exemples principalement).

Ce tableau des matrices considéré isolément définit, par son format même, des partitions pour chaque niveau, qui ne font pas ressortir les incertitudes de délimitation. Celles-ci sont présentes dans les commentaires explicatifs de (Sablayrolles, 2000: 207-245), qui insiste sur les nuances et les fluctuations possibles : la structure de la matrice résulte de choix, souvent délicats, issus de prises de position face à un ensemble complexe et ne pouvant faire l'objet d'un consensus unanime.

<table>
<tr><td>préfixation</td><td>démariage[12]</td><td rowspan="4">affixation</td><td rowspan="4">construction</td><td rowspan="8">morpho-sémantique</td><td rowspan="15">matrices internes</td></tr>
<tr><td>suffixation</td><td>péleriner</td></tr>
<tr><td>dérivation inverse[13]</td><td>glisse (issu de glisser) ; orater (issu d'orateur)</td></tr>
<tr><td>parasynthétique[14]</td><td>dépigeonnisation</td></tr>
<tr><td>composition dont
- synapsie[15]
- quasimorphème[16]</td><td>irakophile
- carte mer
- oenothèque</td><td rowspan="2">composition</td></tr>
<tr><td>mot-valise</td><td>auto-stérilisation</td></tr>
<tr><td>onomatopée
fausse coupe et jeu phonique : paronymie</td><td>spychique
pendicite (la)</td><td colspan="2">imitation et déformation</td></tr>
<tr><td>conversion</td><td>notoire (un)</td><td colspan="2" rowspan="2">changement de fonction</td><td rowspan="5">syntactico-sémantique</td></tr>
<tr><td>construction différente</td><td>syndrome infirmières (aussi rattaché à la composition)</td></tr>
<tr><td>métaphore</td><td>tankers (du parti chiraquien) [remarque : souvent couplé à de la composition]</td><td colspan="2" rowspan="3">changement de sens</td></tr>
<tr><td>métonymie</td><td>verdissement artificiel [appliqué au domaine politique par un emploi figuré : symbole des écolos]</td></tr>
<tr><td>autres figures, restriction de sens, extension de sens, etc.</td><td>dromomanie [passage d'un domaine spéc. (en psychiatrie "impulsion morbide à marcher ou à courir") à un article culturel de critique picturale]</td></tr>
<tr><td>troncation</td><td>quadra (pour quadragénaire)</td><td colspan="2">réduction de la forme</td><td>morphologique</td></tr>
</table>

[12] Les exemples ont été relevés chez (Sablayrolles, 2000:443-533).

[13] obtention d'une nouvelle lexie par suppression et non ajout d'un affixe, pouvant aboutir à une nouvelle catégorie grammaticale.

[14] construction morphologique par préfixation et suffixation simultanées

[15] constitution d'une nouvelle lexie à partir de plusieurs lexies autonomes jointes par des prépositions (Sablayrolles, 2000:223)

[16] Lexies simples constituées d'éléments correspondant à des lexies simples dans la langue d'origine, comme *path(o)*, plus mobiles que des affixes. Le terme quasimorphème désigne chez Sablayrolles les composés dont les éléments sont issus de langues anciennes.

siglaison	SDF			
détournement	*est pas un long fleuve tranquille (la publicité)*	**pragmatique**		
emprunt	*suburbes (adolescent des)* ; *canyoning*	**matrice externe**		

Tableau I.2.1 : Tableau des matrices lexicogéniques présentant une typologie des néologies en fonction des procédés de formation (Sablayrolles, 2000:245).

Avant de discuter plus en détail des fluctuations et de centrer la structure sur la néologie sémantique, nous proposons une première réorganisation partielle de la matrice de Sablayrolles, afin d'être en accord avec notre positionnement théorique.

Suite à notre remarque sur la tripartition classique, nous proposons de préciser ce que recouvrent les emprunts. Une première distinction consiste à dissocier les emprunts susceptibles d'être identifiés comme tels et ceux qui ne le sont pas. Les emprunts non identifiés comme tels sont ceux qui font l'objet d'une francisation ou les calques, comme *haut-parleur* pour *loud-speaker* (Sablayrolles, 2000:234) ou comme *initier* au sens de 'débuter, commencer', issu de l'emploi anglais *to initiate*. Le signifié provient d'une autre langue, mais le signifiant est connu ou analysable en langue française par procédé morphologique. Nous n'incluons pas ce type d'emprunts dans la catégorie 'emprunts', nous y associons seulement ceux dont le signifiant n'est pas présent dans la langue de référence, ou dont les constituants obtenus par décomposition n'appartiennent pas à la langue de référence. Autrement dit, la définition de la catégorie 'emprunts' repose implicitement sur un critère d'existence du signifiant dans la langue de référence, donc sur critère de sa détectabilité. Parmi ce qui est retenu dans la catégorie des emprunts, nous proposons une dissociation entre emprunts externes, provenant de langues étrangères, et emprunts internes, au sein de la langue française, c'est-à-dire incluant toutes ses variations linguistiques, mais hors de la langue de référence (celle-ci peut être, par exemple, une langue de spécialité, ou au contraire exclure certaines terminologies trop spécialisées).

Le deuxième axe de réorganisation concerne les procédés morphologiques. Sablayrolles choisit de privilégier le courant traditionnel en morphologie, il n'adopte pas les positions préconisées par (Corbin, 1987) mais s'inscrit dans le courant de la morphologie morphématique : le morphème est l'unité de base, il correspond à une unité phonologique et sémantique précise ; une unité lexicale et en particulier son sens s'obtiennent par composition de morphèmes. À l'inverse, la position que nous retenons respecte le courant lexématique : l'unité considérée est le lexème, qui se décompose en trois niveaux (phonologie, catégorie grammaticale et sens) ; les changements morphologiques s'analysent comme des opérations à chacun des niveaux. Dans ce cadre, la dérivation inverse et la construction parasynthétique sont analysées selon d'autres schémas de fonctionnement, donc les mécanismes en jeu sont décrits à travers d'autres procédés. La dérivation inverse est associée à la conversion et la construction parasynthétique s'analyse en terme d'opérations successives et non pas simultanées. Par exemple, elle résultera d'une suffixation suivie d'une préfixation. De plus, la conversion sera considérée comme un procédé morphologique et non pas syntaxique. Certes, le changement de catégorie grammaticale implique des changements syntagmatiques, mais l'unité que nous considérons est le lemme, qui comporte une étiquette de catégorie. Un changement de catégorie affectera donc le lexème.

Au niveau de la composition, on privilégiera le palier de la lexie simple. Les synapsies, qui correspondent à la constitution de lexies complexes, et les quasimorphèmes, qui sont des lexies simples, seront dissociés. La constitution d'une nouvelle lexie complexe, ou phraséologisation, sera considérée comme un phénomène syntagmatique, donc syntactico-

sémantique. Cette catégorie intègrera les détournements, qui correspondent aussi à l'apparition d'une nouvelle lexie complexe, même si celle-ci provient d'une autre lexie complexe.

Ces considérations amènent à réorganiser le tableau des matrices comme suit :

<table>
<tr><td>emprunts externes</td><td colspan="2">emprunts externes (emprunts : acception restreinte)</td><td rowspan="2">emprunts (emprunts : acception large)</td><td rowspan="2">matrice externe (hors langue représentée par la ressource de référence)</td></tr>
<tr><td>emprunts internes</td><td colspan="2">emprunts internes</td></tr>
<tr><td>métaphore</td><td colspan="2" rowspan="3">changement de sens (néologie de sens : acception restreinte)</td><td rowspan="3">syntactico-sémantique (néologie de sens : acception large)</td><td rowspan="12">matrice interne (langue représentée par la ressource de référence)</td></tr>
<tr><td>métonymie</td></tr>
<tr><td>autres figures, restriction de sens, extension de sens, etc.</td></tr>
<tr><td>construction différente</td><td colspan="2" rowspan="2">changement syntagmatique</td></tr>
<tr><td>phraséologisation</td></tr>
<tr><td>conversion</td><td colspan="2">changement morpho-syntaxique</td><td rowspan="8">morpho-sémantique (néologie de forme : acception large)</td></tr>
<tr><td>mot-valise</td><td rowspan="2">composition</td><td rowspan="4">construction (néologie de forme : acception restreinte)</td></tr>
<tr><td>autres formes de composition</td></tr>
<tr><td>préfixation</td><td rowspan="2">affixation</td></tr>
<tr><td>suffixation</td></tr>
<tr><td>troncation</td><td colspan="2" rowspan="2">réduction de la forme</td></tr>
<tr><td>siglaison</td></tr>
<tr><td>onomatopée, fausse coupe et jeu phonique : paronymie</td><td colspan="2">changement morpho-phonologique</td></tr>
</table>

Tableau I.2.2 : Tableau des matrices lexicogéniques réorganisé

Partant de ce tableau, on cherche à centrer la structure sur la néologie sémantique et mettre en évidence le flou des frontières.

Au niveau des procédés, la néologie sémantique dans sa définition minimale correspond au changement de sens, qui recouvre la métaphore, la métonymie et les autres figures.

La définition de la néosémie n'est pas stricte, elle pourrait être élargie à partir du palier d'observation supérieur à ce qui est qualifié de syntactico-sémantique. Cet élargissement regroupe les deux volets mis en évidence par (Bastuji, 1974), qui cherche à définir la néologie sémantique :

> « Le néologisme de sens n'est rien sans ses règles d'insertion lexicale dans la phrase et/ou le syntagme ; il n'est rien non plus sans le discours – ou l'interdiscours – où il prend son sens. » (Bastuji, 1974:7)

La néologie de sens peut donc résulter de modifications syntagmatiques, par modification de règles syntaxiques ou phraséologisation, ou de ce qui est qualifié de 'changement de sens' dans la matrice précédente, qui se manifeste par un changement de l'environnement discursif. Dans le prolongement de cette distinction, (Bastuji, 1974) propose deux mécanismes pour décrire les modifications provenant du syntagmatique et du discursif, qui sont respectivement la modification de traits de sélection[17] et la modification de traits inhérents[18].

[17] Traits sémantiques qui définissent les règles de combinatoire syntactico-sémantiques.
[18] Traits distinctifs qui analysent la signification d'une unité, ou *sèmes* (Bastuji, 1974:7).

De même, (Mel'čuk, 2008) propose de voir deux stratégies pour pallier l'insuffisance de signifiants dans une langue, donc de créer de nouveaux signifiés à partir de signifiants connus : la phraséologisation et l'ambiguïsation.

> « ... à cause du nombre de signifiants (dans une langue naturelle) par rapport au nombre gigantesque et toujours croissant de signifiés, les syntagmes perdent souvent leur liberté pour porter un nouveau signifié "uni" : ceci est la phraséologisation, ou figement, de syntagmes (ex-)libres. » (Mel'čuk, 2008:189)

> « La phraséologisation n'est qu'une des deux stratégies dont la langue dispose pour pallier l'insuffisance de signifiants. L'autre est l'ambiguïsation, soit l'association de plusieurs signifiés à un même signifiant, ce qui crée de nouveaux signes. » (Mel'čuk, 2008:189)

Ce point de vue revient à considérer deux facettes complémentaires de la néologie, l'une qui repose sur le syntagmatique, en termes de combinaison d'unités et de perte d'autonomie, l'autre qui est purement sémantique et qui se constitue à partir d'une unité fixée.

La conversion, ou recatégorisation, est quelquefois considérée comme un procédé associé à la néologie sémantique. En particulier, (Bastuji, 1974:14-15) l'intègre à la néologie sémantique en y voyant l'effacement d'un ou plusieurs constituants, par un processus inverse de l'émergence d'une lexie complexe. Le parallèle avec la phraséologisation, avec la formation d'une nouvelle lexie à partir d'une lexie préexistante, peut légitimer le rattachement de la conversion à la néologie sémantique. À l'instar de Bastuji, d'autres auteurs évoqués par Sablayrolles, sur lesquels nous ne nous attarderons pas, choisissent d'associer la conversion à la néologie sémantique, tels que (Guilbert, 1975, cité par Sablayrolles, 2000:418), (Walter, 1984) ou encore (Tournier, 1985). La conversion peut effectivement s'accompagner d'un changement sémantique marqué mais que la morphologie ne suffit pas à expliquer, notamment dans le cas des ellipses (le *portable* pour *téléphone portable*). Ce point est développé dans le prochain paragraphe, où est abordée la question du degré de nouveauté.

Enfin, les emprunts ont de nombreuses facettes (Sablayrolles, 2000:232-237), partiellement évoquées lors de la première réorganisation de la matrice. Selon que leur définition est plus ou moins restreinte, les délimitations de la néologie sémantique varient. L'emprunt à une langue étrangère qui s'accompagne d'un nouveau signifiant se distingue nettement de la néologie sémantique. Par contre, l'intégration des calques à la néologie sémantique peut être contesté : le contenu sémantique qu'apporte la langue d'origine semble ainsi exclu, rattacher les calques à la néologie sémantique revient à rejeter l'apport sémantique que permettrait une ressource externe. Cependant, le fait que l'origine étrangère des calques ne soit pas toujours identifiée comme emprunt témoigne que l'accès au sens indépendamment de ressources externes est possible. Les ressources externes facilitent l'accès au sens dans ce cas, mais n'y sont pas indispensables. De façon similaire, les emprunts internes peuvent être ou non associés à de la néologie sémantique. Ainsi, (Guilbert, 1971 et 1975 cité par Sablayrolles, 2000) associe des changements sociologiques (changement de niveau ou de registre) à des néologies sémantiques, ou encore il considère les emprunts à un domaine apparenté comme de la néologie sémantique. De même, (DHLF 1992, cité par Sablayrolles, 2000) associe le passage d'une langue de spécialité à une autre à la néologie sémantique. Les changements de sens par emprunt à d'autres sociolectes ou langues de spécialité constituent donc des cas limites, susceptibles de relever de la néologie sémantique.

En bref, la définition de la néologie sémantique peut recouvrir un périmètre variable, aux frontières floues, qu'il est possible de résumer comme suit :

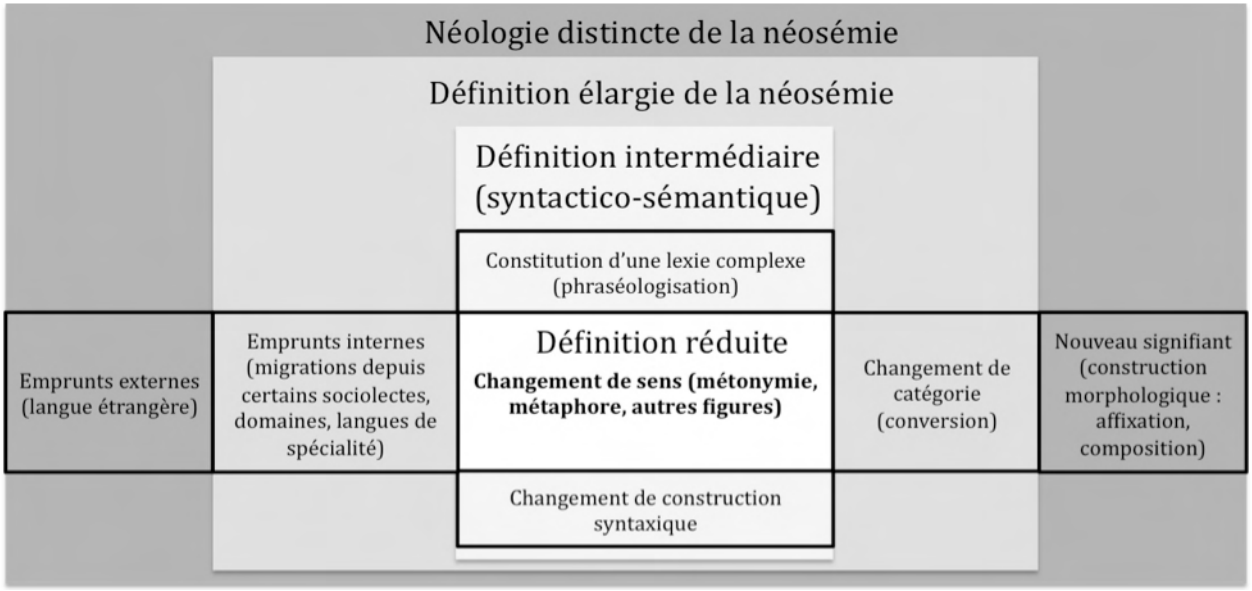


Figure I.2.3 : Articulation de la néologie sémantique aux autres types de néologie

Les fluctuations sont liées à des variations du cadre de référence à deux niveaux :

- Celui de la langue de référence : elle peut connaître des délimitations externes, par rapport à d'autres langues (langue française par opposition à langue anglaise, allemande, etc.) ou internes (exclusion de langues de spécialité ou de sociolectes).
- Celui de l'unité de référence : l'unité propre à la néologie sémantique est une lexie appartenant à la langue de référence, dont le signifiant garde des caractéristiques stables : catégorie grammaticale préservée, indépendance en tant qu'unité conservée. Cependant, si la lexie perd son statut d'unité indépendante en devenant composante d'une lexie complexe, elle peut aussi être considérée comme néologie sémantique. De même, le changement de catégorie peut être considéré comme n'affectant pas, ou peu, le signifiant, donc comme cas de néologie sémantique. Nous ne considèrerons pas la néologie catégorielle comme néologie sémantique, de par notre définition du signifiant, qui inclut la catégorie grammaticale.

D'une certaine façon, la typologie des néologies dessine une hiérarchie des procédés pour accéder au nouveau sens : la priorité est accordée à la décomposition morphologique pour les néologies constructionnelles ; la néologie catégorielle demande de privilégier un jeu sur les étiquettes grammaticales pour pouvoir naviguer dans la langue de référence et accéder au sens ; pour les néologies par emprunts, la construction du nouveau sens peut être obtenue à partir d'une ressource externe, associée à une autre langue que la langue de référence. La néologie sémantique au sens strict est celle où des outils et des ressources externes à la ressource de référence ne peuvent fournir la clé pour accéder au nouveau sens.

La typologie aiguille donc vers des procédés exploitables pour accéder au sens qui ne sont pas exclusifs, mais qui ont un degré de pertinence variable, et dont certains sont plus prometteurs que d'autres. En effet, la construction du sens peut mettre plusieurs procédés à l'œuvre. Ceci apparaît à travers la question du couplage et de l'articulation entre les différents types de néologie, qui se caractérise par des degrés de nouveauté sémantique variables.

3.3 Le degré de nouveauté sémantique : clé pour articuler et nuancer les types issus des procédés

Selon (Rey, 1974:16), la nouveauté sémantique est présente pour tous les types de néologie, que le signifiant soit nouveau ou connu. Le degré de nouveauté sémantique est corrélé au type de néologie, il est plus marqué pour l'emprunt que pour la néologie morpho-sémantique ou que par réduction de la forme :

> « Pour ceux qui présentent une nouveauté formelle, la nouveauté sémantique peut être concomitante et totale dans le système (emprunts), partielle (sémantisme de la préfixation, de la suffixation, de l'agglutination dans les mots complexes, sémantisme du syntagme dans les groupes de mots) ou très faible (le sigle, l'acronyme véhiculent le sens de l'expression qu'ils abrègent, mais en l'abrégeant, ils modifient ses connotations).
>
> S'ajoutent à ces néologismes, les cas de dérivation dite interne, c'est-à-dire de transfert fonctionnel, et les "néologismes de sens", c'est-à-dire les transferts sémantiques, qui peuvent être internes (album (de disques)) ou empruntés (cf. réaliser, au sens de to realize). »

Le degré de nouveauté sémantique dépend donc du procédé qui est à l'origine du néologisme. Indirectement, Rey souligne que l'accès au nouveau sens dépend du contexte de façon plus ou moins marquée selon le type de néologisme.

Cependant, (Gévaudan, 2002) nuance ces corrélations entre type et degré de nouveauté sémantique. Pour un type de néologie tel que décrit dans la sous-section précédente, il existe des différenciations internes. Un néologisme de forme, obtenu par procédé morphologique, pourra présenter un changement de sens que la décomposition morphologique ne suffit pas à expliquer. Cette différenciation rejoint d'une certaine façon le courant lexématique en morphologie. L'unité n'est plus le morphème, mais le lexème, qui constitue une entité abstraite à laquelle un mot-forme est associé. Ce mot-forme dépend de l'énoncé, donc du contexte, à travers des dépendances syntaxiques. Les procédés morphologiques restent pertinents pour construire un signifié, mais le signifié obtenu doit être validé à travers une confrontation au contexte. La description sémantique sera alors susceptible d'être ajustée ou amendée de façon plus ou moins marquée. (Namer *et al.,* 2007) ont réalisé des expériences dans ce sillage ; elles confrontent des informations obtenues par analyse morphologique à des informations provenant d'autres sources, et en retirent des éléments de validation et d'enrichissement. La différenciation entre un néologisme constructionnel de sens amendé par le contexte ou non apparaît chez (Gévaudan, 2002) à travers la distinction entre néologie morphologique et néologie morpho-sémantique. Cette distinction apparaît si l'on considère l'exemple précédemment cité de *bravitude* : le sens obtenu à partir de *brave* et du suffixe -*itude* ne permet pas d'obtenir les connotations liées au parti socialiste qu'on retrouve dans les emplois du néologisme, tandis que le sens obtenu pour *aptitude* à partir de *apte* par construction morphologique est en accord avec les emplois en discours *d'aptitude.* Dans ce cas, la néologie sémantique se combine à de la néologie constructionnelle. De même, la néologie sémantique peut se combiner ou non à un emprunt : le sens du verbe *supporter* en sport peut être obtenu directement à partir des emplois en anglais, analogues, mais, par distorsion, il peut aussi se déduire du verbe français *supporter* au sens de soutenir, être le point d'appui de, par exemple dans le bâtiment (supporter le poids de la charpente) ; inversement, le sens de *cake* provient de l'anglais, mais au sens anglais s'ajoute une restriction de sens pour ne désigner qu'un type précis de produit boulanger ou pâtissier, non les gâteaux en général.

L'apport contextuel a donc un rôle de modulation du signifié indispensable pour la néologie sémantique. Le signifié dont on dispose n'est pas adapté : travailler le signifiant et naviguer

dans la ressource de référence ne suffisent pas à expliquer le nouveau sens. Pour les autres types de néologie, l'apport du contexte existe, mais le signifiant est absent de la ressource de référence et, avant d'étudier l'apport du contexte, il semble profitable d'analyser le signifiant par d'autres moyens. Autrement dit, avant d'examiner l'influence du contexte sur le signifié, il semble judicieux d'associer un premier signifié au signifiant. Pour cela, l'application de règles (par exemple morphologiques) et le recours à des ressources externes priment. La néologie sémantique dépend donc le plus étroitement du contexte. En effet, sans contexte, la néologie passera inaperçue : c'est à travers les interactions entre le signifié normalement associé au signifiant et le contexte que la nouveauté apparaît et que le nouveau sens se construit. Ce qui est mis en œuvre pour accéder au sens en néologie sémantique peut aussi être exploité pour d'autres types de néologies, soit à partir d'un signifié vide, soit à partir d'un signifié obtenu par d'autres procédés dont on pourra vérifier la validité en contexte et qui pourra être modulé si nécessaire.

Le problème peut donc se généraliser à la question de l'allocation de signifié en fonction du contexte. Ce procédé sera bien sûr fondamental pour la néologie sémantique, mais il pourra aussi contribuer aux autres types de néologie.

Les fluctuations entre types peuvent prendre une autre forme. Cette autre forme de fluctuation conforte notre position et amène à se détacher de la question des types pour se centrer sur celle de l'accès au sens à partir d'une source donnée, en l'occurrence l'apport contextuel. Cette situation est celle des variations d'un même néologisme dans le temps entre différents types.

À titre d'illustration, considérons le cas du substantif *malbouffe*, dont l'évolution est décrite par (Boussidan *et al.*, 2009) : le néologisme est apparu sous forme de syntagme, *mal bouffe*, pour évoluer vers un mot-valise *mal-bouffe* en concurrence avec le composé morphologique *malbouffe*, puis ses emplois se sont stabilisés sur la forme construite morphologiquement, *malbouffe*. Plusieurs types de néologies se sont donc concurrencés au cours du temps, jusqu'à ce qu'une graphie s'impose. Cependant, les variations observées relèvent du même phénomène néologique, par-delà les différences de types.

Ajoutons qu'un néologisme peut être étendu d'une lexie isolée à un ensemble de lexies correspondant au même phénomène, mais qui se manifeste à travers les différents éléments d'une famille morphologique. Ce sera le cas du nom propre *Outreau*, étudié par (Lecolle, 2007b), dont le sens a évolué de celui de ville française du Pas-de-Calais vers celui de parangon du scandale judiciaire. Le changement de sens du nom propre *Outreau* s'est effectué à travers des emplois où d'autres types de néologies apparaissaient, à travers des emplois *d'Outreau en* tant substantif (nom commun ?), comme dans *un nouvel Outreau*, ou encore par néologisme constructionnel, avec l'emploi du verbe *outreauiser*. Le regroupement par famille de néologismes associés peut dépasser le cadre des familles morphologiquement proches : il s'agit des néologismes concurrents, participant du ***foisonnement néologique*** *transitoire* de (Dury, 2008 citant Guilbert, 1965) , par exemple la famille {tablette, ardoise, écran (tactile)*, pad}, dont les éléments se concurrencent pour désigner les micro-ordinateurs tactiles et plats, de format intermédiaire (grosso modo, du A3 au A6). Ce type de regroupement dépasse la question de la répartition en types.

Différents types de néologismes peuvent donc participer d'un même phénomène néologique. L'association à un type donné apparaît donc comme un préalable susceptible de fournir les clés d'analyse du néologisme, de fournir les points communs avec les signifiants apparentés et de permettre des regroupements à partir desquels pourra, dans un second temps, s'effectuer une allocation de signifié commune. Dans ces néologismes qui prennent corps à travers une

famille de lexies de type variable, à signifiant instable, l'apport contextuel est une constante, base commune aux occurrences des différentes formes participant du même phénomène néologique et susceptible d'apporter ce signifié commun aux représentants de la même famille.

3.4 Bilan

La tripartition classique emprunt / néologie de sens / néologie de forme est à relativiser : elle constitue une base, support d'une réalité plus complexe faite d'imbrications et de nuances. Il convient de conserver la définition de la néologie sémantique comme signifiant connu et évolution du signifié (par création d'un nouveau sens ou restructuration du signifié) et de se servir de ce critère pour la distinguer des autres formes de néologie.

Pour déterminer ce qu'est un signifiant existant, on s'appuiera sur le palier du lemme, qui constitue le point d'entrée des ressources lexicographiques classiques. L'élargissement aux lexies complexes ou, inversement, aux morphèmes, est possible mais tributaire des ressources et outils complémentaires dont on dispose, c'est-à-dire d'un dictionnaire de collocations dans le premier cas ou d'un analyseur morphologique dans le second cas. Autrement dit, l'obtention d'un signifié avant confrontation au contexte nécessite une étape intermédiaire.

Les néologies privilégiées seront celles pour lesquelles on dispose d'un signifié obtenu directement, c'est-à-dire à partir de la ressource de référence, sans passer par des outils ou ressources complémentaires. Les cas où les signifiés ne sont pas disponibles constituent des cas limites. Il est délicat de les qualifier de néologie sémantique. La distinction entre néologie sémantique et d'autres types de néologie reposera sur le mode de détection : si la détection s'opère parce que l'unité lexicale considérée est absente de la ressource de référence, la néologie ne sera pas qualifiée de néosémie. Cependant, les procédés mis en œuvre pour allouer un nouveau signifié pourront être élargis à d'autres types de néologie, à partir d'un signifié vide ou à partir d'un signifié obtenu indirectement (*cf.* figure 1.2.4 infra). Au demeurant, l'allocation de signifié à partir du contexte offre un regard complémentaire à l'allocation de signifié par construction morphologique : le contexte permet la validation et une restructuration ou un enrichissement éventuels du signifié construit morphologiquement.

On se focalisera sur la question de la contribution du contexte à l'allocation de signifié plus que sur la question de l'affectation ou du respect strict d'un type de néologie. La question centrale n'est donc pas d'attribuer une étiquette 'néologie sémantique', mais de voir comment le contexte contribue à construire un nouveau signifié lorsque l'unité lexicale est codée dans la ressource de référence, avec élargissement aux cas où un signifié est obtenu par d'autres procédés.

Le contexte peut avoir deux modes d'impact sur le signifié, syntagmatique ou sémantique. Les deux approches du changement de signifié sont complémentaires : l'une concerne la phraséologisation, l'autre reste sur un signifiant fixé, dont le contenu est susceptible de se restructurer ou de s'enrichir. Nous ciblons la deuxième approche, où l'influence du contexte est étudiée à travers les thématiques présentes, les domaines représentés ou encore l'environnement distributionnel dans ses variations de surface et ses constantes sémantiques.

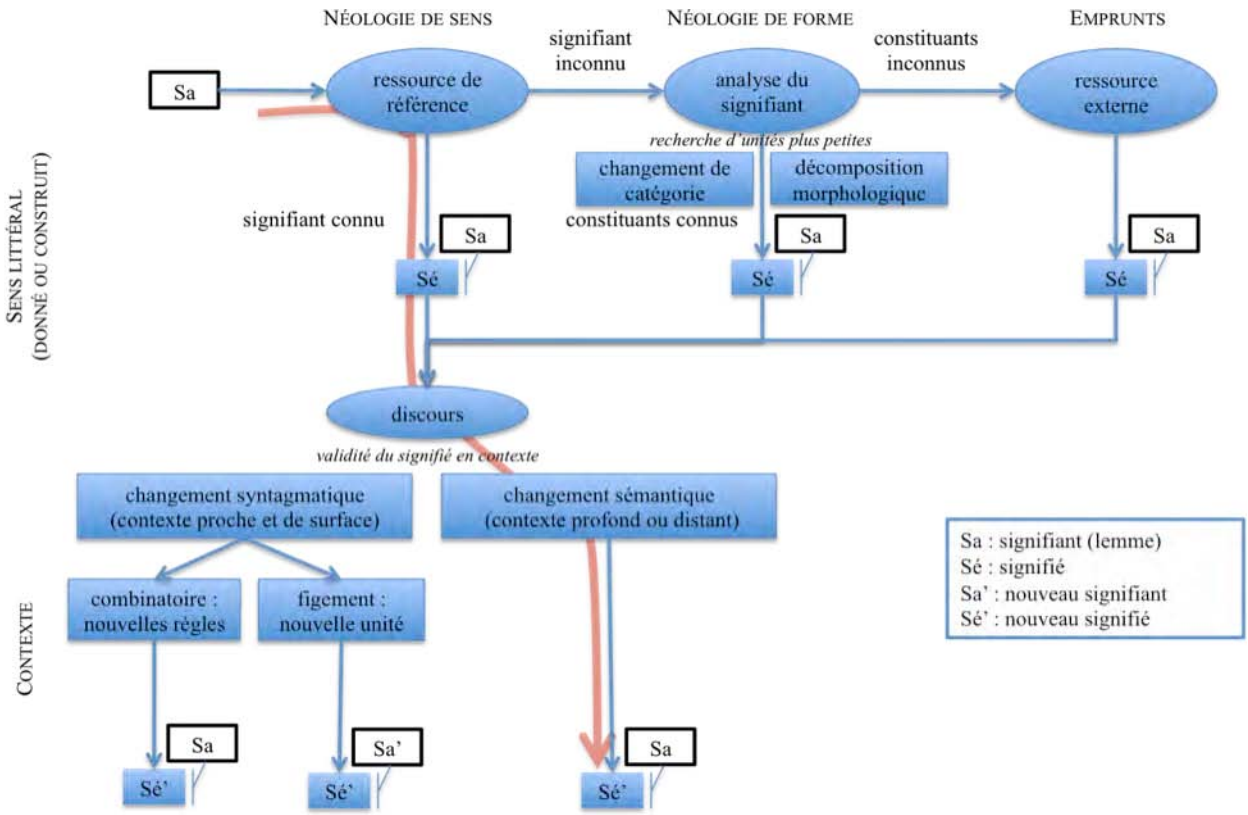


Figure I.2.4 : Articulation du traitement de différents types de néologie à celui de la néologie sémantique

Chapitre I.3
Éléments de modélisation de la néosémie

1. Cycle d'évolution : anticiper sans précipiter

Le lexique est en évolution permanente, cette évolution peut s'observer de façon systémique mais aussi unité par unité. Toute lexie a été néologique ou a connu un processus de néologisation au cours de son histoire. Elle n'est pas pour autant soumise de façon ininterrompue à des variations sémantiques marquées, susceptibles de changer son sens durablement. En particulier, sur de petits intervalles de temps, la plupart des lexies connaissent une relative stabilité sémantique. (Coseriu, 1973) exprime cette idée en ces termes :

> « Finalement, il y aurait, sans doute, une contradiction dans les termes – pour mieux dire, la langue ne pourrait en aucune façon se constituer -, si le changement linguistique était total et perpétuel, si un état de langue n'était *rien de plus* qu'un simple moment éphémère d'une « transition fuyante et fluctuation incessante » »

Ainsi, si on considère l'ensemble des unités lexicales, il y a une inertie du système, ou encore une homéostasie (Nyckees, 2000: 33). Pour la majorité des unités considérées isolément, les transitions sémantiques de ces unités n'ont pas lieu en permanence, mais ponctuellement. Il est donc légitime de supposer qu'une unité lexicale n'est pas soumise à un enchaînement ininterrompu de changements sémantiques, mais que ces changements s'inscrivent entre deux états stables, qui tendront à se maintenir. Une néosémie peut donc se voir, en première approximation, comme un processus qui se déroule selon un cycle en trois étapes, inscrit dans le temps, où interagissent langue et discours :

- Un état sémantique initial, dans lequel est codé le sens des lexies connues. Cet état sémantique est le reflet d'emplois stabilisés dans les discours.
- Un état transitoire, qui correspond à une transgression. Projeté en discours, le sens codé associé à l'unité ou aux unités qui constitueront la nouvelle lexie présente une anomalie. La transgression que représente cette anomalie se répète et se diffuse dans le discours, à la fois dans la durée et à travers différentes communautés linguistiques.
- Un nouvel état sémantique, où une nouvelle lexie s'est stabilisée dans les discours. Ceci permettra de la coder en langue et donnera lieu à son référencement, par exemple par la mise à jour de dictionnaires de référence.

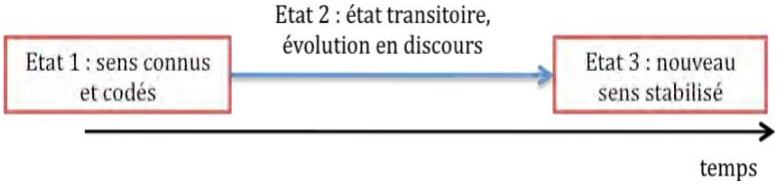


Figure I.3.1 : Schéma d'évolution du sens

À titre d'illustration, considérons l'exemple de *mutualiser*. Ce verbe a connu une évolution de sens au cours des vingt dernières années. En témoigne le *Petit Larousse*, qui propose deux acceptions pour *mutualiser* en 2009 :

> **mutualiser** v.t. 1. Faire passer un risque, une dépense à la charge d'une mutualité, d'une collectivité. 2. Mettre quelque chose en commun, le répartir. *Mutualiser des compétences, des frais*.

Seule la première acception, propre au domaine de l'assurance, est présente dans l'édition de 2005 du *Petit Larousse*.

L'évolution de sens de *mutualiser* a fait l'objet d'une étude particulièrement détaillée sur la période 1993-2009 (Viprey et Schepens, 2010). Jusqu'en 90, l'emploi de *mutualiser* est rare et il renvoie presque exclusivement au domaine de l'assurance. Le nouveau sens de partage de moyens et de ressources émerge et se diffuse dans les années 90, puis les emplois connaissent un basculement marqué entre 2001 et 2005 : la deuxième acception, associée à l'idée de mise en commun, s'impose dans les emplois et supplante la première acception. L'évolution de sens est marquée d'une part par une présence accrue dans les discours de *mutualiser* et de *mutualisation,* substantif associé ; d'autre part par l'apparition de nouveaux contextes qui supplantent progressivement les anciens : pour les corrélats *moyens* et *risques* par exemple, *mutualisation des moyens* s'impose dans l'usage au détriment de *mutualisation des risques* (Viprey et Schepens, 2010:495)). Cette évolution est résumée dans le schéma suivant :

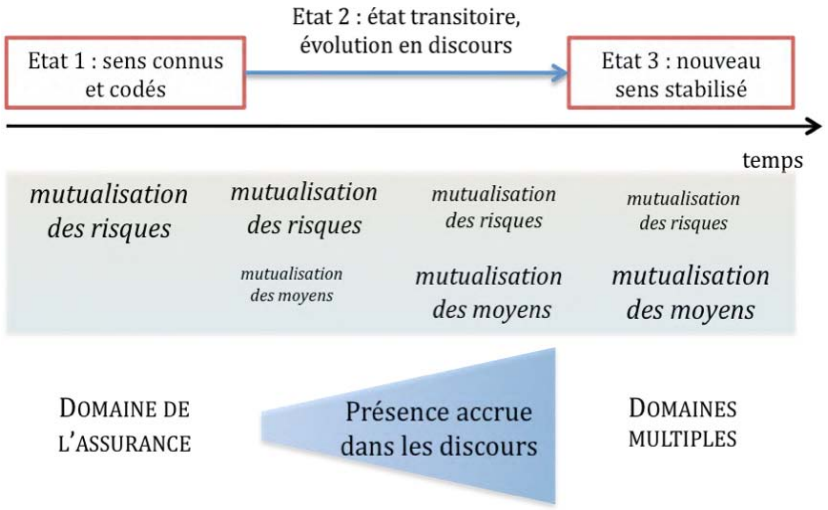


Figure I.3.2 : Schéma d'évolution du sens de mutualisation

On a défini notre objet d'étude comme des variations marquées en cours d'implantation. Ce choix conditionne doublement le positionnement dans le cycle d'évolution : d'une part, il implique l'existence de contrastes répétés entre le sens observé en discours et le sens codé ; d'autre part, il nécessite une régularité au niveau des contrastes. Cette régularité est le reflet d'une stabilisation : les variations sémantiques ne sont pas fragmentées en une foule d'emplois atomisés et hétérogènes, mais elles se répètent de façon similaire.

On se positionne donc en cours de période transitoire, à un stade suffisamment avancé pour qu'il y ait saillance quantitative, mais aussi suffisamment tôt pour que la détection et l'allocation d'un nouveau signifié puissent être précoces, autrement dit, de façon à anticiper l'état stabilisé sans pour autant être dans un événement accidentel. Ce qu'on cherche à observer doit se traduire quantitativement et qualitativement selon des tendances opposées, par une régularité quantitative dans l'irrégularité qualitative.

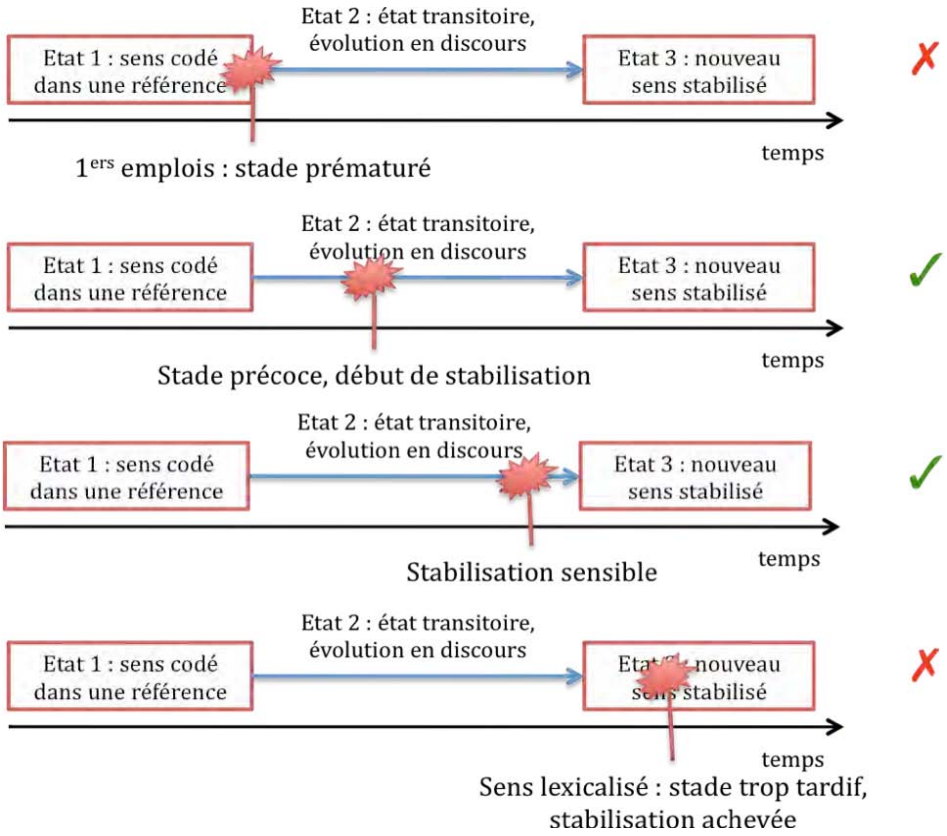


Figure I.3.3 : Positionnement dans le cycle d'évolution du sens

Dans la suite, nous cherchons à proposer des indices témoins de la transition et à analyser leur profil de comportement ainsi que leur accessibilité à un traitement automatique.

2. Modélisation : jeu sur des indices multiniveaux, de la détection à l'allocation

La néosémie s'exprime à différents niveaux. Ces niveaux peuvent être appréhendés à travers des indices organisés et analysés selon deux angles : l'existence d'un changement de sens, où les indices contribuent à la détection de la néosémie (sous-section 2.1) ; la qualification de ce qu'est le nouveau sens, où les indices peuvent être utilisés pour allouer un nouveau signifié (sous-section 2.2). La détection est un préalable incontournable à l'allocation de signifié : caractériser un changement de sens n'est possible que s'il y a effectivement changement de sens. De plus, nous faisons l'hypothèse que les indices qui servent en détection participent doublement à l'allocation de signifié :

- Ils jouent le rôle de pointeurs : ils aiguillent vers de l'information pertinente et guident vers des zones fécondes pour la construction du nouveau sens.
- Pour la plupart des indices, leur utilisation à des fins de détection correspond à une sous-exploitation de leur potentiel. La nature de ces indices ne change donc pas entre la détection et l'allocation de signifié ; en revanche, leurs modes d'observation et d'exploitation diffèrent : ils nécessitent des stratégies d'analyse plus poussées pour l'allocation de signifié.

Pour ces raisons, il nous semble indispensable de faire le point sur la détection avant d'aborder l'allocation, cœur de nos préoccupations. Le premier point mentionné ne donnera pas lieu à des développements expérimentaux : l'objectif à ce niveau est de fournir matière à

réflexion pour des développements futurs. Le second point sera en revanche développé en détail.

Après présentation des indices utilisables pour la détection et pour l'allocation, on débattra de l'autonomie et de l'articulation des niveaux d'observation associés à ces indices pour nous positionner, in fine, de façon à privilégier une approche sémantique et centrée sur la construction de signifié (sous-section 2.3).

2.1 Profiler la détection : des indices conscients aux indices émancipés

2.1.1 Une transgression consciente et affirmée : du sentiment aux indices de rupture

La néosémie est un objet difficilement saisissable, mais que l'on peut appréhender car il s'agit d'une transgression ressentie et, d'une certaine façon, cette nouveauté que l'on ressent *se dit*.

Précisons notre pensée. Pour identifier une néosémie et affirmer son caractère néologique, le sentiment néologique joue un rôle fondamental, comme le souligne notamment (Auger, 2010) qui l'inclut dans les critères principaux de néologicité.

Le sentiment néologique relève de la perception, il est dépendant de la subjectivité des individus, et on peut légitimement s'interroger sur sa pertinence pour guider une entreprise de modélisation.

Différents travaux invitent à penser que le sentiment néologique possède une certaine fiabilité. Cette fiabilité n'est ni absolue, comme le montre une étude de (Gardin *et al.*, 1974), ni aussi incertaine que ce que les résultats de l'étude laissent transparaître.

L'expérience de (Gardin *et al.,* 1974) montre combien il est délicat de s'appuyer sur le sentiment néologique dans une perspective de formalisation. Cette expérience a été initiée selon un postulat, à savoir l'universalité du sentiment néologique. Elle avait pour but de guider la théorie à partir de résultats empiriques, le néologisme étant un objet mal défini, sans statut théorique clairement établi. Les observateurs avaient pour consigne de relever les néologismes sur un corpus de huit pages du journal *Le Point*. Les résultats vont doublement à l'encontre du postulat d'universalité du sentiment néologique : ils témoignent d'importantes divergences entre évaluateurs et d'une inconstance pour un même évaluateur (une même lexie n'était pas systématiquement repérée comme néologique). Les désaccords, particulièrement marqués pour de la néologie de sens, semblent a priori remettre en question la fiabilité du sentiment néologique pour ce type de néologie et son exploitabilité dans une approche formelle ou modélisatrice.

Le manque de fiabilité qui semble résulter de cette étude s'explique en partie par les conditions expérimentales dont l'analyse amène à relativiser les résultats négatifs : 1) le repérage n'était pas spécifique à la néologie sémantique, d'où un risque de dispersion de la vigilance des évaluateurs à son égard (un regard focalisé exclusivement sur un objet, en l'occurrence la néologie de sens, est plus sélectif ; un repérage large de la néologie peut rendre moins attentif à un phénomène moins percutant qu'une néologie de forme, par exemple par procédé morphologique) ; 2) la taille du corpus était limitée et le nombre d'évaluateurs réduit ; 3) aucun critère de diffusion ou d'implantation n'avait été précisé. D'autres études invitent à reconsidérer le manque de fiabilité du sentiment néologique. Ainsi, pour les spécialistes d'une discipline chargés d'identifier les néologismes dans leur langue de spécialité (Auger, 2010) ou pour des personnes 'instruites' au sens qu'elle disposent de connaissances sur ce qui est attesté dans des ressources lexicales (Sablayrolles, 2002), le sentiment néologique présenterait une certaine fiabilité. L'expérience de (Ben Hariz Ouenniche, 2009) témoigne dans ce sens. Cette expérience adopte une démarche proche de

(Gardin *et al.,* 1974) et, bien que modeste en nombre d'évaluateurs, elle indique une convergence au niveau du repérage et de la classification des néologismes pour des évaluateurs ayant une certaine expérience dans l'étude des néologismes.

De plus, le sentiment néologique est corrélé à des indices formels. Nous estimons que ces indices sont des vecteurs de transmission du sentiment néologique, ils contribuent à le faire partager entre un locuteur, qui exprime sa conscience de la transgression, et un interprétant, qui se laisse guider par les indices.

Côté interprétant, l'analyse qualitative des résultats réalisée dans un second temps chez (Gardin *et al.,* 1974) montre qu'il existe des corrélations du sentiment néologique avec des indices plus adaptés à la modélisation. Malgré le caractère local et synchronique de l'étude, deux tendances se dessinent :

1. il existe des liens entre le sentiment néologique et la présence d'indices de surface, de type typographiques, métadiscursifs ou syntaxiques, sur lesquels on reviendra ci-dessous ;
2. une sensibilité au changement sémantique apparaît notamment lorsqu'il est sous-tendu par un processus de métaphorisation, donc lors de rencontres inattendues entre champs sémantiques ou domaines.

Selon nous, la fiabilité du sentiment néologique chez des interprétants est fondée, au moins partiellement, sur l'identification d'une régularité implicite ou explicite d'indices textuels, qui rejoint les tendances identifiées par (Gardin *et al.*, 1974). Nous faisons l'hypothèse que les corrélations qui s'esquissaient dans cette expérience sont généralisables et seraient encore plus marquées dans des études comportant moins de fluctuations du sentiment néologique. Chez un interprétant, le sentiment néologique semble donc guidé par des indices, qui sont émis par un ou des locuteurs.

Côté locuteur, certains de ces indices reflètent le fait que celui-ci a conscience d'une transgression sémantique ou d'un emploi atypique. Les locuteurs utilisent des stratégies récurrentes pour signaler la nouveauté, ou du moins exprimer leur sentiment de rupture. Ils recourent à des indices que nous qualifierons de marqueurs et qui renvoient à ce que (Bauer et Renouf, 2000) qualifient d'*overt or conscious help*. Leur présentation sera illustrée par l'exemple de *brouteur* : alors que le sens répertorié dans le TLFi est « Animal qui broute », ce substantif connaît actuellement de nouveaux emplois pour désigner des escrocs d'Afrique de l'Ouest qui, à travers des échanges Internet, cherchent à soutirer de l'argent à des Occidentaux (duperie affective, fausses donations, fausses ventes immobilières, etc.). Les marqueurs sont :

- des indices typographiques, principalement les guillemets et l'italique ou, moins fréquemment, le gras :

  > *Les « brouteurs » sont aussi des experts en manipulation sentimentale.* (source : programme de l'émission *Envoyé Spécial* du 17/02/2011 sur France 2, qui présente une enquête sur ce nouveau type d'escrocs ; accessible à l'adresse http://teleobs.nouvelobs.com/tv_programs/2011/2/17/chaine/france-2/20/35/envoye-special)

- des gloses, qui introduisent des définitions ou éléments de définition :

  > *Voici le visage des grands brouteurs d'Abidjan* ***(Escroquerie sur internet)***. (source : commentaire d'une émission sur les brouteurs postée sur YouTube, émis par la personne à l'origine du post, accessible à l'adresse http://www.youtube.com/watch?v=xwjrtfOQJpk)

- d'explications introductives, telles que *dit*, *qualifié de*, *que l'on appelle*, etc., ou consécutives, telles que *c'est-à-dire* :

> En ce qui concerne ce cas, d'une personne venant de Côte d'Ivoire, faisant partie d'une bande organisée **que l'on appelle des** « brouteurs » qui écument tous les sites de rencontres en prenant l'identité d'un Français ou d'une Française lambda, n'y aurait-il pas un moyen technique quelconque pour les repérer ? (extrait d'un article du site http://www.pointscommuns.com/si-la-photo-est-bonne-commentaire-musique-94431.html, décrivant le déroulé type d'une escroquerie sentimentale via un site de rencontre).
>
> Tu es en contact avec des « brouteurs », **c'est-à-dire** de jeunes Africains qui passent leur temps dans les cybers à arnaquer les hommes blancs en se faisant passer pour des filles (commentaire d'un forum de discussion intitulé *Arnaque sexe Côte d'Ivoire*, à l'adresse : http://www.commentcamarche.net/forum/affich-12711076-arnaque-sexe-cote-d-ivoire)

- des formules exprimant la nouveauté, qui témoignent notamment qu'un nouveau concept ou un nouveau phénomène s'incarnent linguistiquement à travers de nouvelles lexies. De telles formules sont notamment présentées par (Picton, 2009, 161-162), qui décrit un ensemble structuré de marqueurs de connaissances évolutives, adaptés à un contexte terminologique ; une classe de marqueurs recouvre ce qui nous intéresse ici, à savoir les marqueurs associés à la nouveauté de termes ou concepts, tels que *nouveau, récent* ou *apparaître*. Les formules de nouveauté se distinguent des indices précédents qui sont pour ainsi dire accolés à la lexie néologique. Elles peuvent se situer à une certaine distance linéaire du néologisme, mais elles restent cependant proches à travers des liens sémantico-syntaxiques : des processus d'anaphore ou encore de reprise les maintiennent dans la portée de la lexie néologique. Ces formules peuvent témoigner de l'existence d'un nouveau sens, mais aussi avoir une contribution notable à la qualification du nouveau sens : elles articulent la lexie néologique à des éléments de définition. Considérons l'exemple suivant :

> Dans le milieu de la police judiciaire, on les appelle les « brouteurs ». Ils forment souvent un groupe et agissent en meute. Cette **nouvelle forme de** criminalité via Internet fait fureur à Abidjan et fait beaucoup de victimes. (Article de blog, intitulé « Profession ? Brouteur à Abidjan », 27 avril 2010, accessible à l'adresse http://regardscroises.ivoire-blog.com/archive/2010/04/27/profession-brouteur-a-abidjan.html).

  Nouvelle reste dans la portée de *brouteurs* par l'intermédiaire de la chaîne référentielle supportée par le pronom *ils* et le démonstratif *cette*. L'adjectif *nouvelle* articule ainsi *forme de criminalité via Internet* à *brouteurs* et fournit ainsi des éléments de qualification du nouveau sens. Bien que ce type d'indice soit intéressant en termes de détection et prometteur en termes de qualification, nous ne l'approfondirons pas, car il repose sur des mécanismes de localisation précise de l'information et sur une analyse syntaxique fine, tandis que nous privilégierons des mécanismes de saillances liés à des effets de masse.

Aucune étude, à notre connaissance, n'a établi le profil général d'évolution dans le temps de tels indices. Cela aurait pu rejoindre l'objectif initial du projet APRIL (Analysis and Prediction of Innovation in the Lexicon) (Bauer et Renouf, 2000), qui était de faire émerger des corrélations entre les procédés à l'origine de changements de sens et les indices correspondant à une aide contextuelle, mais cet objectif n'a pas été atteint : une typologie des aides issues du contexte immédiat a été établie, mais pas les profils de corrélation. La présence d'indices et leur type sont connus, mais leur systématicité, leur courbe d'évolution ou leurs alternances relatives les uns par rapport aux autres n'ont pas été établies.

On peut supposer que de tels indices auront d'autant plus tendance à être présents que le changement est récent, car le sentiment de nouveauté ou de rupture sera plus fort lors des premiers emplois néologiques, et qu'ils s'effaceront une fois l'implantation dans l'usage bien amorcée.

Soulignons également que ces indices sont à utiliser et à interpréter avec précaution, notamment dans le cas des indices typographiques et des formules introductives ou consécutives. Ils peuvent témoigner de l'existence d'une néologie, mais cette interprétation n'est pas systématique. À cet égard, l'étude de la plurivocité des guillemets effectuée par (Rinck et Tutin, 2007) est éclairante. Elles montrent que les guillemets témoignent de l'existence d'une rupture, mais que cette rupture peut avoir diverses interprétations : nouveau concept, certes, mais aussi citation, mise à distance, mention d'un titre ou d'un nom propre, etc. Les guillemets ne peuvent faire l'objet d'une lecture univoque et leur interprétation mérite d'être guidée, ce que les auteures s'efforcent de faire à travers un enrichissement en annotations. Parler d'existence d'une rupture néologique est déjà un pas interprétatif et exige quelques précautions, notamment en terme d'exploitation systématique de tels indices. On peut supposer que la récurrence et la diversité d'indices associés à une même lexie sont susceptibles d'offrir plus de garantie sur la néologicité de cette lexie :

- la récurrence parce que la néologie est un processus, inscrit dans la durée : le sentiment néologique perdure, donc les indices exprimant ce sentiment tendent à se répéter ;
- la diversité parce qu'un indice donné peut prendre des valeurs différentes, mais l'ensemble de ces valeurs ne se superpose pas nécessairement à l'ensemble de valeurs associées à un autre indice ; par intersection d'ensembles, la valeur commune à une diversité d'indices tendra à converger vers celle de nouveauté.

Lors de nos investigations, nous privilégierons d'une part les indices typographiques, en particulier les guillemets, indice particulièrement fréquent et aisé à repérer, d'autre part les expressions introductives et consécutives. Nous laisserons de côté les formules de nouveauté et les gloses, qui aiguillent certes vers des éléments de définition pertinents, mais que nous préférons écarter pour deux raisons :

- ils nous ont semblé moins présents que les autres indices lors de nos investigations, donc ils ne sont pas prioritaires pour la détection ;
- pour une réexploitation dans une perspective d'allocation, ils aiguillent certes vers des éléments de définition particulièrement intéressants, mais leur étude répond à une problématique de localisation précise d'information plutôt que de synthèse sur un grand nombre d'emplois, donc ils ne rejoignent pas l'axe d'approche que nous souhaitons privilégier. De plus, ils semblent plutôt des indices propres aux premiers emplois, lorsque le sens n'est pas suffisamment stabilisé et nécessite d'être précisé par le locuteur, donc la définition vers laquelle ils aiguillent peut connaître une dérive au fil des emplois, une fois que le néologisme commence à être plus intégré dans l'usage et à se stabiliser. Cette hypothèse reste à approfondir et pourra faire l'objet d'études ultérieures.

2.1.2 Une transgression émancipée et mesurable : variations quantitatives et diversification

L'existence d'une néologie de sens peut être appréhendée à travers d'autres types d'indices. Ceux-ci ne sont pas des indices volontairement émis par le locuteur pour signaler l'existence d'une rupture ou d'une nouveauté : ils ne disent pas la transgression, mais ils la reflètent indirectement. Ils constituent des traces laissées par le phénomène néologique en lui-même, dans son caractère diffusif et expansif : leur caractère remarquable ne vient pas tant d'une rupture sur une occurrence isolée que d'une évolution sur une série d'occurrences. Autrement

dit, ces indices sont des témoins du processus, comme phénomène qui s'inscrit dans la durée et se construit à travers des récurrences ou des répétitions. Ils présentent ainsi deux caractéristiques :

- ils ne peuvent être saisis que si on accepte de se positionner dans une globalité, au-delà des volontés individuelles ou d'emplois locaux ;
- ils restent accessibles sous la forme de régularités de surface à l'échelle du texte ou du corpus de textes.

La plupart de ces indices n'apportent pas seulement une indication sur l'existence ou non d'un changement, ils peuvent, en outre, servir à le qualifier. On s'attardera ici sur la contribution à la détection de la néosémie, autrement dit, au témoignage de l'existence d'un changement ; la qualification du changement sera abordée à la sous-section suivante.

Les manifestations quantitatives du changement peuvent se répartir en deux types principaux :

- Les variations quantitatives qui renvoient à une variabilité associée à la lexie néologique. Elles reposent sur une **diversification de familles ou regroupements** associés au néologisme. La variation porte sur la taille de la famille, autrement dit sur son cardinal.
- Les variations quantitatives qui reposent sur le **nombre d'occurrences** et font intervenir la **densité dans les discours de l'unité lexicale ciblée ou de l'ensemble d'unités lexicales ciblées**. Elles s'expriment à travers l'évolution d'une fréquence ou d'une distribution de fréquences. L'existence d'une néologie sémantique apparaîtra à travers des variations de fréquence (croissance marquée par exemple, *cf*. infra) ou à des distances ou variations de distances entre environnements distributionnels successifs.

Les familles associées à la lexie néologique peuvent se répartir selon un degré de variabilité par rapport au signifiant. Les familles proposées ci-dessous peuvent s'imbriquer successivement : une nouvelle famille peut se définir par rapport à la lexie néologique simple, c'est-à-dire non regroupée, ou par rapport à une des familles intermédiaires. Par exemple, dans la structuration proposée, les phrasèmes à composante commune (*cf*. infra) associés aux produits et instruments financiers dits *toxiques* peuvent se définir autour de la composante simple *toxique*, avec *actifs toxiques, crédits toxiques,* etc.*,* ou autour des différentes composantes du regroupement morphologique *{toxique, toxicité},* qui ajoutera par exemple *toxicité des actifs*. Nous proposons les regroupements suivants, en précisant pour quel type d'indice d'évolution chacun sert d'observable privilégiée. Les indices d'évolution sont seulement mentionnés, ils seront détaillés ultérieurement.

- ***Lexème isolé :*** le degré de variation est minimal, l'observable est une unique entité, à savoir le lexème néologique. Le lexème est l'observable privilégiée au niveau des *empreintes de fréquences* (*cf*. infra).
- ***Regroupements morphologiques* :** la famille de lexies se constitue autour d'une base morphologique commune, elle correspond au paradigme morphologique associé à cette base. Ce type de famille regrouperait *mutualiser* et *mutualisation* dans un même ensemble ; de même, un regroupement de ce type serait constitué de l'ensemble *Outreau* (nom propre), *Outreau* (nom commun, dans l'expression « un nouvel Outreau »), *outreauiser*, relevés et analysés par (Lecolle, 2007) dans son étude sur l'évolution du sens d'*Outreau* de la désignation d'une ville du Pas-de-Calais à l'erreur judiciaire par excellence. Les regroupements morphologiques sont à la base des *variantes de forme*.
- ***Phrasèmes à composante commune* :** selon que la lexie néologique initiale est simple ou complexe, la famille peut correspondre soit à un ensemble de phrasèmes associés à la lexie simple, soit à un paradigme de lexies complexes possédant une composante

commune avec la lexie complexe initiale. À titre d'exemple, la lexie *cuisine moléculaire* a été introduite comme nouveau phrasème dans le Nouveau Petit Robert 2010[19], mais dans les emplois, l'adjectif *moléculaire* s'insère dans d'autres syntagmes nominaux, toujours en lien avec le domaine de la gastronomie, tels que *gastronomie moléculaire, perles moléculaires, billes moléculaires, spaghettis moléculaires, repas moléculaire*, etc[20]. L'adjectif *moléculaire* sert de pivot pour constituer une famille. Tous les éléments de cette famille participent du même phénomène néologique : pour l'instant, le néologisme en cours d'intégration dans des dictionnaires est une lexie complexe, mais les usages observés témoignent d'une émancipation de *moléculaire,* dont le nouveau sens est domanialisé en gastronomie. Les regroupements phraséologiques à composante commune se situent à l'articulation des regroupements morphologiques et des regroupements cooccurrentiels (voir infra) : ils possèdent non plus une base morphologique commune, mais un mot en commun, que ce mot soit lexie simple ou composante d'une lexie complexe ; ils dépassent le palier du mot et mettent en œuvre des liens syntagmatiques, qui restent limités à la lexie. Ils participent du phénomène de *foisonnement néologique* décrit infra, même s'ils n'en constituent qu'une sous-partie. Le phénomène complet relève de la famille suivante.

- ***Réseau paradigmatique* :** cette famille se définit selon un principe de substitution. Elle recouvre l'ensemble des lexies associées au nouveau signifié. Le signifié constitue donc le point d'articulation de la famille, tandis que le signifiant propre à la lexie néosémique peut n'avoir aucun lien formel avec le signifiant d'autres lexies de la famille. Un exemple de famille est présent chez (Dury, 2008) : elle présente l'ensemble des termes qui ont servi à désigner le pétrole en langue anglaise depuis le XIX^{e} siècle, ensemble qui comporte des unités à composante commune, telles que *oil, Sicilian oil, mineral oil*, et d'autres sans aucun lien au niveau du signifiant, comme *oil*, *naphtha, bitumen*. Ces familles sont étroitement associées au phénomène de *foisonnement néologique*.
- ***Réseau cooccurrentiel* :** l'ensemble des cooccurrents peut définir une famille, même si ce cas constitue un cas limite. La famille est définie par une relation d'appartenance à une même unité textuelle, éventuellement filtrée (exclusion de certaines catégories grammaticales, par exemple). Le réseau cooccurrentiel est un sur-ensemble des variantes phraséologiques, élargi à des unités textuelles supérieures au syntagme[21]. Par exemple, pour *toxique* en contexte de crise financière, des unités lexicales telles que *banque, américain, bilan* ou *sauvetage* s'ajouteront aux substantifs qualifiés par *toxique* tels que *créance, crédit, titre*, etc. Les constituants n'ont pas de contraintes imposées en termes de parenté avec le signifiant de la lexie néologique. Le lien de cooccurrence peut s'appliquer

[19] Source : pages du site web de (Martinez, 2011) qui répertorient les nouveaux mots et sens apparus dans le Petit Robert et le Petit Larousse, à l'adresse http://www.orthogrenoble.net/page-de-camille-club-orthographe-grenoble.html.

[20] Exemples récoltés sur la Toile, essentiellement sur des blogs de cuisine ou des sites commerciaux ; quelques sources : http://www.10zign.fr/kit-moleculaire-spherification-p-252.html ; http://recettes-de-ml.over-blog.com/article-30026750.html ; http://lesfeesmeres.over-blog.com/article-soiree-moleculaire-58923530.html ; http://blog.victoiremag.lesoir.be/larecette/2009/10/10/ludiques-et-surprenants-les-spaghettis-moleculaires/.

[21] Nous faisons la distinction entre variantes phraséologiques et cooccurrents car cela correspond à un saut de palier. (Mayaffre, 2008) invite à distinguer la cooccurrence, générique et pertinente pour décrire la co-présence sur des unités textuelles de taille variable, de cooccurrences sur des paliers réduits et contraintes syntaxiquement – les *corrélats*, qui d'une certaine façon rejoignent les phrasèmes selon que la lexie-pôle y est incluse ou non. Une telle distinction est également faite par (Condamines *et al*., 2004) dans leur étude sur l'évolution temporelle de termes, où sont distingués l'expansion, associée au palier du syntagme, et la distribution, au-delà de ce palier. Dans notre cas, variantes phraséologiques et variations distributionnelles ne sont pas disjointes mais en relation d'inclusion.

à la lexie néologique ou à une des familles associées à la lexie néologique, telles qu'elles ont été définies précédemment.

La gradualité de variation du signifiant (signifiant fixe, modulé ou remplacé d'un élément à l'autre d'une même famille) a permis d'ordonner les familles décrites ci-dessus, mais leur unité repose également sur un partage de signifié, identique ou proche d'une unité à l'autre de la famille, sauf dans le cas limite des cooccurrents où le lien de chaque cooccurrent avec le signifié de la lexie néologique est plus complexe. Ce sont donc bien des familles associées à la lexie qu'on vise idéalement, même si, dans un traitement automatique, le mot-forme prime sur la lexie.

Comme nous allons le voir, un certain nombre de travaux ont défini des phénomènes qui correspondent à des traces du changement de sens. Ces phénomènes tendent à privilégier l'une ou l'autre des familles précédemment décrites. Ils délivrent des informations sur l'existence d'une néosémie, à travers des considérations essentiellement quantitatives et sur ce qu'est le nouveau sens, lorsque le quantitatif est couplé à du qualitatif. On abordera les phénomènes suivants, dont une approche centrée sur du quantitatif permet de mettre en évidence l'existence d'une néosémie : les *empreintes de fréquence*, les *variantes de formes* et le *foisonnement néologique*. On cherchera à mettre en relief quelles informations sont accessibles à partir d'une approche quantitative, découplée du qualitatif. L'apport qualitatif, à même de renseigner sur ce qu'est le nouveau sens, sera abordé dans la sous-section suivante. En particulier, les cooccurrents seront étudiés sous l'angle qualitatif, angle le plus adapté pour exploiter la richesse de l'environnement multi-variable qu'ils constituent.

***Les empreintes de fréquence* (Picton, 2009:106[22]).** Elles correspondent à l'évolution au cours du temps de la fréquence de la lexie ciblée. Elles décrivent le comportement d'une unité lexicale, c'est-à-dire qu'elles contribuent à décrire le comportement d'unités isolées, même si leur observation pourrait se généraliser à des familles plus larges. Dans notre cadre, on cherche à observer l'apparition de nouvelles lexies et leur implantation dans les discours. On fait l'hypothèse qu'une néologie de sens peut s'accompagner d'un accroissement de fréquence, à même de représenter la diffusion d'un nouvel emploi dans les discours.

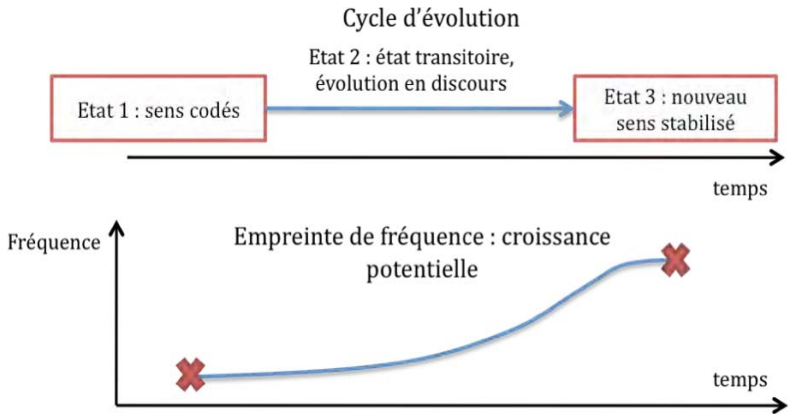
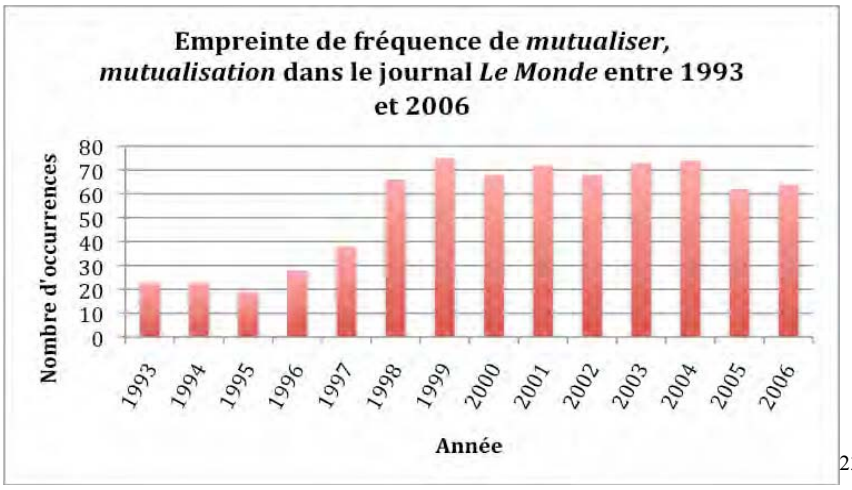


Figure I.3.4 : Cycle d'évolution et empreintes de fréquence

Cet accroissement de fréquence peut se produire lorsque le sens de la lexie évolue d'un emploi propre à un domaine particulier vers un emploi plus général. C'est notamment le cas de *mutualiser*, initialement associé au domaine de l'assurance puis employé dans un cadre plus général à partir de la fin des années 90 et le début des années 2000.

[22] L'auteure reprend et traduit le terme « *frequency signature* » introduit par (Ahmad *et al.*, 2002).



[23]

Figure I.3.5 : Empreintes de fréquence de mutualiser, mutualisation

Les informations purement quantitatives délivrées par les empreintes de fréquence doivent toutefois être manipulées avec précaution. En effet, l'existence d'un accroissement de fréquence ne garantit pas qu'un nouveau sens s'implante. D'autre part, tous les néologismes de sens ne présentent pas nécessairement ce profil. Un nouvel emploi peut rester mineur par rapport aux emplois préexistants, ou un nouveau sens peut remplacer un sens préexistant, donc en l'absence de désambiguïsation, la fréquence de l'unité lexicale peut stagner, voire diminuer.

Une empreinte de fréquence qui correspond à un accroissement est donc un indice d'existence d'une néosémie, mais pas une preuve, puisque ce type de profil n'est ni systématique, ni un garant de néosémie. Un tel indice n'est donc pas à rejeter complètement, mais à exploiter intelligemment, en ayant conscience des limites et en complétant les tendances qu'il délivre par une analyse qualitative ou par des recoupements avec d'autres indices. Moyennant ces précautions, les empreintes de fréquence sont pertinentes pour détecter la néosémie, car elles correspondent à un phénomène quantitatif qui peut s'étudier indépendamment de variations susceptibles d'apporter de l'information qualitative. Leur exploitation se réduit à de la détection tant qu'elles sont décorrélées d'un profilage sémantique des discours. Elles peuvent aussi contribuer à la qualification du nouveau sens dans le cas inverse, ce qui nous a amenée à conserver ce type d'indices dans notre champ d'investigation. Nous reviendrons ultérieurement sur la question de l'apport des empreintes de fréquence à la qualification du nouveau sens.

L'émergence de variantes de formes. On entend par variantes de formes des unités lexicales qu'on peut associer à la même famille morphologique que l'unité lexicale observée ou qui constituent des variantes phraséologiques de l'unité lexicale ou d'une partie de l'unité lexicale considérée. Ces variantes de formes peuvent relever de trois phénomènes : un **renforcement** de la lexie néologique ; une **émancipation** ; une **concurrence** avec d'autres lexies, dont les variantes de formes ne sont qu'un sous-phénomène et qu'on abordera spécifiquement au point suivant. De façon plus détaillée :

- Le **renforcement** comporte deux facettes : il peut servir à créer une redondance pour mieux guider l'interprétation ; il peut être destiné à moduler le nouveau sens pour en faire ressortir une de ses propriétés. La redondance apparaît chez (Renouf, 2000) comme les *root or base repetition*, qui correspondent à une reprise de la lexie néologique par une de

[23] L'histogramme de fréquences a été réalisé à partir des archives du Monde, sur des tranches de temps annuelles, à partir d'une requête à partir de l'expression régulière "mutualis[ea]*".

ses variantes dans un voisinage proche, ce qui se traduit par une densité locale d'unités de la famille ; la variante donne un nouveau contexte d'interprétation, donc elle consolide l'accès au nouveau sens en offrant un éclairage similaire, à proximité de la lexie néologique. Dans l'extrait d'article ci-dessous, la reprise de *brouteur* par *broutage*, puis la nouvelle occurrence de *brouteur* permettent de définir la forme d'escroquerie associée, puis de revenir et d'insister sur les aspects de la manipulation :

> « L'ETAU SE RESSERRE SUR LES **BROUTEURS** D'ABIDJAN
>
> Qui n'a jamais reçu un spam l'appelant à envoyer de l'argent à l'autre bout du monde ? En Côte d'Ivoire, d'où sont envoyés une grande partie de ces mails, on appelle cela du "**broutage**". Il s'agit d'appâter un internaute européen à la suite d'un échange de courriers électroniques en lui faisant miroiter une somme d'argent en échange d'un service rendu, un investissement immobilier ou encore en se faisant passer pour une femme en quête d'un mari. Les **brouteurs**, qui renouvellent sans cesse leurs stratagèmes, sont passés maître en l'art de la persuasion, grâce notamment à leur capacité à produire de faux documents pour rassurer leur "pigeon" sur leur identité ou leur bonne foi. » (Source : France 24, 28/07/2009, accessible à l'adresse http://observers.france24.com/fr/content/20090728-cote-ivoire-brouteur-brouter-internet-arnaque-victime-abidjan-cybercriminalite-spam)

Le renforcement semble plutôt se manifester lorsque le sentiment de nouveauté est fort, donc en début de phase transitoire, de façon à guider par redondance l'interprétation qui pourrait être encore hésitante ou incertaine. La modulation intervient du fait que des variantes morphologiques sont aussi sémantiquement liées. Jouer sur les variantes permet de mettre en relief différents aspects ou propriétés du nouveau. La variante peut intervenir à proximité de la lexie néologique, mais elle peut aussi apparaître isolément dans un discours donné, comme c'est le cas dans un article du *Figaro* du 10/10/2008[24], où il est question de *toxicité des créances* et non de *créances toxiques*, alors que *toxique* est l'unité lexicale usuellement porteuse de néosémie, comme qualificatif de produits ou instruments financiers (pour plus de détails, voir (Reutenauer, 2010)) : seul le substantif *toxicité* est présent, l'adjectif *toxique* n'apparaît pas dans l'article.

- L'**émancipation** de la lexie néologique peut se faire lorsque le nouveau sens commence à être intégré dans les usages, donc à un stade de stabilisation relativement avancé, où la lexicalisation se profile. La famille peut s'enrichir en néologismes de forme. Ce phénomène s'observe par exemple pour le nom propre *Outreau* (Lecolle, 2009), dont le sens initial renvoie à une ville du Nord-Pas-de-Calais, puis qui évolue au cours de l'affaire d'Outreau, de fin 2001 à 2006, vers le sens de "scandale judiciaire par excellence". Conjointement à l'évolution de sens du nom propre *Outreau* apparaissent des variantes morphologiques (*outreauiser, outreauisme*) et des emplois par antonomase (*un nouvel Outreau, des Outreau)*. Ces nouvelles formes sont présentes à partir de 2005, c'est-à-dire à la fin de la période au cours de laquelle s'est progressivement construit le nouveau sens. Elles sont le témoin d'une certaine diffusion et d'un certain degré d'implantation du nouveau sens (Lecolle, 2009:94-95). Bien que limitées en nombre d'occurrences, elles participent à un accroissement notable de la famille morphologique associée au nom propre *Outreau*.

[24] Source : article «Le moment est idéal pour investir sur l'or», http://www.lefigaro.fr/sicav/2008/10/10/04006-20081010ARTFIG00358-profiter-de-l-heresie-des-marches-.php

Les variantes de forme sont accessibles et, d'une certaine façon mesurables, car elles correspondent à une régularité au niveau des signifiants accompagnée d'un accroissement de la taille de la famille morphologique ou phraséologique associée. Le profil général de telles variations est plus difficile à établir que pour les empreintes de fréquence ou l'environnement distributionnel : il peut y avoir un accroissement temporaire de la famille des variantes du fait d'une concurrence de formes lexicales au sein d'une même famille, ou un accroissement qui apparaît avec la stabilisation, du fait du développement de nouveaux potentiels.

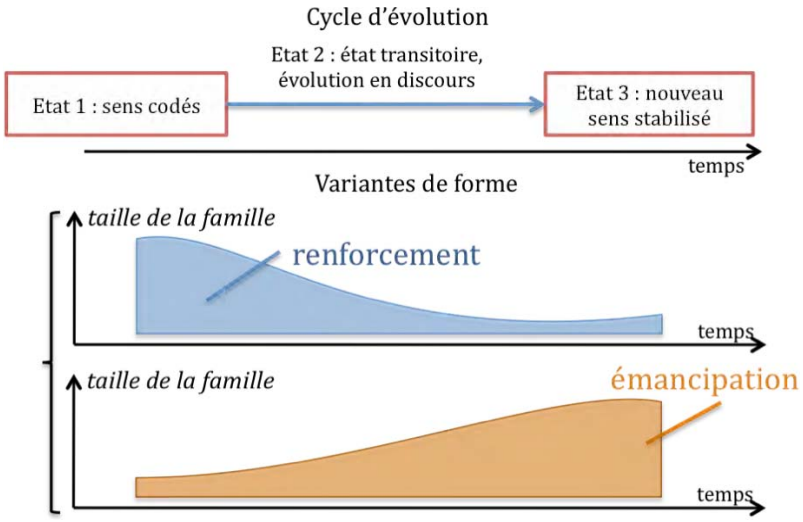


Figure I.3.6 : Cycle d'évolution et émergence de variantes de formes

Au cours de nos observations, les phénomènes associés aux variantes de forme ne nous ont pas paru centraux pour l'étude de la néosémie : ils semblaient à la fois être moins présents et jouer un rôle moins important que d'autres indices pour témoigner de l'existence d'une néosémie et pour guider vers le nouveau sens. De ce fait, nous n'avons pas approfondi leur étude et nous ne les retiendrons pas ici comme indices de détection automatisable.

***Le foisonnement néologique transitoire* (Guilbert, 1965:331, cité par Dury, 2008:13) .** Il correspond à un **concurrence de formes**, autrement dit, il y a la présence alternée, à une même période, de plusieurs formes lexicales susceptibles de correspondre au nouveau signifié. Au cours du temps, les emplois tendent à converger et une forme lexicale finit par s'imposer. La famille associée au phénomène néologique est le réseau paradigmatique, constitué de lexies au moins temporairement synonymes. Quantitativement, le foisonnement néologique correspond à un accroissement rapide, éventuellement massif, de la taille de la famille, puis à sa décroissance au fur et à mesure que le néologisme se stabilise, en s'incarnant de façon privilégiée dans une lexie. (Dury, 2008) présente de telles variations quantitatives du paradigme néologique à travers l'étude des termes employés pour désigner le pétrole entre le XIXe siècle et l'époque actuelle : 36 termes ont servi à désigner le pétrole dans les premiers temps, lorsque ce concept correspondait à un phénomène émergent et encore mal connu, puis la famille s'est progressivement réduite. Seul un petit nombre d'entre eux sont encore employés et le sens des termes encore usités s'est spécialisé pour désigner des réalités distinctes. De même, (Dassi, 2003) a montré un phénomène de foisonnement néologique sur les dénominations associées au téléphone portable entre 1998 et 2003 au Cameroun, dans la langue orale. En 1998, le substantif *cellulaire* était employé quasi exclusivement. Une concurrence avec le substantif *portable* est progressivement apparue et elle s'est imposée comme telle vers 2002, puis *portable* s'est imposé dans les usages, jusqu'à représenter 9 emplois sur 10 en fin de période étudiée. L'évolution quantitative montre une concurrence croissante, puis décroissante au fur et à mesure que *portable* devient l'emploi

dominant. De façon générale, le foisonnement néologique s'accompagne de variations quantitatives de type croissance-décroissance de la taille de la famille de lexies concurrentes.

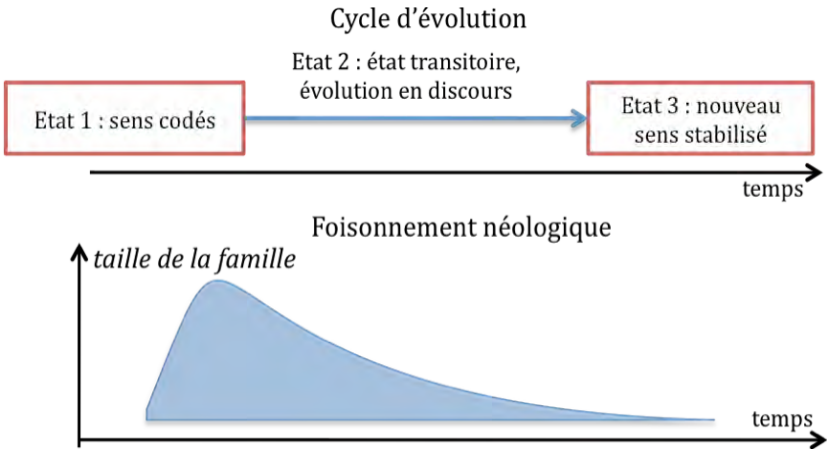


Figure I.3.7 : Cycle d'évolution et foisonnement néologique transitoire

Outre ses apports en termes de détection, le foisonnement néologique a un fort potentiel en termes de qualification du nouveau sens. Cet indice a donc été retenu dans notre champ d'investigation, malgré certaines difficultés liées à la constitution des paradigmes du foisonnement lors d'un traitement semi-automatique.

Pour résumer, les indices quantitatifs peuvent donc se concevoir comme des témoins du processus, dont l'évolution peut être approximativement schématisée comme suit :

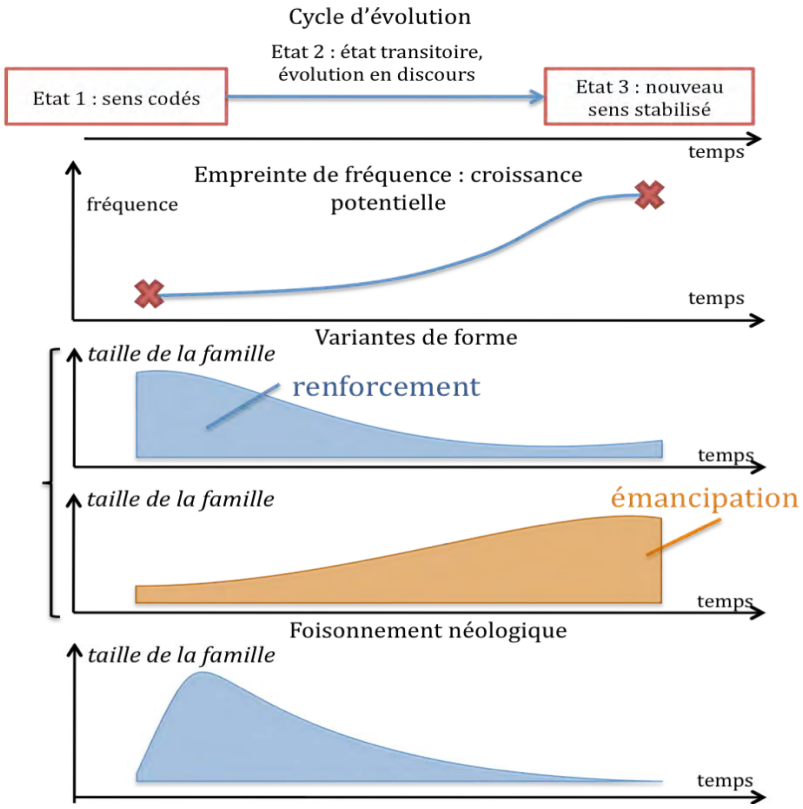


Figure I.3.8 : Bilan de l'évolution des indices quantitatifs témoignant de l'existence d'une néosémie

Les indices présentés sont des témoins d'existence d'un changement de sens lorsqu'ils sont observés à travers des critères essentiellement quantitatifs, et faiblement qualitatifs. Ces indices restent des indices, c'est-à-dire qu'ils n'offrent aucune garantie : ils ont des comportements potentiels, mais ces comportements ne sont pas vérifiés systématiquement et ils peuvent aussi correspondre à d'autres phénomènes. Ils sont donc à manipuler avec précaution. Plutôt que d'en faire une exploitation isolée et absolue, il convient de les croiser pour faire émerger des convergences, de faire des vérifications par un retour au texte et de compléter leur étude par une analyse qualitative, qui ne peut que consolider et enrichir les informations qu'ils délivrent.

2.2 Allocation de signifié : exploiter le potentiel qualitatif des indices quantitatifs

À certains niveaux d'observation, les indices repérables ne renseignent pas seulement sur l'existence d'un changement de sens mais aussi sur ce qu'est ce changement. Ils sont susceptibles de fournir des informations sur le nouveau sens.

Les données quantitatives ne sont pas exclues, mais elles sont au service du qualitatif et servent à mettre en relief des unités ou des relations linguistiques qui sont éclairantes sur le plan sémantique. Les informations sur l'évolution de sens sont accessibles à travers la reconfiguration d'unités, qu'il s'agisse de saillances inédites, de disparition d'unités particulières, de regroupements d'unités en affinités sémantiques, ou d'apparition de motifs composés d'unités hétérogènes. Ces informations apparaissent aussi à travers de nouvelles contraintes sémantiques exercées sur ou par l'unité lexicale, avec par exemple une évolution de dépendances sémantiques.

Ces changements qualitatifs renseignent sur le changement de sens à plusieurs titres : ils peuvent faire ressortir des facettes sémantiques déjà présentes dans l'ancien sens, manifester des incompatibilités avec d'autres facettes sémantiques qui seront alors exclues, ou faire émerger de nouvelles facettes sémantiques étrangères à l'ancien sens. Ces actions constituent le trio activation-inhibition-enrichissement, qu'on abordera plus en détail en sous-section 3.3 à partir de notions théoriques empruntées à la sémantique textuelle, notre cadre théorique pour représenter le sens.

Un emploi isolé ou des affinements successifs par recoupements d'emplois isolés peuvent apporter des informations, mais ce type d'analyse relève plutôt d'une approche manuelle, en finesse. Face à une grande quantité de données, les approches manuelles se heurtent à leurs limites. Dans le présent travail, on cible des informations acquises par un traitement automatique ou semi-automatique, capables d'émerger par effet de masse, à travers des régularités et des convergences.

Les familles évoquées dans la sous-section précédente sont toutes susceptibles de contribuer à la qualification du changement sémantique, mais à des degrés divers. On s'attardera sur celles qui nous paraissent offrir un potentiel particulièrement riche, à savoir : les cooccurrents, ou plus largement l'environnement distributionnel ; les paradigmes associés au foisonnement néologique. On reviendra également sur les empreintes de fréquence.

2.2.1 Cooccurrents et environnement distributionnel : de la matière à structurer

L'environnement distributionnel présente une richesse pour qualifier le changement sémantique à plusieurs égards :

- ***Nature des liens variables*** **:** les cooccurrents peuvent faire émerger des unités qui entretiennent différents types de liens avec la lexie néologique. Ces liens peuvent correspondre aux relations classiques de synonymie, antonymie, hyperonymie et

hyponymie, mais ils ne s'y réduisent pas car ils peuvent mettre en relief des relations de type prédicat-argument, ou, plus largement des affinités d'association, qui n'imposent pas d'emblée une nature particulière de lien :

> « (…) dans les mailles du filet cooccurrentiel s'accrocheront des lexies composées, des syntagmes figés, des expressions semi-figées mais d'autres associations lexicales plus libres et plus inattendues seront aussi remontées à la surface *sans que l'on veuille* a priori *ni les exclure, ni les privilégier* » (Mayaffre, 2008b).

L'absence de contraintes sur la nature du lien peut introduire un certain flou, mais d'une part, ces liens peuvent être analysés a posteriori, d'autre part, cette absence de contrainte ouvre un horizon d'exploration vaste ;

- ***Information corrélée à différents paliers textuels*** **:** l'environnement cooccurrentiel peut être délimité par différentes unités textuelles (syntagmes, phrases, paragraphes, article pour du discours journalistique, etc.). De ce fait, la définition du changement sémantique peut jouer sur différentes échelles textuelles. L'information délivrée par les cooccurrents reflète alors une cohérence propre à l'unité textuelle retenue. Nous nous intéresserons particulièrement aux paliers textuels larges, à même de refléter l'influence de la textualité.
- ***Information issue de différents niveaux d'analyse linguistique*** **:** l'environnement distributionnel peut être défini par la cooccurrence lexicale, mais aussi par d'autres unités. Celles-ci peuvent être issues d'un filtrage ou d'un enrichissement en information du plan lexical, voire d'unités obtenues indépendamment du plan lexical.
    - *Filtrage :* il peut se faire à partir de contraintes d'ordre grammatical, sémantique, syntaxique... Par exemple, (Villemonte de La Clergerie, 2011) filtre sur critères syntaxique et de taille de contexte : les contextes de cooccurrences sont des triplets de dépendances syntaxiques, ces dépendances étant de différents types. Un autre critère de sélection peut être de ne conserver que les cooccurrents qui appartiennent à une certaine ressource. (Jacquey *et al.*, 2010) procède ainsi à un filtrage lexical préalable des cooccurrents paragraphiques d'un terme-cible, en ne conservant que les termes appartenant à un thesaurus de linguistique, Thesaulangue[25].
    - *Enrichissement* : les unités lexicales peuvent être enrichies en informations issues d'une annotation, voire se faire remplacer par ces annotations. L'enrichissement peut être de nature variable. Il peut notamment correspondre à une annotation en domaines ou en thèmes, comme c'est le cas chez (Rayson *et al.*, 2004) : celui-ci enrichit son corpus par une annotation en domaines issus du *Longman Lexicon of Contemporary English.* L'enrichissement peut consister à ajouter une unité de sens particulière, comme c'est le cas chez (Valette, 2008) : dans un extrait de *Bouvar et Pécuchet* de Flaubert, il ajoute les traits sémantiques /botanique/ et /ornemental/ aux unités lexicales qui en sont porteuses et il analyse la récurrence de ces traits sémantiques. Ce type d'information, issu d'un enrichissement sémantique, nous intéressera tout particulièrement par la suite. On en proposera une systématisation, car les interactions avec des thèmes, des domaines, ou avec des unités de sens particulières jouent un rôle fondamental dans la néologie sémantique. Le cadre théorique retenu sera particulièrement favorable pour mettre en valeur de tels niveaux d'enrichissement sémantique. Cette question sera abordée en sous-section 3.3.

[25] Ce thesaurus est disponible sur le portail Termsciences, accessible à l'adresse http://www.termsciences.fr/.

- *Recherche d'information en dehors des unités lexicales :* l'environnement distributionnel peut être défini à partir d'unités qui ne proviennent pas des voisins lexicaux de la lexie néologique, mais sont issus d'autres niveaux d'information. C'est par exemple le cas des rubriques de journaux qui peuvent guider l'interprétation d'une unité lexicale. Ou encore, dans une perspective TAL, les mots-clés issus de métadonnées de pages web peuvent constituer un environnement propre à la page web où apparaît la lexie.

Aussi bien en termes de nature des unités que de relations entre celles-ci, l'environnement cooccurrentiel se caractérise par une diversité importante et un fort degré de liberté. Ces propriétés permettent l'émergence de structures variées d'unités. On mentionnera deux types de regroupements susceptibles de se dessiner :

- ***Des regroupements possédant une affinité sémantique.*** Ces regroupements correspondent notamment aux métaphores conceptuelles (Ferrari, 2006:212 sqq). Cette notion renvoie à une conception de la métaphore comme une relation conceptuelle du type « x est y » entre un concept cible x et un concept source y, où x et y peuvent être exprimés par une diversité d'unités lexicales. Ainsi, dans les discours de presse sur la crise financière de 2008, une connotation négative se greffe au sens de *finance*, notamment à travers des métaphores conceptuelles telles que « la crise financière est un cataclysme », lexicalisée à travers les syntagmes *tsunami financier, tempête financière, bourrasque financière, tornade financière*, *tourmente financière,* c'est-à-dire des *x financier,* où *x* ∈ *{tsunami, tempête, bourrasque, tornade, tourmente}*. Les regroupements par affinité sémantique peuvent aussi correspondre à des classes sémantiques associées de façon récurrente à la lexie néologique, ou encore aux taxèmes de la sémantique interprétative, cadre théorique que nous préciserons ultérieurement. Le noyau sémantique commun aux unités des regroupements est susceptible d'affecter le nouveau sens, soit en activant une facette sémantique préexistante de la lexie néologique, soit en ajoutant une nouvelle facette sémantique. L'effet sémantique a d'autant plus de chance d'être marqué qu'il est porté par un ensemble d'unités, et non par une unité isolée.
- ***Des regroupements composés d'unités disparates*.** Ces regroupements sont constitués d'éléments qui, isolément, ne donnent qu'une approche très partielle et peu éclairante sur le nouveau sens, mais dont la combinaison forme un tout cohérent et guide de façon précise et pertinente la qualification du nouveau sens. Ce type de regroupement peut par exemple correspondre aux motifs définis par (Longrée *et al.*, 2008), puis précisés par (Longrée *et al.*, 2010), qui sont des ensembles formant une micro-structure récurrente et dont les éléments constitutifs peuvent être hétérogènes. Dans le cadre théorique de la sémantique interprétative qui sera le nôtre, ces regroupements seront associés à la notion de forme sémantique, que l'on précisera en sous-sections 3.3 et 3.4.

2.2.2 Paradigmes issus du foisonnement néologique : renforcements, exclusions et transferts sémantiques

Les paradigmes issus du foisonnement néologique servent à qualifier le sens selon des procédés similaires à ceux qui interviennent pour les cooccurrents. Ils peuvent se structurer en sous-ensembles qui font ressortir certaines propriétés sémantiques de la lexie néologique ou qui définissent une nouvelle facette sémantique. On mentionnera trois phénomènes :

- *nouvelle facette sémantique*. Au fur et à mesure que le foisonnement néologique se restreint, la lexie néologique peut intégrer des facettes sémantiques de ses anciens concurrents. L'analyse de (Dury, 2008) sur les concurrents du pétrole (*oil*) en témoigne. L'auteur identifie un ensemble d'unités concurrentes de type toponymique (*Persian rock*

oil, Sicilian oil, Trinidad bitumen, ...). Ces unités constituent près d'un tiers de la famille et ont aujourd'hui complètement disparu. Elles ont permis de préciser le concept de pétrole à une époque où il était encore mal défini. Les précisions qu'elles apportaient sont devenues progressivement partie intégrante du sens de pétrole. De même, (Cusin-Berche, 2003) montre que *voiture* intègre certaines facettes sémantiques de son concurrent *automobile* au fur et à mesure que ce dernier disparaît des emplois au profit de *voiture* :

> « La disparition de la « concurrence » est susceptible de favoriser un glissement sémantique, comme l'illustre l'exemple analysé par A. Martinet [1969:37]:
>
> > « tant qu'il y avait des voitures à chevaux, on parlait d'*automobiles* ; aujourd'hui, il n'y a plus, de nouveau, que des *voitures* »
>
> qui suggère que *voiture* s'est chargé des sèmes précédemment attribués à *automobile* »

- *actualisation de facette sémantique par analogie.* La lexie néologique a des affinités sémantiques avec celles du paradigme, l'actualisation de son sens se fera de façon similaire aux unités qui lui sont proches. C'est en particulier le cas lorsque la lexie s'insère dans un réseau correspondant à un des concepts, source ou cible, d'une métaphore conceptuelle. Ainsi, pour la métaphore conceptuelle « la crise financière est un cataclysme », *tsunami* s'insère dans le paradigme associé au concept source de cataclysme. L'intégration dans la métaphore conceptuelle entraîne une analogie des mécanismes d'actualisation. Des unités du paradigme telles que *tempête* et *tourmente* sont employées fréquemment de façon métaphorique, dans une grande diversité de domaines, pour évoquer les idées de catastrophe et de violence, non pas celles de phénomène naturel ou météorologique. De même, pour *tsunami*, le sens métaphorique, rare jusqu'en 2004, se construit de façon similaire aux autres éléments du paradigme : l'idée de catastrophe ressort, tandis que la dimension climatique s'efface, tout comme pour les autres unités du paradigme. De même, toujours dans l'idée que « la crise financière est un fléau », l'adjectif *toxique* connaît des emplois similaires à *douteux* et *pourri*, ce qui active le sens de néfaste.
- *actualisation de facettes sémantiques par opposition*. Certaines unités interdisent à d'autres unités du même paradigme d'exprimer certaines nuances de sens, ou au contraire en font ressortir par contraste. Considérons le cas de *manageur* (Cusin-Berche, 2003:30), substantif attesté depuis 1961 dans le Petit Robert mais dont les emplois ont évolué du domaine sportif ou artistique vers le monde de l'entreprise. Dans ce contexte d'emploi, *manageur* s'insère dans le paradigme de concurrents {*manageur, directeur, décideur*}. Le trait sémantique /décisionnel/ est ainsi inhibé pour *manageur*, par opposition aux autres éléments du paradigme.

Notons que toutes les unités issues du foisonnement néologique ne renseignent pas nécessairement sur le nouveau sens. Ainsi, dans l'étude sur les substantifs servant à désigner les téléphones mobiles (Dassi, 2003), *portable* et *cellulaire* sont en concurrence, mais le sens de l'un n'éclaire pas sur le sens de l'autre, ou de façon très ténue. Les deux appellations renvoient à des propriétés distinctes des appareils, elles ne renforcent pas mutuellement certaines facettes sémantiques.

2.2.3 Empreintes de fréquence : couplage à un espace sémantique

Les empreintes de fréquence ne contribuent à qualifier le nouveau sens que si elles sont corrélées à des profilages sémantiques des discours, par exemple à des profilages thématiques ou domaniaux. Les cartographies dynamiques de la plateforme ProxiDocs (Roy, 2007) sont un exemple de profilage domanial. Les domaines sont définis par l'utilisateur à partir d'ensemble de lexies puis projetés dans les textes. Cette projection permet de dégager la façon

dont se structurent textes et domaines, notamment comment se répartissent les domaines dans les textes, période par période. La plateforme peut ainsi générer des cartes de domaines propres à chaque période. Les empreintes de fréquences d'une lexie peuvent s'établir relativement à ces cartes thématiques, par restriction aux textes associés à un domaine donné, et ce pour chaque domaine. L'information sur le changement de sens s'obtient à domaine fixé, notamment à travers un accroissement dans un domaine donné (*tablette* dans le domaine de l'informatique, avec par exemple l'émergence des *tablettes numériques*), ou par interaction des profils associés à plusieurs domaines (généralisation des emplois métaphoriques de *tsunami* sensible à travers la diffusion dans plusieurs domaines, avec des accroissements en finance, en politique, dans le domaine médiatique, etc.).

L'exemple de *toxique* illustre l'apport qualitatif d'un profilage domanial. Les empreintes de fréquence ont été obtenues à partir de la banque de données d'actualité internationale Factiva. Le corpus est journalistique[26], il recouvre la période du 01/01/2004 au 31/12/2010. Les sources textuelles sont classées ont fonction de thèmes auxquels elles se rattachent – les *sujets*. Nous en avons sélectionné quelques-uns pour établir les empreintes de fréquence de *toxique* relatives à ces thèmes[27], présentées ci-dessous :

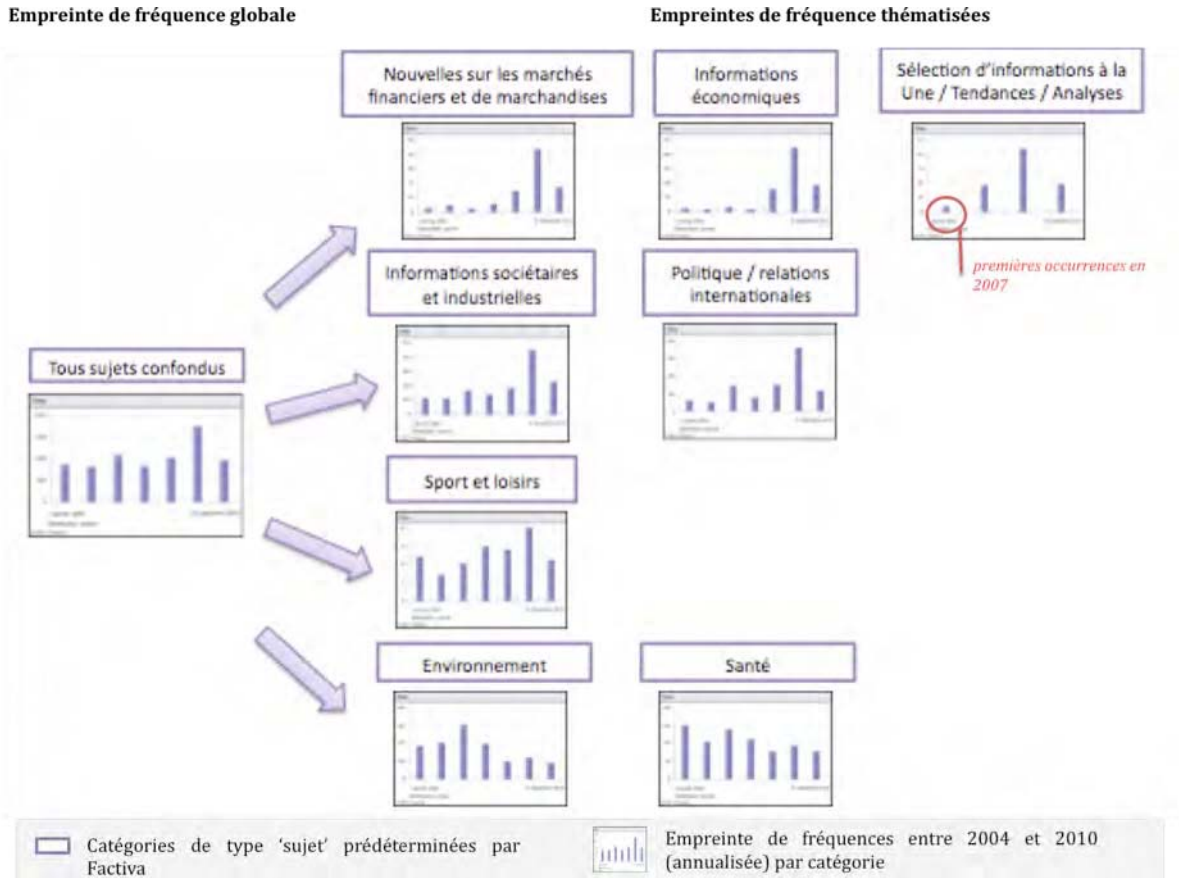


Figure I.3.9 : Empreintes de fréquence de toxique en fonction des domaines d'emploi

L'empreinte de fréquence globale fait apparaître un pic en 2009, mais sinon, pas de réel accroissement des emplois de *toxique* entre la période 2004-2007 et la période 2008-2010. En

[26] Les journaux retenus sont les suivants : *Le Figaro, Les Échos, Libération, La Tribune, Ouest-France, L'Équipe, L'Express, L'Expansion*. Ils ont été sélectionnés sur le critère de leur date d'introduction dans la banque de données : seuls ont été conservés des journaux présents sur l'ensemble de la période étudiée, c'est-à-dire intégrés à la base avant 2004.

[27] Factiva génère automatiquement des histogrammes de fréquences annualisés, en fonction du corpus sélectionné (date, sujet, etc).

revanche, selon le sujet, le profil d'évolution varie fortement : les discours associés à de l'économico-financier et en lien avec de l'événementiel (informations à la Une) montrent non seulement un pic en 2009, mais encore un accroissement sensible sur l'ensemble de la période 2008-2010 par rapport à la période 2004-2007. Nous faisons l'hypothèse que ce profil d'évolution provient des emplois de *toxique* dans le contexte de la crise financière pour qualifier des produits et instruments financiers à l'origine du krach (*actifs toxiques, crédits toxiques*, etc.). La même tendance d'évolution temporelle se reflète, bien que de façon nettement moins marquée, dans les articles associés au monde politique et au monde de l'entreprise ; il semble plausible d'interpréter cela comme une diffusion des emplois précédemment décrits dans des domaines en liens thématiques avec les domaines précédents. En sport et loisirs, une évolution proche se dessine, mais la distance thématique avec les domaines déjà mentionnés laisse supposer un phénomène sémantique distinct du précédent. Au niveau de l'environnement et de la santé, domaines en lien avec l'ancien sens de *toxique*, on assiste plutôt à une baisse des emplois. Ceci pourrait expliquer pourquoi l'empreinte de fréquence globale ne montre pas d'accroissement notable des occurrences de *toxique* : il se crée une sorte de balancier, avec une augmentation marquée dans certains types d'emplois – notamment dans le domaine économico-financier – et une baisse dans d'autres.

2.3 Agencer les niveaux : sémantique et allocation au centre

2.3.1 Une complémentarité plutôt qu'une autonomie

Il n'y a aucune garantie que, considérés isolément, les différents niveaux d'approche d'une néosémie suffisent pour repérer et caractériser le changement de sens. Plusieurs arguments incitent à ne pas exploiter exclusivement un niveau, mais à jouer sur une complémentarité des niveaux.

1. ***Lois de comportement incertaines***. On ne dispose pas de loi de comportement propre aux différents niveaux d'observation. À l'heure actuelle, les indices associés à chaque niveau ont été identifiés, listés, et on a proposé une tendance générale d'évolution de ces indices. Toutefois, il n'y a aucune garantie que cette tendance soit effective. Les indices ont jusque-là été exploités, mais toujours sous l'œil vigilant du linguiste, de façon à faire un choix – rejet ou adoption du candidat repéré à l'aide des indices – et non pour établir un profil d'évolution des indices : le caractère néosémique d'une lexie est souvent incertain, il demande à être étayé et n'a pas la robustesse nécessaire pour servir de base à l'établissement de lois de comportement des indices. On ne cherchera pas à établir des lois de comportement précises et universelles pour chaque niveau – leur existence est improbable car les comportements de néologismes sont trop variables. Il semble préférable de modéliser des tendances générales et de recouper les tendances des différents niveaux : on privilégiera la complémentarité et la convergence, plutôt que la précision et l'exclusivité.

2. ***Des signaux faibles plutôt qu'un signal faible***. Étudier la néosémie revient à étudier des signaux faibles, pour plusieurs raisons. D'abord, le changement de sens résulte souvent d'une polysémisation, le sens ciblé n'est pas le sens principal, mais un nouveau sens, qui n'est pas et ne sera pas nécessairement le sens dominant dans l'usage. Par ailleurs, le nouveau sens est un sens émergent, à ses débuts : plus on cherche à repérer le changement tôt, moins on dispose de matière linguistique. De plus, la néosémie n'est pas un phénomène majeur dans un discours, à l'inverse des thèmes par exemple. Ce problème peut néanmoins être partiellement contourné par le choix d'un corpus approprié. Enfin, nous nous situons en linguistique de corpus, dans une perspective applicative. Ceci implique de travailler sur des données textuelles, de taille limitée, et d'être confronté au problème récurrent d'éparpillement des données (*data sparseness*). Un changement de sens et, par conséquent, les indices témoins de ce

changement, ont donc de fortes chances d'être des signaux faibles. Certains indices sont particulièrement concernés par ce problème, tels que les variantes de forme (comme *outreauiser, outreauisme ; cf.* (Lecolle, 2009:95, note de bas de page)), qui peuvent tout simplement ne pas exister dans les données d'analyse. D'autres sont moins affectés (il existera toujours un environnement cooccurrentiel par exemple), mais ne sont pas pour autant totalement préservés (moindre robustesse des informations, problème de validité de saillances, en particulier statistiques, etc.). D'un changement de sens à l'autre, le niveau défaillant peut varier, donc croiser les niveaux d'observation augmente les chances de faire émerger des informations et de pallier les absences et incertitudes dues aux signaux trop faibles. Il n'est pas exclu de donner la préséance à un niveau d'observation. Dans ce cas, il convient de ne pas se focaliser exclusivement sur ce niveau, mais d'utiliser les autres niveaux à titre complémentaire, pour étayer les informations déjà collectées et suppléer aux insuffisances liées au caractère signal faible.

3. ***Un jeu multiniveau encouragé et validé***. Un certain nombre de travaux incitent à multiplier les niveaux d'observation et à les combiner. Du côté de l'analyse des discours, (Cusin-Berche, 2003) analyse en finesse l'évolution de sens de *décideur* en s'appuyant à la fois sur des réseaux morphosémantiques et sur des relations sémantiques au sein de paradigmes sémantiques. Elle montre à travers son étude que « le sens se construit à partir de l'interférence de plusieurs microsystèmes » et prône ainsi les recoupements et articulations entre niveaux d'observation. (Lecolle, 2007b) adopte une position similaire dans son étude sur l'évolution de sens du nom propre *Outreau* :

> « Avec [nos] angles d'analyse nous proposons autant de points de vue, différents mais complémentaires, destinés à résoudre le problème interprétatif posé – polysignifiance, évolution du sens du Npr –, par le biais d'un environnement étroit ou élargi.»

De fait, ses sept angles d'analyses jouent sur les paliers textuels (du syntagme au texte), les types de relations (cooccurrents qui informent sur les "actants" associés à *Outreau*, tels que *enfants, juge* ; unités lexicales du même paradigme sémantique, comme *affaire Dutroux*), la diffusion du nom propre (indices de notoriété selon la position dans l'article, présence dans les titres d'article…), etc. Les convergences entre angles d'analyse et leurs particularités permettent de tirer des conclusions robustes et nuancées, donc témoignent en faveur d'une articulation des niveaux d'observation. De telles études n'invalident pas l'autonomie des niveaux, mais elles montrent l'intérêt de les recouper.

4. ***Une question d'angle de vue***. Le sens n'est pas directement accessible, mais il nécessite de passer par des représentations, qui restent soumises aux opérations interprétatives. Chaque niveau d'observation donne à voir le changement de sens sous un certain angle. Ainsi, jouer sur plusieurs niveaux revient à fournir à l'interprétant des vues multiples, d'autres éclairages du même phénomène. Donner à l'interprétation des points d'ancrage variés est d'autant plus important que la néologie sémantique est difficile à saisir. Certains éléments guident de façon latente l'interprétation et le sentiment néologique. Matérialiser ces éléments à travers différents niveaux revient à les donner à voir comme la lumière à la sortie d'un prisme : les composantes sont dissociées en faisceaux distincts, chacune a sa couleur – *i.e.* sa contribution – propre, et seul l'ensemble des faisceaux permet de reconstituer la source. Ces idées sous-tendent la réalisation de la plateforme de multi-annotation de (Loiseau, 2007) :

> « Les annotations (…) ont elles-mêmes un statut d'interprétation (…) En résumé, le contexte est entendu ici comme l'articulation, dans un dispositif de codage, d'une pluralité d'annotations qui réifient des interprétations ».

Jouer sur du multiniveau permet donc d'objectiver certains guides latents de l'interprétation et de les mettre en relation : le contexte est disséqué puis recréé sous une autre forme, qui donne un nouvel éclairage interprétatif.

2.3.2 Une articulation délicate

Il y a certes convergence entre les différents niveaux, mais chacun a son apport propre et peut être plus pertinent tantôt pour la détection, tantôt pour la caractérisation. Les profils d'évolution ne sont pas nécessairement les mêmes. De plus, deux à deux, les niveaux ne s'articulent pas nécessairement de la même façon. Il convient donc de s'interroger sur la façon de mettre en relation les différents niveaux. Trois questions se posent pour l'articulation : l'ordre d'exploitation, la hiérarchie des niveaux et la dépendance. Ce paragraphe a pour objectif d'expliciter ces questions. Nous nous positionnerons par rapport à ces questions dans le paragraphe suivant.

1. ***Ordre d'exploitation des niveaux***. Le repérage des néologismes est un préalable incontournable à la caractérisation du nouveau sens. Les indices conscients (indices typographiques, en particulier les guillemets, très répandus ; expressions introductives ou consécutives ; gloses ; formules de nouveauté) sont particulièrement adaptés pour une présélection de candidats à la néologie sémantique. Les autres indices sont à exploiter dans un second temps, pour faire passer du statut de candidat-néologisme à néologisme avéré et pour construire le nouveau contenu sémantique. Dans nos cas d'étude, l'ordre d'exploitation ne sera pas une question prioritaire, car leur finalité ne sera pas de présenter une démarche complète, mais d'approfondir des points particuliers adaptés à nos centres d'intérêt. Cependant, une telle question ne peut être négligée dans un protocole qui se veut complet, dans l'esprit de la plateforme de détection et de caractérisation des néologismes Neologia (Cartier et Sablayrolles, 2008)[28]. Pour l'allocation de signifié, il semble préférable de chercher des changements en allant du macro vers le micro, autrement dit, d'exploiter avant tout des niveaux qui donnent des informations sur des changements "macroscopiques" (nouveau domaine d'emploi par exemple, comme dans le cas de *caviar* (Rastier et Valette, 2009) employé dans le domaine du football pour désigner une belle passe), puis seulement après de rechercher des changements sémantiques plus fins (maintien d'un contexte d'emploi relativement proche et nouvelles nuances de sens ; par exemple, le sens de *niche fiscale* était initialement rattaché à l'idée de vide législatif, qui a évolué vers l'idée de lois à caractère dérogatoire à travers le discours médiatique (Détrie, 2011)).
2. ***Hiérarchie des niveaux***. Aucun niveau n'a priorité absolue sur les autres. La hiérarchie des niveaux dépend du type de néologisme ciblé et des informations considérées comme prioritaires. Par exemple, si la force du sentiment néologique est un critère important pour considérer comme néologique un changement de sens, le poids des indices conscients sera à renforcer. De même, la priorité peut être donnée à une rupture avec des informations sémantiques globales (incompatibilité domaniale par exemple) ou locales (infraction aux règles sémantico-syntaxiques et à la combinatoire). Comme nous le verrons, les priorités que nous nous sommes données nous amèneront à privilégier à l'environnement cooccurrentiel. Dans un cadre légèrement différent, (Yatsko et al, 2010) défendent l'idée que les résumés automatiques de textes sont à adapter selon le genre, et pour ce faire, ils attribuent des pondérations variables aux paramètres (par exemple, la longueur des phrases, les pronoms interrogatifs, les marqueurs de la temporalité, etc.) en fonction du genre. Dans notre cadre, l'allocation de signifié est,

[28] Cette plateforme est à l'heure actuelle encore au stade prototypique.

d'une certaine façon, un résumé ou une synthèse d'emplois. Par analogie, on peut considérer que, selon le rattachement générique du néologisme, les poids accordés aux indices – nos paramètres – sont à adapter. La notion de rattachement générique reste à définir : elle peut renvoyer au procédé à l'origine de la néologie, à la nature des contextes d'apparition, ou encore à l'importance de la rupture néologique. On ne s'attardera pas sur une définition du rattachement générique des néologismes dans l'absolu, qu'on laisse à la réflexion du lecteur, mais on se contentera de proposer une solution adaptée à nos objectifs.

3. ***Dépendance des niveaux.*** Au (chapitre I.1, 1.2.1), nous avions souligné l'importance des interactions entre global et local, et nous avions tendance à privilégier la détermination du local par le global. Ces liens entre global et local existent entre niveaux d'observation et peuvent se manifester sous forme de corrélations ou de dépendances. Par exemple, l'environnement distributionnel peut faire émerger des tendances saillantes propres au niveau global, par exemple des thématiques ou des informations concernant le domaine d'emploi. Ces informations peuvent éclairer l'interprétation de paradigmes issus du foisonnement néologique ou de variantes phraséologiques. Ainsi, pour l'adjectif *toxique* en contexte de crise financière, le paradigme des substantifs participant aux variantes phraséologiques observées dans un corpus sur la crise financière comporte les éléments suivants : {*titres, produits, actifs, créances, crédits*}. Les cooccurrents de *toxique* indiquent par ailleurs que le domaine d'emploi est la finance. Le recoupement de cette information avec le paradigme permet d'interpréter *titres, produits* et *actifs* comme produits ou instruments financiers, ce qui renforce ainsi la cohérence du paradigme et donc la caractérisation du sens de *toxique*. Mettre en relation les niveaux nécessite de disposer de descripteurs qui permettent cette articulation. La solution retenue, à savoir une description des phénomènes à l'aide de sèmes, sera présentée en section 3.

2.3.3 Trois clés d'articulation : allocation, étendue et sémantique

a- Détection vs allocation : privilégier l'allocation

Dans les travaux et réflexions sur l'approche automatisée de la néologie sémantique, la préoccupation principale est celle de la détection, non de l'allocation. Ainsi, dans l'article *Néologie et traitement automatique*, (Mejri, 2010) présente le traitement automatique de la néologie sémantique comme un problème non résolu et il évoque des pistes de recherche de son laboratoire, le LDI…

> « … pour faire en sorte que des indices formels servent de support **au repérage de sens** »

Bien que le titre de l'article ouvre des perspectives plus larges, la visée des travaux est clairement la détection, pas l'allocation de signifié. De même, (Cartier et Sablayrolles, 2008) pointent les faiblesses de Neologia, plate-forme de gestion des néologismes développée par le LDI. Ils mentionnent l'incapacité du système à détecter automatiquement de néologismes sémantiques :

> « Enfin, les néologismes syntaxiques ou sémantiques, pas plus que les néologismes homonymiques ne peuvent **être repérés** de cette façon. »

…mais restent silencieux sur l'absence de procédures automatisées participant à la construction du nouveau sens.

La détection automatique connaît des développements relativement poussés. (Jannssen, 2009) fait le point sur les méthodes de détection des néologismes et met en évidence trois

modes principaux de détection : les listes d'exclusion (non adaptées à la néologie sémantique), les patrons lexico-syntaxiques et les méthodes statistiques. Même si les listes d'exclusion sont les techniques les plus fréquemment mises en œuvre, les deux derniers modes ont été exploités dans des projets, rares pour les patrons linguistiques, mais nombreux pour les méthodes statistiques. Ainsi, (Renouf, 2010) repère les néologismes sémantiques à partir d'écarts statistiques entre profils collocationnels. Ces profils sont obtenus à partir de fenêtres glissantes, dont la taille est de plus ou moins quatre mots. Elle observe par exemple qu'*ethnic* apparaît fin 1992 dans la liste des principaux cooccurrents de *cleansing* dans des emplois du journal *The Times*, alors qu'il était absent de la liste de la période 1989-1992, composée de *street, agent, department, process, thorough, activities*. De même, (Nazar et Vidal, 2010) recourent aux empreintes de fréquence pour détecter la néologie. Pour la néologie sémantique, ils couplent ces empreintes de fréquence à des regroupements de contextes afin de différencier les emplois et gérer le problème de la polysémie.

En revanche, l'allocation automatique de signifié est peu avancée. Les démarches les plus abouties utilisant des procédures automatiques ne mènent pas à bien la question. Ainsi, les systèmes développés par (Renouf, 2010) et par (Nazar et Vidal, 2010) retournent des ensembles évolutifs de cooccurrents qui constituent de la matière brute, mais cette matière reste à raffiner. Certes, les cooccurrents retournés ont un contenu informatif potentiellement riche, mais les apports de ce contenu ne sont pas précisés : les cooccurrents sont-ils tous pertinents pour qualifier le sens ? Quel rapport entretiennent-ils avec le signifié de la lexie ciblée ? Se structurent-ils sémantiquement, par exemple à travers des regroupements ou des oppositions, et comment cette structure affecte-t-elle le sens de la lexie ? Ou encore, quels traits sémantiques doit-on intégrer au nouveau sens ? En bref, ni le statut des cooccurrents, ni les relations des cooccurrents avec le nouveau et l'ancien signifié ne sont précisés. Parallèlement, les travaux dont les efforts ont permis une description plus poussée du nouveau signifié sont essentiellement manuels. Nous avons déjà évoqué les études de (Lecolle, 2009) sur *Outreau*, de (Cusin-Berche, 2003) sur *manageur*, de (Rastier et Valette, 2009) sur *caviar*, dont l'analyse est détaillée, mais manuelle. Le recours à des outils informatiques se limite à l'usage de concordanciers. Dans la même veine, (L'Homme *et al.*, 1999) se concentrent sur la recherche de nouveaux termes à l'aide d'outils informatiques, mais d'une part l'instrumentation de leur méthode est surtout développée en phase de détection (recours à des listes d'exclusion) et, à ce niveau, inadaptée à la néologie sémantique ; d'autre part, la construction du nouveau sens repose sur la lecture et l'analyse manuelle de contextes obtenus à l'aide d'un concordancier, puis sur la recherche d'information – à nouveau manuelle – dans des documents complémentaires. Les auteurs structurent l'information sémantique issue des contextes pour l'articuler à la définition du terme (répartition en éléments définitoires, relations explicites avec des concepts connexes, attestation de relations synonymiques obtenues à partir d'autres sources, etc.), mais cette tâche reste manuelle.

La question de la détection automatique est plus avancée que celle de l'allocation de signifié car elle est plus abordable : il s'agit de trancher s'il y a ou non néologie sémantique. Certes, la réponse peut être nuancée et associée à une échelle graduelle, mais elle reste unidimensionnelle. En revanche, l'allocation de signifié est une question plus délicate, à laquelle peu de réponses ont été apportées. Elle touche à la question de l'accès au sens et de la représentation du sens. Or l'interprétation n'est pas de nature mathématique ou formelle, vouloir automatiser l'allocation de signifié revient à faire le lien entre des tendances a priori incompatibles. De plus, l'allocation de signifié est indissociable des problèmes de désambiguïsation, d'autant plus que nous ne nous situons pas en terminologie, mais en langue générale : la question de la plurivocité y est tenue pour acquise et elle fait partie des impératifs à gérer. L'accès au sens est déjà une tâche délicate en désambiguïsation, elle l'est donc a

fortiori pour construire de nouveaux sens. La difficulté est d'autant plus grande qu'aux problèmes déjà présents en désambiguïsation s'ajoute la question de l'enrichissement sémantique.

L'allocation de signifié est donc complexe, moins avancée que la détection de néologie. C'est précisément pour cette raison que nous y porterons nos efforts.

b- Refléter l'étendue du phénomène

Ce qu'on cible est un phénomène de diffusion, qui a une triple étendue : en temps, dans la langue et en quantité de données.

L'implantation d'une nouvelle lexie est un phénomène qui a une histoire et pour lequel positionnement par rapport au passé est incontournable. Il n'est pas seulement question d'acquérir un sens, mais de le situer par rapport aux sens antérieurs (dans le cas de *caviar* en football, le trait sémantique /remarquable/ ou /haut de gamme/ pour souligner la qualité de la passe est hérité des emplois passés, aussi bien en gastronomie que dans des emplois métaphoriques). Les indices propres à la phase transitoire – indices typographiques, expressions introductives ou consécutives, formules de nouveauté, gloses, foisonnement néologique – ainsi que les variantes morphologiques peuvent guider sur le sens en construction, mais ne mettent pas en relation sens nouveau et sens ancien. En revanche, les réseaux cooccurrentiels (dont les variantes phraséologiques), les réseaux paradigmatiques et les empreintes de fréquences thématiques peuvent renseigner aussi bien sur le sens en construction que sur le sens en amont du changement, puisqu'ils ne sont pas propres à la phase transitoire. Ceci nous incite à mettre plus fortement l'accent sur ces descripteurs.

De même, nous ne nous situons pas en terminologie, mais en langue générale. En langue de spécialité, les thématiques et les domaines sont à peu près définis. En langue générale, ils sont multiples, variables. Or un des mécanismes principaux à l'origine de la néosémie est étroitement lié aux migrations thématiques ou domaniales. La langue n'est pas un tout homogène, elle se construit à partir d'une multitude de pratiques discursives. Ces pratiques discursives peuvent s'associer de façon plus ou moins marquée à un domaine. Parallèlement, certains domaines sont plus productifs en néologismes que d'autres – une étude de (Alaoui, 2008), issue d'un travail d'analyse des domaines producteurs de néologie effectué fin 2000 chez Larousse, identifie l'informatique, ce qui se rattache à la vie sociale, l'économie et la santé comme domaines les plus productifs. Ajoutons que, selon nous, certains couples 'domaine source (celui du discours visé) – domaine cible (domaine de provenance de la néologie)' sont particulièrement favorables à de la production néologique. L'étude des métaphores conceptuelles de (Roy *et al.*, 2005) témoigne de l'association privilégiée de certains domaines – l'emploi métaphorique dans le domaine boursier de lexies issues du domaine de la météorologie est qualifié par les auteurs de "métaphore conceptuelle conventionnelle", ce qui souligne l'affinité particulière de ces deux domaines. Inversement, l'étude de (Drouin *et al.*, 2006) cherche à extraire des néologismes d'un corpus sur le terrorisme et produit peu de néologismes sémantiques ; ce faible retour pourrait s'expliquer par la restriction au domaine du terrorisme, peut-être moins réceptif aux phénomènes métaphoriques, ou trop réduit pour accueillir une diversité des migrations domaniales – hypothèse qui reste à vérifier. La multiplicité des couples 'domaine source - domaine cible' fait de la langue générale un terrain particulièrement fertile pour la néologie sémantique. Le choix d'une approche non terminologique est donc aussi celui d'intégrer ce jeu de configurations domaniales propre à la langue générale, particulièrement favorable à la néosémie. Les indices à cibler en priorité doivent permettre d'accéder au niveau d'information global – c'est-à-dire relevant de la thématique ou encore des domaines. En cela, les cooccurrents issus de paliers textuels larges (comme le paragraphe) semblent particulièrement

à même de refléter de l'information globale, marquée par des caractéristiques textuelles. Les limites des paliers réduits apparaissent chez (L'Homme *et al.*, 1999) : les contextes obtenus par les concordanciers, archétypes de paliers réduits, ne suffisent pas pour proposer une définition des nouveaux termes détectés par leur plateforme et amènent les auteurs à chercher de l'information complémentaire dans des ressources dictionnairiques ou sur le web. De plus, les paliers plus larges sont souvent négligés, alors qu'ils méritent d'être explorés. Cette conclusion ressort chez (Renouf, 2010), lorsqu'elle fait état des limites de son approche : du fait de la taille réduite des fenêtres contextuelles, les cooccurrents éloignés ne sont pas retournés par la procédure de traitement.

Enfin, un des enjeux de l'allocation semi-automatique de signifié est d'aider le linguiste face au foisonnement de données. Deux stratégies peuvent être adoptées : une stratégie de synthèse de l'information (réduction de la masse) et une stratégie d'aiguillage précis vers de l'information pertinente. Les deux stratégies ont leur légitimité et leur intérêt. Nous faisons le choix de privilégier l'approche synthétique, car nous faisons l'hypothèse, certainement discutable, que le regard du linguiste a plus de difficulté à être synthétique (notamment pour faire émerger des informations latentes) que sélectif. Il semble plus utile d'apporter une aide là où la difficulté est maximale. De ce fait, les indices de type 'guides de navigation' (essentiellement les indices dits conscients) dans les données seront relégués au second plan. Les indices qui se caractérisent par une diversité et un foisonnement marqués (les réseaux cooccurrentiels, et dans une moindre mesure, les réseaux paradigmatiques) seront prioritaires.

c- Privilégier le niveau sémantique

Notre objectif est l'allocation de signifié, le sens est donc au cœur de nos préoccupations et il convient de donner priorité à des niveaux d'observation susceptibles d'apporter des informations sémantiques. Par la suite, notre démarche sera orientée par trois considérations :

- Le sens n'a pas toujours d'assise formelle (Mejri, 2010:99) : vouloir allouer automatiquement un signifié exige de choisir un modèle de représentation du sens suffisamment structuré pour se prêter à un traitement automatique et adaptable aux procédés ou mécanismes identifiés comme fondamentaux dans la néologie sémantique.
- La sémantique peut se manifester à tous les niveaux (même des guillemets peuvent apporter de l'information sémantique, par exemple en créant une mise à distance ou, dans des emplois ironiques, en invalidant des traits sémantiques habituellement associés à une unité lexicale), même si elle n'y apparaît pas directement. L'exploitation de tout niveau de description doit se faire de façon à se ramener à de l'information sémantique.
- Le niveau lexical est insuffisant pour décrire la néosémie. Il doit être dépassé pour deux raisons :
    - La néologie sémantique nécessite de décrire des phénomènes en deçà du niveau lexical, puisqu'il s'agit de décrire comment se comporte le contenu sémantique d'une unité lexicale et des phénomènes au-delà du niveau lexical, dans la mesure où interviennent des informations sémantiques caractéristiques de textes ou de discours, tels que les domaines ou les thèmes.
    - La description en unités lexicales n'est pas une description directe du sens. Nous avons évoqué précédemment le problème des cooccurrents, dont la liste ne suffit pas à donner le sens de la lexie ciblée. Certes, quel que soit le mode de représentation du sens, il existera toujours un seuil qu'aucun traitement automatique ne pourra franchir, le seuil de l'interprétation, mais choisir d'autres descripteurs que les unités lexicales est peut-être une solution pour éviter l'amalgame entre unité de sens de la lexie ciblée et unité lexicale associée à la

lexie ciblée. Ces unités lexicales peuvent certes donner des informations sur le sens, elles ne sont pas pour autant partie intégrante du sens de la lexie.

L'enjeu est donc de trouver des outils de représentation du sens capables de décrire le contenu sémantique d'une unité lexicale et de s'appliquer à tous les paliers de la textualité. Pour cela, nous nous positionnons dans un cadre théorique centré sur la sémantique, qui propose un modèle de représentation du sens et qui intègre la variété des paliers sémantiques. Le cadre retenu est celui de la sémantique interprétative.

3. Éléments de modélisation en sémantique interprétative

Pour aborder la néologie sémantique, nous nous appuierons sur le modèle de la sémantique interprétative (Rastier *et al.*, 1994). Cette théorie présente deux avantages principaux :

- elle permet de placer l'information sémantique au centre des considérations ;
- elle propose des modèles de description qui articulent tous les niveaux de la textualité.

La sémantique interprétative est différentielle. Elle accorde une place centrale à la textualité et à la dimension herméneutique. Elle recourt aux sèmes, unités de sens minimales, pour décrire le contenu sémantique de signifiés ainsi que des phénomènes sémantiques survenant à tous les niveaux de la textualité.

Le choix de ce modèle est lié au contexte dans lequel cette thèse s'est mise en place. Elle s'inscrit dans le cadre d'une politique scientifique de laboratoire et participe d'un travail d'équipe, dont un axe d'investigation préalablement défini était l'apport de la sémantique interprétative à la lexicographie. Nous avons épousé ce choix d'orientation pour l'étude de la néologie sémantique et de l'allocation de signifié.

Nous verrons dans un premier temps pourquoi la sémantique interprétative répond à la problématique du multiniveau évoquée à la section précédente, puis en quoi l'utilisation des sèmes s'inscrit dans le respect d'une certaine éthique, ensuite comment la dynamique du sens décrite en sémantique interprétative est transposable à la dynamique néologique, et enfin sur quelle base – lexicale ou sémique – il convient de construire une démarche applicative. L'exemple de *tsunami* illustrera les différentes phases de notre argumentation. Ce substantif se prête à notre problématique car il a connu une diffusion massive accompagnée d'une métaphorisation, pour qualifier des événements marquants de type catastrophe ou chamboulement.

3.1 Les sèmes pour unifier les niveaux et dépasser le lexical

La perspective de veille lexicale du présent travail tend à privilégier le niveau lexical pour accéder aux nouveaux sens. Cependant, les considérations du chapitre I.2 ainsi que celles de la section précédente invitent à relativiser l'apport du niveau lexical pour y intégrer celui d'autres niveaux. Nous verrons dans un premier temps en quoi les sèmes créent une articulation entre les différents niveaux de description de la néologie ; nous nous attarderons ensuite sur l'apport des sèmes sur un plan infra-lexical, c'est-à-dire pour décrire le contenu d'une lexie, puis sur le supra-lexical, pour décrire les apports de la textualité, et nous évoquerons brièvement le rôle des sèmes à d'autres niveaux (syntaxiques, morphologiques, typographiques).

3.1.1 Intégrer les niveaux plutôt que fragmenter

Traditionnellement, il existe différents niveaux de description des phénomènes linguistiques, notamment la syntaxe, la morphologie et la sémantique, particulièrement

prégnants à l'écrit. D'une certaine façon, les niveaux de description linguistique font écho à ce que nous avons qualifié de niveaux pour caractériser la néologie sémantique : ils proposent des angles d'approche distincts et, selon la perspective, ils permettent soit de décomposer un phénomène en modules disjoints (perspective dissociatrice), soit de proposer une description du phénomène selon des facettes articulées et formant un tout cohérent (perspective intégratrice). La sémantique interprétative s'inscrit dans une perspective intégratrice et (Rastier *et al.*, 1994) souligne sa visée d'unification :

> « Nous souhaitons donner à la sémantique linguistique toute la place qui lui revient, en unifiant la description du lexique, de la syntaxe profonde, et des structures textuelles. À chacun des trois paliers traditionnels de la description linguistique (mot, phrase, texte), nous faisons alors correspondre trois paliers de la théorie sémantique (*micro-, méso-* et *macrosémantique*) en unifiant leur conceptualisation. » (Rastier *et al.*, 1994:25)

> « La linguistique connaît une organisation en niveaux, et la plupart des théories distinguent les niveaux phonologique, syntaxique, sémantique, pragmatique. Pour beaucoup de théories, au premier rang desquelles la grammaire chomskyenne, ces niveaux justifient la construction de modules théoriques ou composantes autonomes. La plupart des systèmes classiques de traitement automatique du langage qui se réclament de l'IA étagent ainsi des modules, non sans difficultés théoriques et pratiques. Pour éviter ces difficultés oiseuses, nous avons proposé d'unifier la syntaxe profonde, la sémantique et la pragmatique intégrée. » (Rastier *et al.*, 1994:35-36)

L'ambition de la sémantique interprétative rejoint notre volonté de mettre en relation les différents niveaux de modélisation de la néologie sémantique et de les aborder comme parties d'un ensemble cohérent.

De plus, la problématique des niveaux est transformée en problématique des paliers :

> « La réduction des niveaux permet de mieux aborder l'organisation en paliers. Sur chaque niveau linguistique, on peut pratiquer l'analyse à trois principaux paliers de complexité : le mot, la phrase et le texte. Les deux premiers paliers ont jusqu'ici été privilégiés par la tradition grammaticale et linguistique. » (Rastier *et al.*, 1994:36)

Les différents paliers sont en interaction. Pour les articuler et mettre en évidence des phénomènes qui relèvent aussi bien de la sémantique de l'unité lexicale (*i.e.* des états ou phénomènes sémantiques internes à l'unité lexicale) que de la sémantique du texte, il est fondamental de disposer de descripteurs de granularité sémantique variable. Les descripteurs proposés, à savoir les traits sémantiques ou *sèmes*, sont adaptés à tous les paliers : ils peuvent s'appliquer aux paliers inférieurs à l'unité lexicale – ce qui est indispensable pour la néosémie puisqu'il faut disposer d'un contenu sémantique d'une unité lexicale – et aux paliers larges – qui sont également nécessaires pour refléter l'influence du contexte. Les niveaux s'articulent autour des paliers, c'est-à-dire selon qu'ils relèvent du global ou du local. Ainsi, la sémantique interprétative et en particulier le recours aux sèmes offrent une clé d'articulation des niveaux d'indices participant à la néologie sémantique.

L'approche de la sémantique interprétative se démarque d'autres approches en cela qu'elle accorde un rôle prépondérant au palier textuel :

> « … pour une sémantique interprétative, le palier du texte est primordial, puisque c'est la connaissance des caractéristiques du texte qui permet d'assigner du sens à la phrase et au mot. » (Rastier *et al.*, 1994:36)

Le rôle majeur du palier textuel se reflète dans le mode d'interaction des paliers, selon le principe fondamental que le global détermine le local. Ce principe est fondamental en

sémantique interprétative, de même qu'il l'est dans notre cadre, puisque c'est à travers des emplois répétés de l'unité lexicale dans des contextes inattendus, cadre global, que s'impose le nouveau sens, transformation locale.

Les interactions entre global et local décrites par la sémantique interprétative et sa capacité à unifier les différents niveaux sont non seulement affirmées en théorie, mais aussi pensées en termes applicatifs (Tanguy, 1997 ; Pincemin, 1999) et attestées dans des cas d'étude. Ainsi, dans le projet PRINCIP, (Valette, 2004) s'appuie sur le formalisme de la sémantique interprétative pour contraster sémantiquement des sites racistes et antiracistes. Il recourt à tous les paliers de la textualité : macrosémantique à travers un profilage en genres par exemple, reflets d'un "implicite inconscient", ou d'informations sémiotiques infratextuelles telles que la police de caractère ou l'utilisation d'images ; mésosémantique avec des récurrences de traits sémantiques (par exemple, celles de /maladie/ dominent dans les textes racistes) ou la mise en évidence de thèmes ; microsémantique à travers la décomposition de lexies et l'étude de morphèmes ou lexèmes particuliers (morphème *-phobe* plus présent dans les sites antiracistes, lexème *démocr-* typique des sites associés au racisme, qu'ils soient pour ou contre, et équitablement réparti entre les deux types de sites). Les multiples angles d'approche s'inscrivent dans un tout cohérent, fondé sur la sémantique interprétative, où interagissent les différents paliers d'observation.

3.1.2 Sèmes et infra-lexical : de la matière et une structure pour le contenu sémantique

Allouer un signifié nécessite, évidemment, une représentation du signifié, donc cela exige de se situer à un niveau de description infra-lexical. Les sèmes sont un moyen d'accéder à ce niveau.

Le contenu sémantique des unités lexicales, c'est-à-dire le contenu de leur signifié, est constitué d'un ensemble de sèmes. Cet ensemble de sèmes est qualifié de *sémème* s'il s'agit du contenu du signifié en langue et de *sémie* pour le contenu du signifié en discours (Valette, 2009:24).

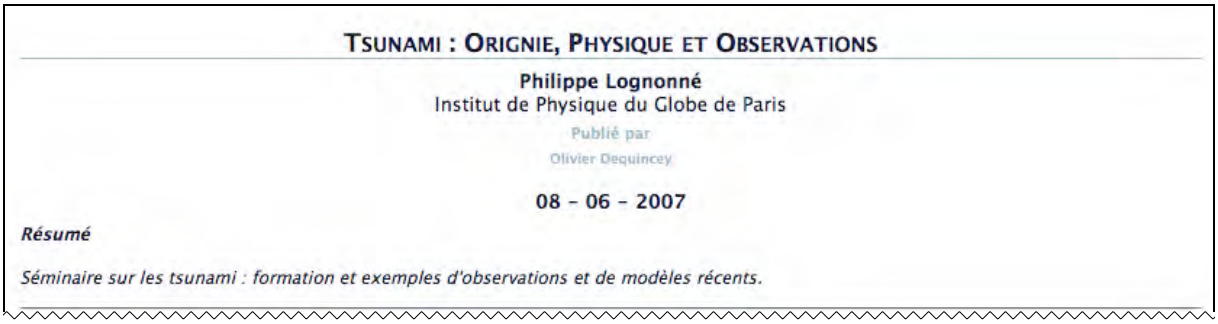
Par exemple, la définition lexicographique de *tsunami* dans le *Trésor de la langue française informatisée* peut être considérée comme une lexicalisation de son sémème :

> **TSUNAMI, subst. masc.**
> *SC. DE LA TERRE.* Onde océanique solitaire, immense vague ayant pour origine un tremblement de terre, une éruption volcanique sous-marine ou la chute dans la mer de grands pans de falaises ou de glaciers, et provoquant de graves dégâts quand elle déferle sur une côte. Synon. *raz* de marée*. *Lorsque les observations séismologiques ont signalé une éruption ou un séisme sous-marin dans le Pacifique, il est possible, à l'aide de cartes de propagation, de déterminer l'heure à laquelle le* tsunami *engendré atteindra les îles Hawaii* (*La Terre*, 1959, p. 475 [Encyclop. de la Pléiade]). P. métaph. *Quelle immense marée d'informations bonnes la science même pourra-t-elle intégrer avant de rouler dans la barbarie, emportée par cet horrible tsunami d'articles?* (M. Serres, *Genèse*, 1986 [1981], p. 218).

L'emploi de *tsunami* est associé au domaine des sciences de la terre. Le sens métaphorique est signalé, mais semble correspondre à un potentiel plus qu'à un emploi intégré dans la langue : d'une part, il ne possède pas de définition propre, d'autre part, l'exemple est rédigé dans un style très littéraire, extrêmement travaillé, où *tsunami* intervient pour filer une métaphore déjà amorcée.

Le sémème associé à cette définition comporterait des sèmes étroitement liés aux sciences naturelles, tels que /mer/, /vague/, /terre/, /côte/, ou, plus généralement, //phénomènes et éléments naturels//, et des sèmes associés à un caractère //négatif//, comme /danger/, /provoquer des dégâts/, ou à l' //ampleur// du phénomène, pour lequel tout est /immense/, /grand/, /grave/. Dans l'emploi ci-dessous de *tsunami* provenant d'un site d'enseignement des

sciences de la Terre[29], la sémie met au premier plan les //phénomènes et éléments naturels// et autres sèmes associés :



TSUNAMI : ORIGNIE, PHYSIQUE ET OBSERVATIONS

Philippe Lognonné
Institut de Physique du Globe de Paris

Publié par
Olivier Dequincey

08 - 06 - 2007

Résumé

Séminaire sur les tsunami : formation et exemples d'observations et de modèles récents.

Dans cet exemple, le tsunami apparaît comme un phénomène physique, objet scientifique, observable et mesurable. Autrement dit, il apparaît comme quelque chose de concret, objectivé (les aspects connotatifs, tels que le caractère négatif, sont éclipsés) et rattaché aux sciences de la nature. À l'inverse, dans l'exemple qui suit, la sémie est très éloignée de celle qui vient d'être décrite. L'auteure d'un blog culinaire accaparée par la création en parallèle d'une boutique en ligne (la 'cook-shop') écrit :

> « Il semble que ce blog ait subi la désertion de son auteure. Pas faux. Le **tsunami** cook-shop qui a bousculé ma vie, m'a mise à plat, remisant au placard tout le superflu et même l'essentiel. »
>
> (site : http://www.lignepapilles.com/2011/01/04/et-une-nouvelle-annee-une/)

L' //ampleur// ressort fortement, voire, éventuellement, un côté négatif, du type /provoquer des dégâts/ (blog délaissé au profit de la boutique en ligne), mais les //phénomènes et éléments naturels// et autres sèmes associés aux sciences de la terre sont exclus de la sémie.

Dans notre cadre, la distinction entre *sémème* et *sémie* peut être source de confusion, voire même être inopérante : la néologie sémantique est un phénomène qui fait interagir sémème et sémie, puisqu'il se crée un nouveau signifié en langue sous l'effet des valeurs prises en discours. La néosémie est un processus d'interaction, ou plus exactement de rétroaction des sémies sur le sémème. La position en phase de transition implique donc un flou terminologique. Par la suite, on aura tendance à parler de *sémème* (contenu sémantique lexicalisé ou fruit d'une mise à jour) ou de *signifié* (dans le cadre d'interactions) , car cette terminologie est plus adaptée à la perspective de stabilisation et de lexicalisation : on part d'un sens codé pour aboutir à un nouveau sens codé, autrement dit, ce dont on part et ce qu'on cherche à obtenir est un sémème, non une sémie.

Disposer d'une représentation du contenu sémantique est fondamental pour formaliser la néologie sémantique : le nouveau sens s'établit généralement par polysémisation, donc à partir de pivots, parties de l'ancien signifié qui créent le lien avec le nouveau sens, et à partir de nouveaux traits sémantiques. La représentation en sèmes permet de voir la néosémie comme un jeu sur des ensembles : des sous-ensembles du sémème sont activés en discours et, à l'issue du processus, l'ensemble de départ s'enrichit en nouveaux sèmes, en lien avec les sous-ensembles activés en discours. Considérons les tablettes numériques, aussi appelées *ardoises numériques*. Cette désignation renvoie à des supports multimédias : ce sont des micro-ordinateurs plats, tactiles, respectant un format qu'on pourrait qualifier d'intermédiaire, c'est-à-dire qui restent portables tout en dépassant le format poche. Le substantif *ardoise* dans le domaine du numérique conserve les sèmes /plat/, /format rectangulaire/, /transportable/,

[29] Site source : http://planet-terre.ens-lyon.fr/planetterre/XML/db/planetterre/metadata/LOM-tsunami-Lognonne_conf.xml

/interface/ ou encore /support d'écriture/ de l'ardoise de l'écolier et se voit ajouter /informatique/, /nouvelles technologies/, /tactile/. Les traits sémantiques /craie/ et /école/ disparaissent dans ces emplois.

Le sémème peut se concevoir comme un ensemble d'unités différenciées, doté d'une organisation interne. La structure correspondante n'est pas rigide, elle reste déformable lors des actualisations en discours. En sémantique interprétative, on distingue généralement les sèmes spécifiques, qui singularisent une unité lexicale et permettent de l'opposer à d'autres unités lexicales, et les sèmes génériques, communs à un ensemble d'unités lexicales. Les notions de généricité et de spécificité n'ont normalement de sens que dans des corpus, où interviennent plusieurs unités lexicales. Cependant, nous considérons ces notions comme pertinentes pour structurer le signifié d'une unité lexicale pour trois raisons :

1) L'idée que certains sèmes servent de clés de regroupement à d'autres sèmes reste sensée et elle est transposable au sémème d'une unité lexicale.

2) Le sens d'une unité lexicale n'est pas codé isolément, il s'intègre à un ensemble plus vaste (un dictionnaire par exemple) qu'on peut considérer comme un corpus, comme le font (Valette *et al.,* 2006) lorsqu'ils structurent en classes sémantiques des définitions dictionnairiques.

3) Nous verrons plus loin que les textes font ressortir des groupements structurés récurrents, les *formes sémantiques*, susceptibles de se lexicaliser de façon privilégiée à travers la lexie néosémique. Lors de la lexicalisation, la conservation d'une telle structure plutôt qu'une déstructuration complète semble plus cohérente : comme toute unité a été néologique à un moment de son histoire, elle a intégré une structure sémantique, donc le sémème peut être vu comme doté d'une structure interne.

Le sémème d'une unité lexicale peut se structurer en classes. Ces classes sont représentées par des sèmes génériques de plus ou moins grande généralité (Ballabriga, 2005), macrogénériques pour le degré de généralité maximal (pour l'opposition /abstrait/-/concret/, /animé/-/non animé/ etc.) , mésogénériques au stade immédiatement inférieur (domaines d'appartenance) , microgénériques pour les plus petites classes sémantiques. Les sèmes qui ne contribuent pas à regrouper mais qui servent à différencier sont spécifiques. Pour une lexie, ces différences entre sèmes participent de la structuration du sémème. Ainsi, *tsunami* comporte le sème macrogénérique /concret/ – la métaphorisation se reflète entre autre à travers un jeu sur ce sème ; le sème mésogénérique //sciences de la terre// pour évoquer le domaine de rattachement ; des sèmes microgénériques tels que //phénomènes et éléments naturels// ; des sèmes spécifiques tels que /vague/ ou encore /déferlement/. Le sémème de *tsunami* pourrait être représenté avec la structure présentée en figure I.3.10.

La structuration pourrait répondre à d'autres critères et refléter une autre configuration. Par exemple, on pourrait intégrer la distinction entre traits de sélection et traits inhérents, c'est-à-dire des sèmes de liaison participant aux règles sémantico-syntaxiques et des sèmes de contenu (Bastuji, 1974). Nous ne nous attarderons pas sur ces questions, car nous avons choisi d'appuyer notre analyse du sens sur des contrastes et saillances, non sur des dépendances syntaxiques et l'approche linéaire qu'elles impliquent.

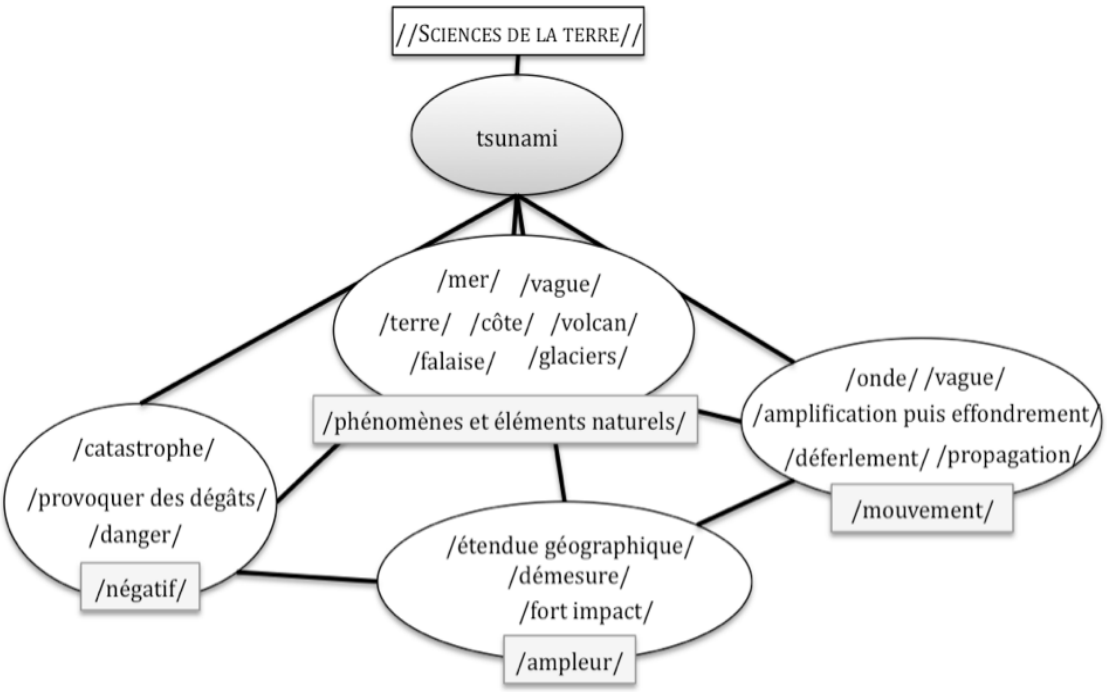


Figure I.3.10 : Sémème structuré de tsunami

3.1.3 Sèmes et supra-lexical : fonds et formes sémantiques pour exprimer le global

Les sèmes ne sont pas seulement destinés à exprimer le contenu sémantique d'une lexie, ils sont adaptés pour décrire la sémantique de tous les paliers de la textualité. Ils peuvent donc exprimer du contenu sémantique caractéristique de paliers supérieurs à celui de la lexie, notamment du contenu sémantique propre à des unités textuelles larges.

La sémantique des unités textuelles est exprimée à travers les *fonds sémantiques* et les *formes sémantiques*.

Les *fonds sémantiques* expriment un contenu sémantique qui relève du niveau global (thèmes, domaines, genres), pour le palier du corpus de textes, ou des paliers plus réduits, comme un texte ou le passage d'un texte. Ils se manifestent sous forme diffuse à travers des isotopies (récurrences de sèmes) et des faisceaux d'isotopies, ou sous forme de structures à travers les thèmes génériques[30] (Rastier, 2006). Les isotopies peuvent relever d'une globalité large, mais aussi d'une globalité plus réduite, auquel cas on parlera d'isotopies locales. Par exemple, les isotopies peuvent être caractéristiques d'un voisinage tel que le paragraphe, comme dans l'exemple déjà évoqué de /botanique/ dans un extrait de Flaubert (Valette, 2008).

Les *formes sémantiques* sont des "figures qui contrastent sur des fonds", elles apparaissent à travers des groupements structurés de traits saillants, les *molécules sémiques* (Rastier *et al,* 1004:130). Elles peuvent être diffuses ou compactes (Rastier, 2006). Par exemple, (Roy *et al.*, 2005) étudient conjointement trois métaphores conceptuelles, ayant pour même domaine cible la bourse et dont les domaines sources sont la météorologie, la guerre et la santé. Ils cartographient l'évolution des domaines sources dans un corpus boursier. De cette façon, ils

[30] Un thème est une "structure stable de traits sémantiques (ou *sèmes)*, récurrente dans un corpus, et susceptible de lexicalisations diverses" (Rastier, 1996). Un thème se rattache aux fonds ou aux formes sémantiques selon qu'il est générique ou spécifique, et il reste associé à une certaine globalité, c'est-à-dire à des paliers textuels moyens ou larges. Un thème générique est un "fond sémantique constitué par la récurrence d'un ou plusieurs sèmes génériques", un thème spécifique est une "molécule sémique relevant du palier mésosémantique" (Missire, 2006). Les thèmes génériques se distinguent des isotopies par leur structure.

étudient la configuration d'une forme sémantique définie par le regroupement de traits sémantiques /météorologie/, /guerre/, /santé/ sur le fond sémantique de la /bourse/.

Les fonds et formes sémantiques ne s'inscrivent pas dans une logique compositionnelle, ils permettent de s'affranchir de la linéarité textuelle et de faire ressortir des saillances globales. Les interactions entre fonds et formes sémantiques mettent en œuvre un jeu de contrastes, qui eux-mêmes déterminent l'interprétation locale.

Les fonds et formes sémantiques s'accommodent de lexicalisations diverses, et en cela, ils permettent de dépasser le niveau lexical, donc de se rattacher à un plan supra-lexical. Ainsi, dans le paragraphe suivant extrait d'un article rattaché au thème de la crise financière de 2008, on peut dégager un faisceau d'isotopies rattachées à la //finance//, à l'idée de /provoquer des dégâts/ et à celle de /propagation/. L'isotopie financière, par exemple, se lexicalise à travers les unités lexicales *capitalisme, économie réelle, zone euro, dollars, banques, crédit, financement* et *argent*.

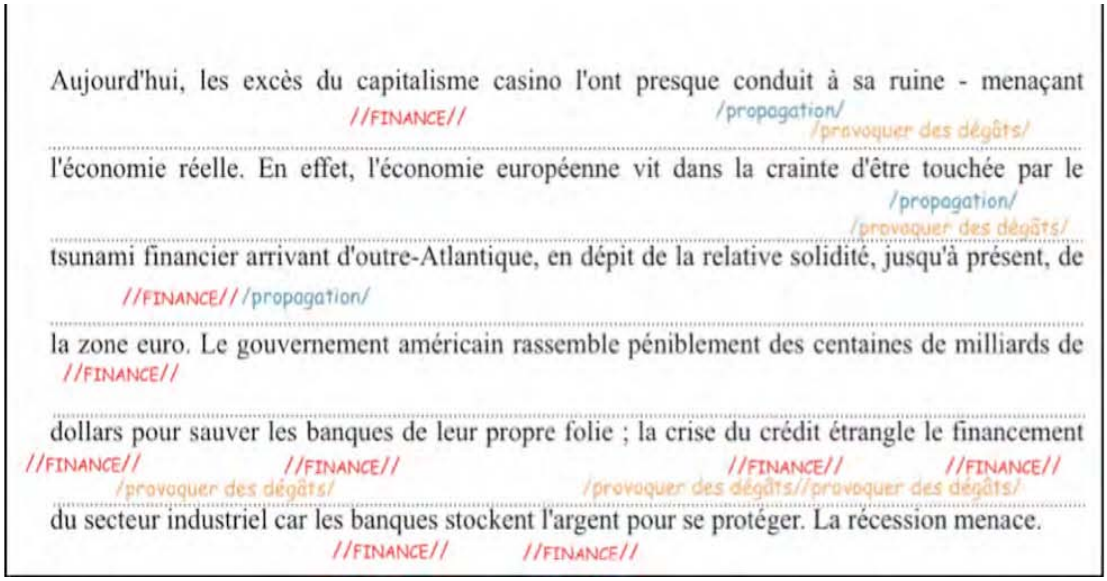


Figure I.3.11 : Faisceau d'isotopies dans un paragraphe contenant économie réelle

Les fonds et les formes sémantiques sont des outils de description des phénomènes sémantiques adaptés à l'étude de la néologie sémantique :

- Ils constituent des unités caractéristiques du niveau global. Les contextes d'emplois jouent un rôle déterminant dans la néosémie, il est donc particulièrement intéressant de disposer d'entités caractéristiques de ces contextes d'emplois. Ces entités sont une façon de représenter des sources de rétroaction du global sur le local.
- Ils se veulent le reflet de la perception sémantique. La perception sémantique n'est pas sans évoquer le sentiment néologique, qui participe aussi d'un ressenti et s'appuie sur des choses latentes, difficiles à repérer et à localiser. À travers les formes et fonds sémantiques, la sémantique interprétative s'attaque à ce qui est diffus et qui peut affecter le ressenti. Cette volonté descriptive n'est pas une garantie d'efficacité, mais elle témoigne d'une convergence dans la démarche de la sémantique interprétative et la nôtre.
- Ils construisent la sémantique des textes à travers des jeux de contrastes. La néosémie fait aussi intervenir des jeux de contrastes, par exemple à travers des ruptures domaniales.
- Ils peuvent être instanciés par une variété d'unités lexicales. Cette variabilité de la lexicalisation n'est pas sans évoquer le foisonnement néologique ou la question de la structuration des cooccurrents. Formes et fonds sémantiques sont une façon de créer une

unité. De plus, ils s'inscrivent pleinement dans une démarche qualitative. Comme le souligne (Rastier, 1996) :

> « La qualification des cooccurrents est cruciale, car elle permet le passage du quantitatif (les cooccurrents) au qualitatif (les corrélats). […] Ainsi, les cooccurrents ne sont élevés à la dignité de corrélats que s'il est possible d'établir une relation d'isotopie ou de paratopie[31] avec d'autres cooccurrents. Par exemple, parmi les cooccurrents d'*ennui*, *dimanche* et *araignée* se sélectionnent mutuellement, dans le contexte d'*inaction*. Ils lexicalisent un des composants du thème recherché, et c'est à ce titre qu'ils sont qualifiés. »

- Les unités méso- et macrosémantiques permettent donc de créer des liens sémantiques entre unités lexicales, donc de qualifier les liens. Elles ouvrent des perspectives pour la structuration des unités les plus riches mais aussi les plus diverses qui interviennent dans la néosémie, à savoir les cooccurrents et les concurrents provenant du foisonnement néologique.

3.1.4 Sèmes et dépendances

Jusque-là, nous avons vu que les sèmes permettent une description du sens non linéaire, qui joue sur des contrastes entre fonds et formes. Les interactions entre le global et le local en font des outils adaptés à une perspective gestaltiste ou impressionniste. Cependant, l'appareil théorique de la sémantique interprétative a un potentiel de description plus riche : il peut exprimer des mécanismes linéaires, à travers des jeux de dépendances.

(Bastuji, 1974) propose deux sortes de traits sémantiques intervenant dans la néologie sémantique : les traits inhérents et les traits de sélection. Les traits de sélection sont définis à partir de propositions de (Chomsky, 1965) : ils sont des traits syntaxiques et ils régissent la combinatoire des unités lexicales, notamment les relations prédicats-arguments. Ils sont associés aux verbes et adjectifs, mais pas aux substantifs.

Cette dichotomie de (Bastuji, 1974) amène à envisager les sèmes comme l'expression de dépendances sémantiques, perspective complémentaire de l'expression de contenus. Cette perspective implique une restriction à du local, mais aussi une intégration des relations syntaxiques. De même que, pour les saillances globales que sont les cooccurrents, les sèmes servent à qualifier les liens cooccurrentiels et à les transformer en corrélations, de même, pour les associations locales des relations syntaxiques, les sèmes peuvent servir à qualifier les liens. Le nouveau signifié proviendrait alors d'une modification des liens sémantico-syntaxiques.

Les fonctions lexicales de la théorie Sens-Texte (Mel'čuk, 1997) sont une façon d'envisager les sèmes en tant que traits de sélection. Elles décrivent des liens paradigmatiques, comme la synonymie, ou des liens syntagmatiques entre des prédicats et des collocatifs, par exemple des liens d'intensification ou des liens causatifs. Une définition de dictionnaire, donc la représentation d'un signifié, s'exprime comme l'ensemble des liens sémantiques (paradigmatiques ou syntagmatiques) caractéristiques d'une lexie.

Ce formalisme permettrait de concevoir sous un autre angle les mécanismes sémantiques à l'œuvre. Par exemple, reprenons la variation sémantique du *corbeau blanc* de (Rastier *et al.*, 1994) : l'adjectif *blanc* contribue à inhiber le sème inhérent /noir/ de *corbeau.* Le lien contrastif entre /noir/ et /blanc/ permet l'inhibition et c'est l'élimination forcée du trait /noir/ qui focalise l'attention sur la couleur. Les mécanismes de variation sémantique pourraient se transposer en termes de fonctions lexicales :

[31] paratopie : "relation entre les diverses lexicalisations partielles d'une même unité mésosémantique ou macrosémantique." (Missire, 2006). Les formes et fonds sémantiques sont des unités méso- ou macrosémantiques.

corbeau noir = Couleur(corbeau) ;

corbeau blanc = corbeau Anti(noir) = Anti(Couleur(corbeau)) = Anticouleur(corbeau).

Le changement devient explicite : le formalisme des fonctions lexicales met en relief à la fois le focus sur la couleur et l'inhibition du trait sémantique usuellement associé à la couleur du corbeau.

Pour de la néosémie, l'allocation d'un nouveau signifié correspondrait à une modification des liens syntagmatiques et/ou paradigmatiques. Dans ses emplois métaphoriques, *tsunami* est prédicat, puisqu'il sert à exprimer quelque chose au sujet de grands événements ou d'entités : finance / crise financière (*tsunami financier*), victoire électorale de l'UMP aux présidentielles de 2007 (le *tsunami bleu*), etc. Dans *tsunami financier,* l'adjectif *financier* est un adjectif relationnel, l'emploi du complément du nom (*tsunami de la finance*) exprimerait la même idée. *Financier* joue donc le rôle d'actant plus que de prédicat. Le lien entre *tsunami* et *financier* pourrait s'exprimer à l'aide de fonctions lexicales : l'ampleur et le caractère négatif de *tsunami* au sens métaphorique ne sont pas sans évoquer les fonctions Magn et Antibon, même si le formalisme précis pour relier *tsunami* à *financier* reste à définir : il faudrait préciser la combinaison des deux fonctions et ajouter d'autres éléments pour une relation définie rigoureusement. Le signifié de *tsunami* s'enrichit donc en liens syntagmatiques à travers les fonctions Magn et Antibon. De même, les liens paradigmatiques changent : alors que *raz-de-marée* était le synonyme attitré de *tsunami*, d'autres viennent le concurrencer dans les emplois. Ainsi, *tempête* et *tourmente*, dont les emplois métaphoriques sont similaires à ceux de *tsunami* dans le contexte de la crise financière, pourraient s'ajouter à la liste des unités en lien de synonymie avec *tsunami*. Le changement de sens peut donc se voir non plus seulement comme une modification de contenu, mais encore comme une modification de la combinatoire.

Notre priorité ici est l'articulation de la sémantique avec le lexical, en privilégiant les saillances et les contrastes, on ne développera donc pas plus les aspects évoqués dans ce paragraphe. Cependant, il convient de garder en mémoire que la description en sèmes peut être étendue à d'autres niveaux que le niveau lexical. Les informations apportées par ces autres niveaux sont complémentaires de notre perspective. La contribution de chaque niveau et leur articulation à la sémantique interprétative sont des questions à part entière, qui mériteraient d'être approfondies dans des travaux ultérieurs.

3.2 Des descripteurs pour guider l'interprétation

Les sèmes sont des outils pour l'analyse, qui respectent une certaine philosophie. Face à la question de la représentation du sens et de l'accès au sens, ils ne sont pas proposés comme la solution, mais comme des auxiliaires – d'où une certaine « humilité du sème » (Valette, 2010b:186-187) :

> « Ils ne sont pas des universaux, ils sont des traits sémantiques, des valeurs modestes élaborées, construites par le linguistique 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 indénombrabilité (…) et surtout leur dépendance déterminante à la langue, la parole, la pratique et la culture. »

Les sèmes sont donc pensés comme relatifs et non pas absolus :

- Ils ne sont pas des universaux. Ils sont au contraire relatifs, au moins à chaque langue (Pincemin, 1999:282). Ils se veulent le fruit d'une interprétation, ce qui s'oppose à toute

universalité (Tanguy, 1997:44). (Rastier *et al.*, 1994:88) insiste sur le fait que la construction d'un lexique sémantique doit s'appuyer sur l'usage et les discours :

> « Dans l'approche interprétative, la différence trouve sa justification en contexte et il est impossible de fonder linguistiquement des différences entre des éléments arbitraires hors de tout contexte réel. (…) La description du lexique sémantique repose sur l'étude d'un corpus représentatif des textes à traiter par l'application »

- Cette relativité des sèmes amène à concevoir l'allocation de signifié selon une certaine ligne de force : le nouveau sens codé apparaît comme le fruit d'un consensus, il est le reflet d'un savoir partagé, pas d'un absolu. Il témoigne d'une convergence interprétative et se veut le produit de l'usage.
- Ils ne sont pas des primitives. Les sèmes ne s'inscrivent pas dans une perspective atomiste, où ils constitueraient des unités premières et irréductibles, mais plutôt dans une perspective connexionniste. Les sèmes sont des unités de différenciation, contingentes et non minimales. Le sème est la plus petite unité de signification définie *par l'analyse* (Rastier *et al.*, 1994:224), pas la plus petite unité dans l'absolu, et « de niveau de détail ajusté aux besoins de l'analyse » (Pincemin, 1999:282).
- Ils ne sont pas donnés, mais construits. Comme produits de l'interprétation, ils résultent d'un processus dynamique. A priori, cet aspect de la nature des sèmes s'accorde mal avec une approche informatisée, nécessairement déterministe (Pincemin, 2002). Cependant, il est pertinent si on adopte une vue plus surplombante : les sèmes utilisés à un temps t par un programme informatique sont le résultat de constructions antérieures. De plus, la construction des sèmes à partir des pratiques discursives répond à la problématique d'intégration des contextes d'emploi au sens d'une unité lexicale, autrement dit d'intégration locale du global.
- Ils sont potentiellement en nombre infini, car toute paraphrase peut formellement correspondre à un sème. Les sèmes sont certes des unités conçues avec souplesse, avec un potentiel de variation infinie, mais dans toute application informatique, ils sont définis au préalable, ce qui fixe leur nombre et leur retire cette souplesse de configuration. Le caractère intrinsèquement indénombrable des sèmes amène à relativiser les représentations obtenues : celles-ci résultent d'un choix, les sèmes ne sont pas des absolus mais des guides.

La nature théorique du sème respecte d'une part le fait que le sens échappe à la formalisation (la conception du sens mathématisable de (Harris, 1968) est révolue, on est aujourd'hui conscient des limites de la formalisation, en particulier du fait que l'interprétation reste l'apanage de l'humain), d'autre part le rôle central accordé à la dimension interprétative. En effet, les sèmes apparaissent comme des descripteurs, ils ne correspondent pas à une vision imposée de la langue, et plus largement du monde, mais ils sont des auxiliaires pour l'interprétant.

Par ailleurs, les sèmes se caractérisent par une souplesse et un caractère relatif. Plutôt que d'y voir une incompatibilité avec une approche automatisée, qui s'accommode difficilement de flou et d'incertitude, il faut se servir de ces caractéristiques pour l'interprétation des résultats. Les sèmes doivent être envisagés non pas comme des points, mais comme des espaces ouverts ; non comme le sens, mais comme des potentiels de sens ou encore des tendances sémantiques ; non comme des absolus, mais comme des guides.

L'humilité du sème se mesure donc moins en termes d'efficacité pour une démarche applicative que de philosophie dans laquelle s'inscrit leur utilisation. En pratique, les sèmes restent des unités complexes, qui présentent autant de difficultés que les unités (lexicales, morphologiques, syntaxiques) utilisées pour la représentation du sens.

3.3 La dynamique sémique au service de la dynamique néologique

En sémantique interprétative, le rôle central accordé à la textualité est aussi une prise de position en faveur d'une dynamique du sens : le sens n'est pas donné a priori, mais il est en construction permanente, à travers les *parcours interprétatifs*. La notion même de parcours interprétatif exclut toute idée de statisme.

La dynamique sémique se manifeste à la fois à l'échelle textuelle et à l'échelle d'une unité lexicale. Le sens d'une unité lexicale se construit par interaction entre son signifié et les unités sémantiques représentatives du global – c'est-à-dire les fonds et formes sémantiques. Deux mécanismes principaux sont à l'œuvre : l'actualisation et la propagation.

3.3.1 Actualisation et reconfiguration du signifié

L'actualisation du sens d'une unité lexicale se manifeste à travers l'activation ou l'inhibition de sèmes. Les activations et inhibitions sémiques peuvent participer à une reconfiguration du signifié à l'origine d'une néosémie. Ainsi, dans l'exemple de *mutualiser*, l'inhibition récurrente du domaine de l'/assurance/, couplée à une activation de la /mise en commun/ ont participé à la néosémie.

Si un sème participe à une isotopie locale, il aura tendance à être activé. Au contraire, l'absence de récurrence d'un trait sémantique pourra avoir tendance à inhiber celui-ci. Dans l'exemple de *tsunami*, annoté manuellement, un certain nombre de sèmes de son signifié sont repris par des isotopies locales : certaines évoquent le caractère /négatif/ (/provoquer des dégâts/, /danger/), d'autres le /mouvement/ (/effondrement/, /propagation/, /amplification/), d'autres encore l'/ampleur/ du phénomène (/démesure/, /fort impact/, /étendue géographique/). Ces sèmes auront tendance à être activés. À l'inverse, les unités lexicales du cotexte ne sont pas porteuses de sèmes rattachés aux /phénomènes et éléments naturels/, les sèmes correspondants du sémème de *tsunami* seront donc inhibés.

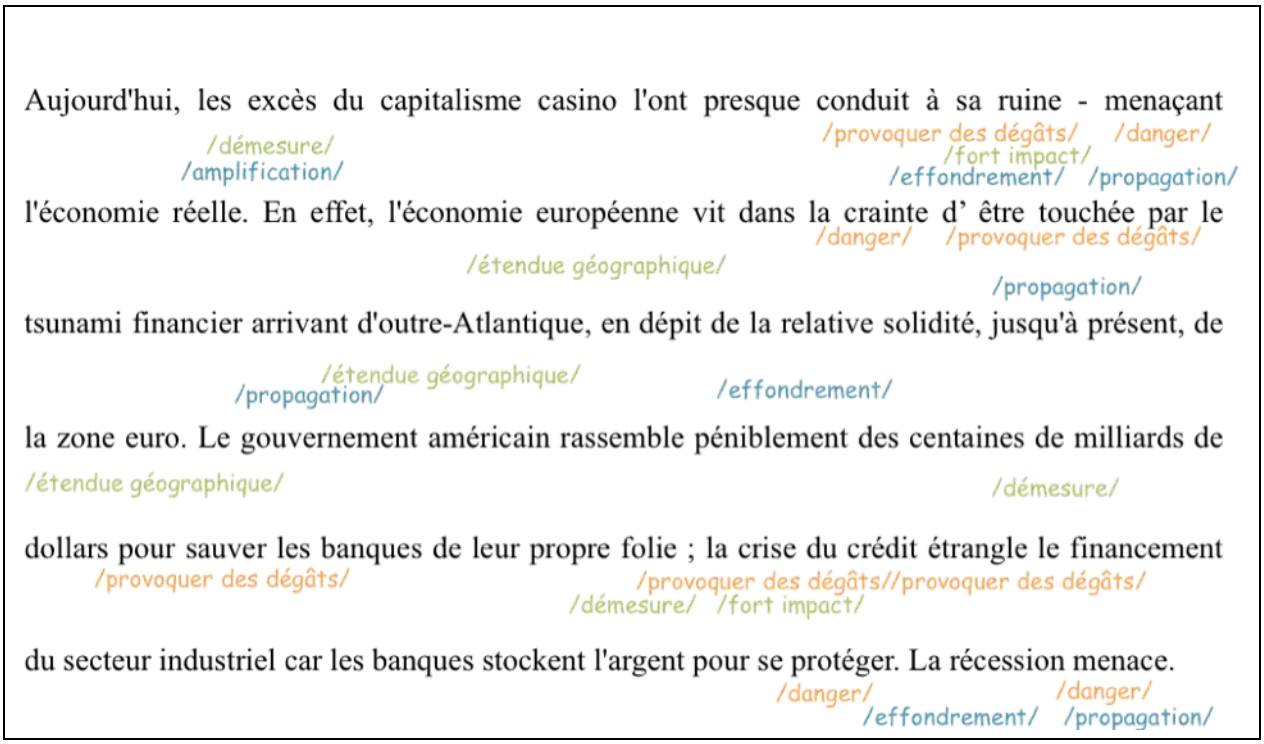


Figure I.3.12 : Isotopies locales au voisinage de tsunami participant à la reconfiguration du sémème

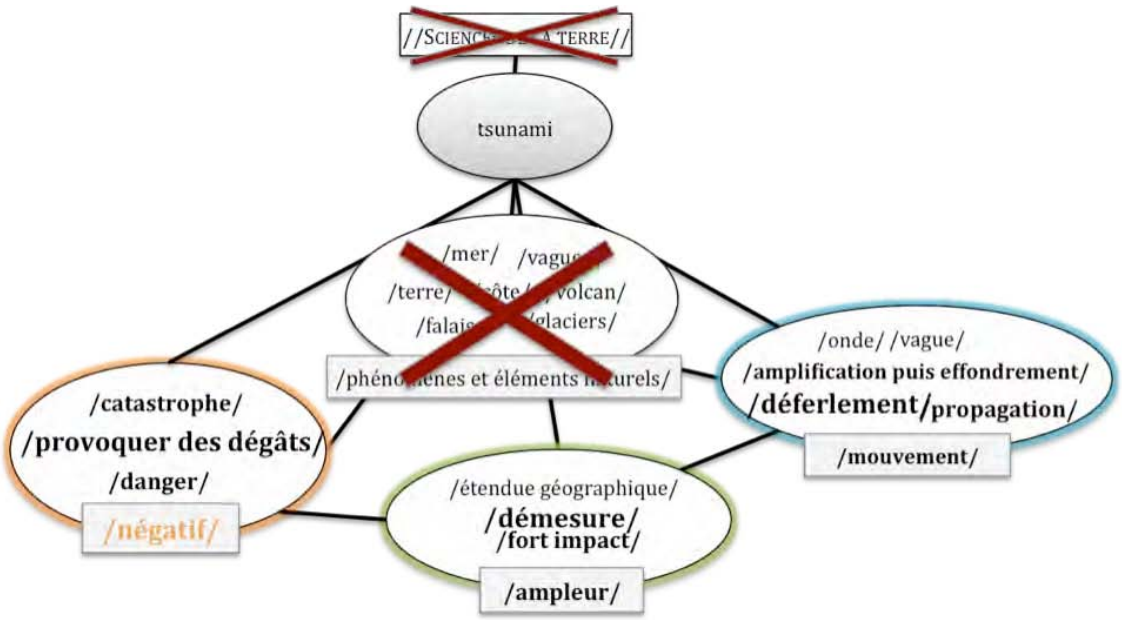


Figure I.3.13 : Reconfiguration du sémème de tsunami induite par les isotopies locales

3.3.2 Propagation de sèmes et enrichissement de signifié

Les sèmes se propagent en contexte. Certains sèmes présents dans le cotexte d'une unité lexicale pourront ainsi se greffer au sens de cette unité. L'intégration de sèmes provenant du cotexte au signifié d'une unité lexicale constitue le phénomène d'afférence : un nouveau sème, contextuel, est adjoint au signifié propre à l'emploi en discours associé – donc à la sémie (Rastier, 1994). Si la même afférence se reproduit de façon récurrente dans les emplois et dans le temps, elle tendra à s'intégrer au sémème.

Un autre phénomène de propagation sémique peut contribuer à l'enrichissement du signifié : le phénomène de *sommation* (Rastier, 2006). Ce phénomène relève de la dynamique des échanges entre fonds sémantiques et formes sémantiques. La *sommation* correspond à une propagation de traits des fonds sémantiques vers les formes sémantiques (Rastier, 2006). Si on adopte l'hypothèse de (Valette, 2010) que la néosémie résulte de la lexicalisation privilégiée d'une forme sémantique, la sommation participe bien de l'enrichissement propre à la néosémie. En effet, elle apporte des traits à la forme sémantique qui s'intégrera au signifié de l'unité lexicale ciblée.

Dans l'exemple ci-dessous, l'isotopie de la //FINANCE// est particulièrement saillante, elle est couplée à d'autres isotopies locales moins saillantes, mais dont certaines entretiennent des relations sémantiques avec elle (on peut donc voir une paratopie qui se crée à partir d'un faisceau d'isotopies locales), comme la récurrence de /banques/, de /marché/ ou d'/instruments monétaires/. Ces unités de sens sont absentes du sémème de *tsunami* et saillantes dans son cotexte, elles auront donc tendance à enrichir le sens de *tsunami.*

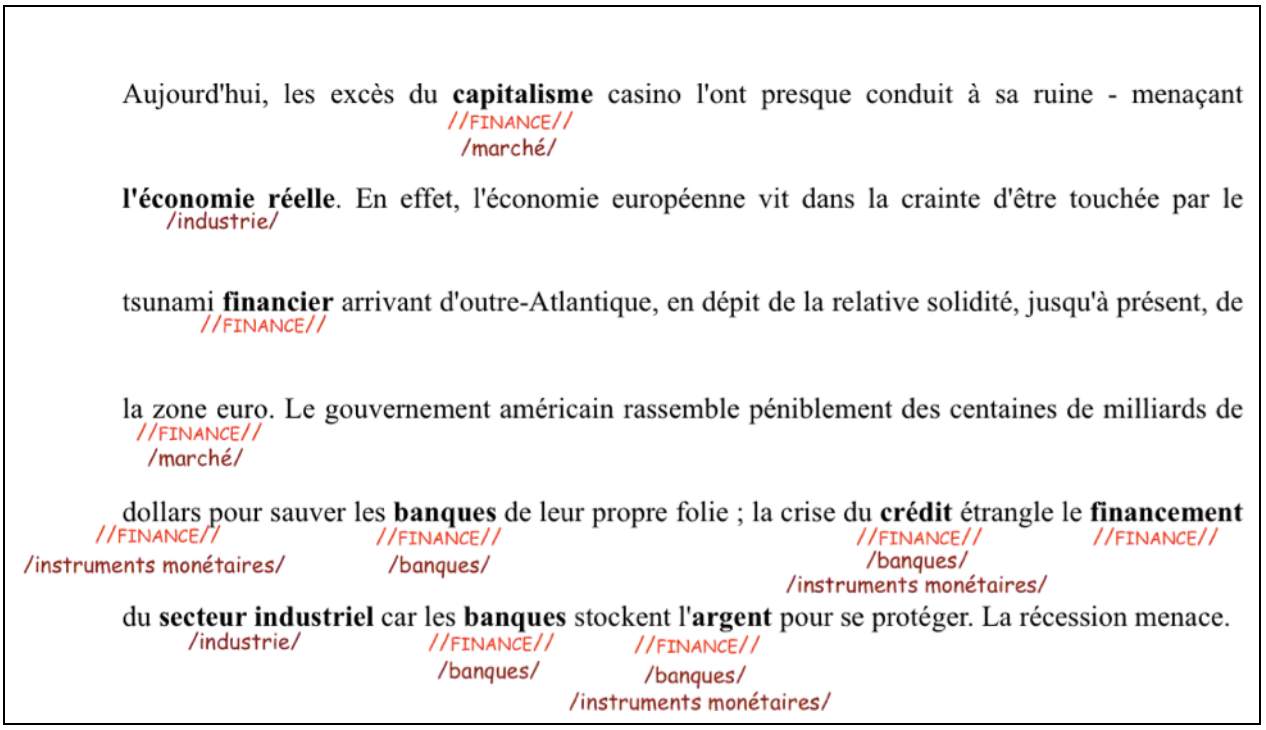


Figure I.3.14 : Isotopies locales au voisinage de **tsunami** *participant à l'enrichissement du sémème*

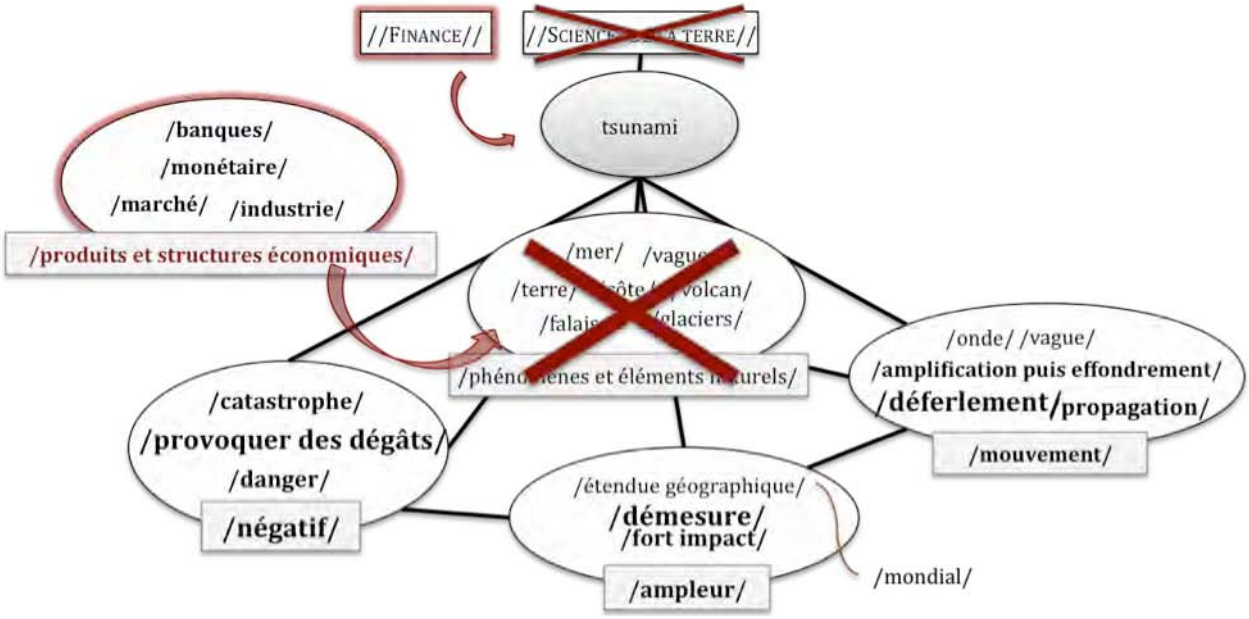


Figure I.3.15 : Enrichissement du sémème de* tsunami *induit par les isotopies locales

Pour que la dynamique ait un impact sur le sens en langue et non seulement en discours, il faut que l'activation, l'inhibition et l'enrichissement ne se limitent pas à un emploi particulier, mais se reproduisent dans le temps de la même façon.

> « 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. (…) tout mot placé dans un texte en reçoit des déterminations sémantiques, et modifie potentiellement le signifié de chacun des mots qui le composent. » (Valette, 2010b:179)

La modification du signifié ne peut évoluer du caractère potentiel au caractère effectif que si le changement en discours se répète suffisamment et de façon régulière.

3.4 Bilan : jongler entre plan lexical et plan sémique pour s'adapter à une perspective applicative

Nous avons dégagé plusieurs pistes pour modéliser l'allocation de signifié. Pour mettre en œuvre ces éléments de modélisation, un traitement automatique doit être axé sur un fil conducteur, c'est-à-dire sur une observable fixée qui permettra de suivre l'évolution de sens. La palette d'indices de néosémie et l'appareil théorique offrent deux choix d'orientation : privilégier un ancrage dans la surface textuelle ou s'en affranchir.

De façon générale, deux angles d'approche sont en concurrence pour définir l'observable privilégiée :

- pister un signifiant : allouer un signifié à un signifiant revient à construire ou reconstruire un signifié pour un signifiant donné ;
- pister le futur signifié : cela revient à choisir comme observable une forme sémantique, c'est-à-dire un groupement stable de sèmes présents en discours et qui n'est pas nécessairement lexicalisé. Allouer un signifié à un signifiant revient à déterminer quel signifiant émerge pour accueillir une forme sémantique donnée.

Le choix du signifiant comme point d'entrée est classique. L'observable est une chaîne de caractères. Cette chaîne de caractères peut éventuellement être modulée par une expression régulière, c'est-à-dire qu'une légère variabilité de signifiant est possible. Lors des actualisations en discours, donc lors des occurrences du signifiant, il s'agira d'étudier la variabilité du signifié préexistant, de répartir les emplois entre les cas de désambiguïsation et les cas de néologie, puis de chercher à dégager des régularités du cotexte des occurrences néologiques, de modéliser la négociation de sens entre le sens cotextuel et le sens préexistant et d'établir un patron d'évolution de ces occurrences.

Le choix d'une forme sémantique comme point d'entrée est plus complexe. Une forme sémantique est un groupement stable de sèmes, dont les occurrences en discours peuvent aboutir à une lexicalisation. (Valette, 2010) voit dans la lexicalisation d'une forme sémantique un processus-clé de la néosémie et il propose le modèle d'évolution suivant :

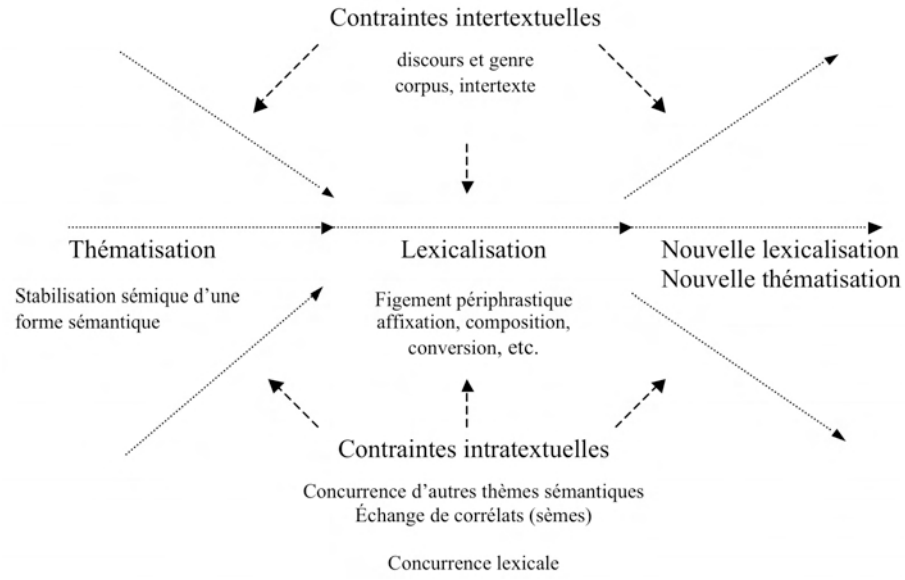


Figure I.3.16 : Lexicalisation d'une forme sémantique (Valette, 2010)

Nous nous intéressons ici à la première partie du schéma de (Valette, 2010), jusqu'à la lexicalisation. Le point de départ est un ensemble stable de sèmes, que la trame textuelle fait ressortir :

> « Avant la lexicalisation, une représentation (un concept) peut en effet exister textuellement de façon plus ou moins ténue, à l'état de thème(s) en cours de structuration. Elle se caractérise par une instabilité sémantique et une certaine complexité textuelle. Elle est enchâssée dans un réseau complexe d'expressions et de phraséologies. » (Valette, 2010:4)

La forme sémantique connaît ensuite un figement, elle prend corps à travers des formes lexicales concurrentes, jusqu'à ce qu'émerge une lexie privilégiée.

Ainsi, lorsque la forme sémantique sert de point d'entrée, la perspective est centrée sur un contenu sémantique doté d'une relative constance, mais qui comporte aussi une certaine modularité et des contours flous. Sur le plan lexical, les signifiants sont soumis à une variabilité qui peut être marquée, qui correspond notamment à la présence de formes lexicales concurrentes. Ceci permet l'intégration du foisonnement néologique et la structuration de cooccurrents, qui sont alors qualifiés et agencés sémantiquement par l'intermédiaire de la forme sémantique. Pister une forme sémantique nécessite de disposer d'une représentation propre au niveau infra-lexical caractéristique du sens de fragments textuels considérés. À l'heure actuelle, les formes sémantiques sont encore à un stade de description théorique, il n'existe pas de modèle robuste adapté à du traitement automatique. Choisir comme point d'entrée une forme sémantique revient à travailler à partir d'une observable indéfinie pour du TAL.

Les deux points de vue sont complémentaires, la façon dont ils se conjuguent pourrait s'appréhender en transformant le schéma précédent de (Valette, 2010) comme suit :

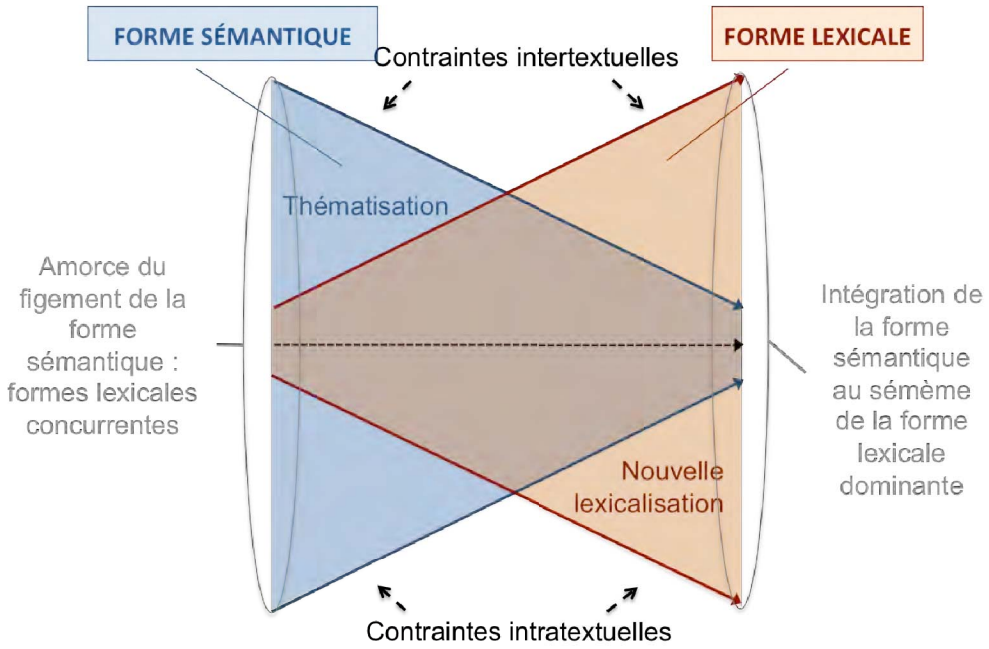


Figure I.3.17 : Complémentarité entre la lexicalisation d'une forme sémantique et l'émergence d'une nouvelle forme lexicale

Les deux approches peuvent s'enrichir mutuellement, elles présentent chacune des difficultés pour un traitement automatique, mais situées à des niveaux différents.

Les développements proposés viseront à coupler les deux approches. Le signifiant et, plus généralement, le niveau lexical auront priorité, du moins dans un premier temps, car ils constituent un point d'entrée incontournable pour des traitements automatiques. Cependant, on cherchera aussi à mettre en place des outils de suivi du signifié et des formes sémantiques.

L'accès aux plans infra-lexical et supra-lexical est doublement nécessaire : d'une part, pour permettre les reconfigurations de signifié (homogénéité oblige : la modulation des sèmes du sémème doit pouvoir se faire à partir d'unités comparables) ; d'autre part, pour représenter l'impact du global (représentation de thèmes, domaines, isotopies, etc., dont on a vu l'importance à la sous-section précédente). Pour décrire les formes sémantiques et autres phénomènes sémiques, on cherchera des regroupements (hypothèse de partage de sèmes, donc isotopie locale potentielle), des réseaux (structure pouvant refléter le squelette d'une forme sémantique) ou encore des saillances sémantiques (isotopies, expression de thèmes ou autres tendances globales) dans le voisinage cooccurrentiel et temporel de la forme lexicale ciblée.

L'enjeu est donc de réussir à accéder au niveau sémique, d'une part pour exprimer du contenu sémantique caractéristique du palier global, d'autre part pour structurer localement l'information sémantique.

Partie II.

Ressources et outils adaptés à l'allocation de signifié

L'allocation de signifié repose sur un jeu de contrastes multiniveaux entre deux espaces structurés, celui des sens codés et celui des sens in vivo, c'est-à-dire en discours. Jusqu'à présent, cette question a été abordée d'un point de vue théorique. Pour passer du cadre théorique au cadre applicatif, il est nécessaire de disposer de ressources, d'outils mathématiques et d'un protocole. Nous proposons dans les chapitres II.1, II.2 et III.1 un modèle applicatif, étayé par une série d'expériences.

Le chapitre II.1 a pour objectif de définir des espaces de représentation du sens : ressource lexicographique pour le sens codé, ressources textuelles pour le sens in vivo.

Le chapitre II.2 propose quelques outils mathématiques adaptés aux jeux de contrastes et utilisés dans notre cadre expérimental.

Au chapitre III.1, nous proposerons un protocole d'allocation de signifié à partir des ressources et outils retenus. On cherchera à mettre en place une démarche structurée, où le jeu sur les différents paramètres cherchera à respecter l'évolution du global vers le local et à mettre en œuvre des affinements successifs de la caractérisation sémantique.

Au chapitre III.2, nous proposerons des développements de ce modèle, aussi bien au niveau des outils et techniques mathématiques utilisés que de l'exploitation du potentiel des ressources.

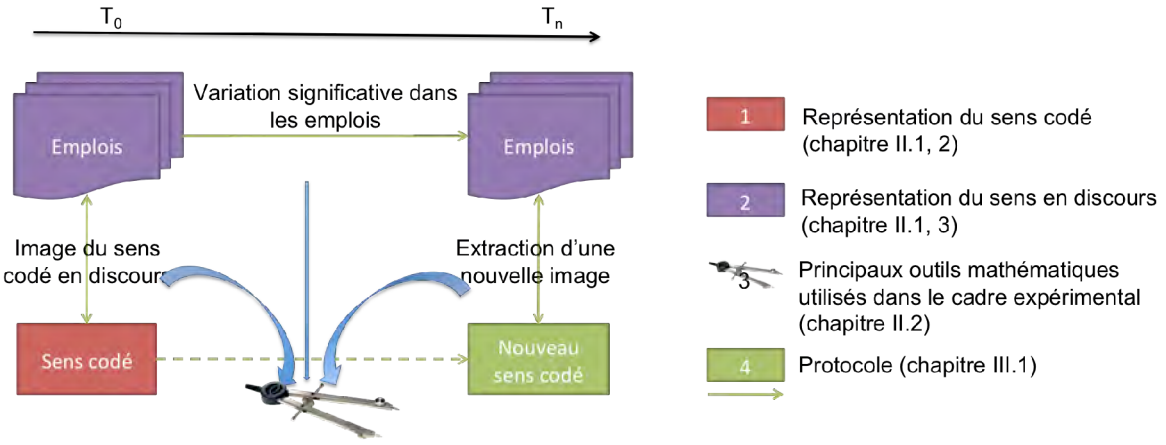


Figure II.a : Grandes étapes proposées pour définir une procédure d'allocation de signifié

3' 302
u 3'.1. 3'.2. 3'.3. 3'.4. 3'.5. 3'.6. 3'.7. 3'.8. 3'.9. 3'.10. 3'.11. 3'.12. 3'.13. 3'.14. 3'.15. 3'.16. 3'.17. 3'.18. 3'.19. 3'.20. 3'.21. 3'.22. 3'.23. 3'.24. 3'.25. 3'.26. 3'.27. 3'.28. 3'.29. 3'.30. 3'.31. 3'.32. 3'.33. 3'.34. 3'.35. 3'.36. 3'.37. 3'.38. 3'.39. 3'.40.

Chapitre II.1
Représentation du sens codé et du sens en discours

La représentation du sens s'est appuyée sur deux types de ressources, lexicographiques et textuelles. Dans un premier temps, nous présenterons la ressource lexicographique que nous avons utilisée pour extraire une image du sens codé, à savoir le *Trésor de la Langue Française informatisé*. Dans un second temps, nous aborderons les caractéristiques recherchées dans des ressources textuelles pour extraire une image du sens en discours, en précisant les corpus utilisés en pratique. Enfin, nous analyserons la question de l'articulation entre les images du sens retournées respectivement par les ressources lexicographiques et textuelles.

1. Représentation du sens codé

La représentation du sens codé est extraite d'un dictionnaire lexicographique, le *Trésor de la Langue Française informatisé* (TLFi ; Dendien et Pierrel, 2003). Le formalisme des entrées est converti, de façon à obtenir un dictionnaire sémique : le contenu rédigé est restructuré sous forme d'un ensemble de traits sémantiques, plus facile à articuler aux ressources textuelles, et adapté à des extractions automatisées de contenu sémantique.

1.1 Le TLFi comme ressource de référence

1.1.1 Une couverture adaptée : description linguistique, vocabulaire étendu et contenu synthétique

a- Un dictionnaire de langue

Le TLFi est la version informatisée du *Trésor de la Langue Française.* Celui-ci est un dictionnaire de langue, non un dictionnaire encyclopédique. Il propose une représentation proprement linguistique du sens des mots, c'est-à-dire qu'il est destiné à fournir à l'utilisateur les éléments nécessaires à l'interprétation, pas à se positionner par rapport aux concepts et à une certaine représentation du monde. Articuler l'information sémantique à une représentation du monde n'est pas une garantie de vérité ou de plus grande justesse, qu'il s'agisse du monde réel, des mondes possibles ou, plus largement, des univers de croyance : un tel positionnement est relatif, il n'exclut pas une certaine subjectivité et il n'échappe pas au problème du flou sémantique (Martin, 1992).

Nous faisons l'hypothèse qu'une approche linguistique retourne des éléments d'information suffisamment précis et pertinents pour faire émerger les caractéristiques d'un nouveau sens. En cela, l'utilisation d'un dictionnaire de langue paraît adaptée.

b- Une portée large

Le TLF a été réalisé sur 30 ans, des années 60 aux années 90, avec les parutions respectives du premier et du dernier tomes en 1971 et 1994. Il s'agit d'une œuvre d'envergure, destinée à devenir le dictionnaire de référence du français :

> « Ses objectifs, définis par Paul Imbs, étaient multiples : bâtir un dictionnaire de référence du français, le doter d'une dimension historique (chaque mot bénéficierait de sa rubrique étymologique) et linguistique (avec une description du mot en contexte). Un corpus d'exemples particulièrement riche viendrait étayer les définitions, et permettrait d'offrir une analyse concrète des usages du mot dans la langue. » (Bernard et Montémont, 2010)

Le TLFi est, de fait, le seul dictionnaire de référence, de langue générale, sur la période des XIX^{e} et XX^{e} siècles et qui résulte d'une approche philologique, par analyse systématique d'un corpus de textes.

Ce dictionnaire dispose d'une couverture large. Doté de près de 100 000 entrées et 270 000 définitions, il constitue une « somme d'informations sans pareille sur la langue française » (Pruvost, 2002:14), obtenue à partir de moyens humains et matériels de grande ampleur.

Les ressources utilisées sont multiples. Trois types de ressources ont servi de support à la rédaction (Henry, 1996) : des dictionnaires, des bases de données textuelles et des études complémentaires sur le lexique, dont on trouvera le détail notamment dans (Fléchon, 1998). Le recours aux ressources textuelles répond à la volonté d'ancrer les définitions dans l'usage. Le corpus de textes était essentiellement constitué de la base textuelle Frantext (www.frantext.fr ; Bernet et Pierrel, 2005), composée de textes littéraires ; à Frantext s'ajoutent des textes scientifiques et techniques, qui représentent 20% du corpus. L'utilisation de concordanciers permettait d'accéder aux contextes d'emploi. Signalons également le recours aux *Archives du Français Contemporain*, un relevé d'emplois de mots nouveaux dans la presse de langue générale (900 000 attestations en 1977). Cette base pouvait être utilisée en l'état, mais elle a aussi amené les lexicographes à consulter les articles journalistiques à l'origine des fiches d'attestations.

Au niveau des destinataires du TLFi, Imbs qualifie le public ciblé d' « homme cultivé moderne », s'intéressant à l'homme en général, dans ses préoccupations quotidiennes, dans son organisation en société et ouvert aux sciences et techniques. Ce public légitime la sélection d'un vocabulaire étendu, qui reste général mais qui intègre également du vocabulaire plus spécialisé. Au niveau du produit final, l'ouverture scientifique et technique se traduit par des définitions dépendant de domaines spécialisés, établies à partir d'un *Thesaurus des domaines scientifiques et techniques*. Comme nous le verrons, cette information a eu un rôle déterminant dans nos expériences pour jouer sur la granularité sémique et accéder à de l'information supra-lexicale. Par ailleurs, Imbs assimile l'homme cultivé moderne à « ce que naguère on nommait les élites » : dans les définitions, cet aspect se reflète à travers le choix d'un certain niveau de langue et d'un vocabulaire diversifié. Les rédacteurs n'ont pas privilégié un vocabulaire simple, fortement récurrent d'une définition à l'autre ; le principe de non-circularité a aussi limité la redondance du vocabulaire d'une définition à l'autre. Selon nous, dans le protocole expérimental que nous présenterons au chapitre III.1, cette tendance a brouillé un phénomène que nous cherchions à identifier, la récurrence de sèmes, tributaire du vocabulaire utilisé dans les définitions. D'un point de vue plus technique, cela a amplifié la dispersion des données et le phénomène des matrices creuses. Ce point a probablement joué sur la qualité des résultats expérimentaux.

c- Un contenu synthétique

Le TLF recouvre un vocabulaire large, mais les entrées n'ont pas été rédigées dans un but d'exhaustivité :

> « Il s'agit moins d'un livre qui aurait à enseigner une matière entièrement nouvelle que d'un aide-mémoire au sens étymologique du mot, c'est-à-dire d'un livre rappelant et précisant ce qu'on sait déjà ou ce qu'on peut induire ou déduire de ce qu'on sait. » (Imbs, 1971)

Le choix d'un contenu synthétique s'explique par le public destinataire, caractérisé par une culture humaniste, pour lequel des guides d'interprétation sont jugés suffisants et sont préférés à des explications développées. Ce choix s'explique aussi par le type de dictionnaire, à savoir un dictionnaire de langue, non pas un dictionnaire encyclopédique. Signalons que la rédaction synthétique n'est pas seulement dictée par ces arguments théoriques : la taille des définitions lexicographiques a aussi été conditionnée par des contraintes éditoriales (Henry, 1996:14). La version informatisée du TLF offre, certes, un espace a priori illimité, mais le TLF a été avant tout conçu comme une ressource au format papier, limitée en place : l'informatisation n'a été décidée qu'au moment où la rédaction du TLF touchait à sa fin.

Le caractère synthétique des définitions est un principe en accord avec le positionnement que nous avons choisi pour l'allocation de signifié (*cf.* chapitre I.3, 2.3.3), à savoir retourner une représentation compacte de l'information ainsi que des éléments pertinents, plutôt que d'orienter vers une masse de contenu sémantique où l'interprétant doit faire sa sélection.

1.1.2 Une ressource favorable à un cadre applicatif

a- Un contenu informatisé

Le TLFi est une ressource informatisée qu'on peut considérer comme auxiliaire de traitements automatiques (Martin, 2001) :

> « La thèse ici défendue (…) est que le dictionnaire automatisé peut accroître sensiblement les performances d'un analyseur sémantique. » (Martin, 2001:51)

Il répond à l'exigence de disposer d'une ressource lexicale de qualité et accessible de façon pérenne. Le format informatique offre l'opportunité d'intégrer les données lexicales à des traitements automatiques, notamment pour l'étude de données textuelles.

Le TLFi est doté d'une structure adaptable à des traitements automatiques. Il dispose d'un balisage XML conséquent, qui identifie notamment les différents objets des entrées lexicographiques (définitions, domaines, exemples, code grammatical, etc.)[32]. Ce balisage permet d'accéder à des dépendances, par exemple celles des définitions à des domaines.

b- La base d'un chantier dynamique de nouvelles ressources

Le TLFi peut se concevoir comme composante centrale d'un réseau de ressources, actuellement en cours de construction. Le TLFi fait ou a fait l'objet d'un ensemble de projets scientifiques destinés à générer des ressources dérivées.

- ***La base SEMEME*** (réalisée par Étienne Petitjean, projet interne de l'ATILF) : cette base est dérivée du TLFi. Elle réalise la décomposition des définitions en sèmes, en jouant sur une relation d'inclusion, et sur laquelle on revient en 2.2.2. Cette base de données a été

[32] « Quelques chiffres peuvent donner un aperçu de la finesse de ce balisage : après validation sur l'ensemble des seize tomes, 36 613 712 balises XML ont été positionnées : 17 364 854 balises typographiques, 1 070 224 balises décrivant la hiérarchie, 18 178 634 balises repérant les objets textuels, dont 92 997 entrées et 64 346 locutions faisant l'objet de 271 166 définitions et illustrées par 427 493 exemples » (Pierrel, 2007)

réalisée dans le cadre du projet DIXEM (Valette *et al.*, 2006), projet interne au laboratoire ATILF visant à l'élaboration d'un dictionnaire sémique. Elle comporte près de 100 000 fichiers XML, où chaque fichier correspond à une entrée du TLFi. Pour une présentation de SEMEME, on pourra notamment se reporter à (Gheorghita, 2011).

- ***Le projet Definiens*** (Barque *et al.*, 2010) : ce projet a pour objectif la construction d'une ressource dérivée du TLFi, avec des définitions de type paraphrase, structurées en genre prochain et différences spécifiques. Techniquement, la structuration s'effectue à travers un balisage XML des définitions du TLFi. Ce projet est actuellement en cours de réalisation. Notons qu'il se situe dans l'esprit des ambitions de Paul Imbs (décrire le sens de manière analytique, en genre prochain et différences spécifiques), dont la concrétisation s'était heurtée à des difficultés de terrain lors de la rédaction et dont la transcription informatique avait été écartée, car jugée trop complexe.
- ***Le projet RLF*** (Lux-Pogodalla et Polguère, 2011) : ce projet lexicographique d'envergure a pour ambition de construire un dictionnaire comme système lexical représenté sous forme d'un réseau, ou graphe, de grande dimension. Le modèle lexical sous-jacent repose sur la lexicologie explicative et combinatoire (Mel'čuk *et al.*, 1995). En particulier, les fonctions lexicales de ce modèle sont une base de définition de la structure de graphe. Le TLFi, et de façon intermédiaire la base de données issue de Definiens, constituent les sources principales d'information lexicographique.

Les projets en cours participent à l'enrichissement du contenu du TLFi. La dynamique d'amélioration de la ressource s'autoalimente : le développement de modules permet à d'autres modules de se perfectionner, et la génération de ressources dérivées peut améliorer le potentiel applicatif de la ressource initiale[33]. Le présent travail participe de cette dynamique.

1.1.3 Le paramètre temps : une nécessaire actualisation et une hétérogénéité interne

Les dates de rédaction du TLF, du début des années 60 au milieu des années 90, sont doublement problématiques.

D'abord, le dernier tome du TLFi est paru il y a presque vingt ans. L'information présente dans le TLFi n'est pas à jour. Elle nécessite une actualisation, à laquelle doit participer le Supplément au TLFi, dont la réalisation a été amorcée en 1992, mais qui, à l'heure actuelle, n'a toujours pas été publié. Cependant, le Supplément ne remédiera que partiellement au décalage, et le délai entre l'amorce et la fin de son élaboration ne témoigne pas en faveur d'une grande réactivité par rapport à l'évolution de la langue. La mise à jour nécessite, d'une part, l'introduction de nouvelles entrées ; par exemple, dans la classe sémantique des régimes alimentaires, *hyperprotéiné, édulcorant* ou encore *aspartame* n'existent pas. D'autre part, il est nécessaire de réviser des entrées existantes. Ainsi, dans le domaine du bien-être, *zen* n'est pas défini comme synonyme de serein, détendu, mais il est rattaché exclusivement à du philosophique ou du religieux ; de même, *diffuseur* apparaît dans des emplois rattachés à du visuel (*diffuseur de lumière*) ou de l'auditif (*diffuseur radio),* ou encore dans un sens technologique, fortement marqué par une dimension mécanique, tandis que n'apparaît pas le diffuseur qui génère des senteurs (*diffuseur d'huiles essentielles* par exemple), produit de bien-être, propre à un usage personnel ou ménager. La nécessité d'une actualisation se fait particulièrement sentir dans des domaines en fort développement dans la société dans les années 90 et 2000, comme l'informatique ou la télématique. La mise en place d'outils pour accélérer l'actualisation est donc particulièrement pertinente pour le TLFi. Cette faiblesse de la ressource est un atout dans notre cadre : elle renforce les possibilités d'observer des

[33] Les produits dérivés sont parfois plus à même de fournir un contenu pertinent et adapté à un objet de recherche : « Après transformation, les dictionnaires les plus savants ne sont pas nécessairement les plus utiles comme matière première pour des expériences » (Habert, 2000), cité par (Loiseau *et al.*, 2010).

variations sémantiques par rapport au sens codé pour des corpus des années 90 ou 2000. Cependant, cet atout n'est valide que jusqu'à un certain stade : dans les domaines marqués par une évolution importante (informatique, etc.), un trop grand nombre de lexies propres à ces domaines est affecté par un décalage sémantique ou par une absence de la ressource. Pour expliquer le sens d'une unité lexicale à partir du sens d'autres unités lexicales, il faut que le sens de ces autres unités soit correctement défini, donc qu'il soit à jour.

Le deuxième problème du TLFi est lié non à la date de fin, mais à la durée du projet. L'étalement sur 30 ans s'est accompagné d'une évolution des normes et de variations du contenu des définitions. Ajoutons que la rédaction a été effectuée par ordre alphabétique et non pas thématique, pour des raisons éditoriales. L'hétérogénéité ne peut donc être couplée à une structure sémantique, elle est susceptible de se manifester au sein de tout regroupement thématiquement cohérent, pour peu que ce regroupement comporte une certaine dispersion alphabétique. En particulier, la caractérisation domaniale des définitions a été affectée par des évolutions temporelles. Avec le changement de direction, Bernard Quemada remplaçant Paul Imbs, les rédacteurs ont eu pour consigne de systématiser l'introduction d'étiquettes de domaines, alors qu'auparavant, les domaines considérés comme évidents pour certaines définitions n'étaient pas explicités[34]. Rappelons que les étiquettes de domaines jouent un rôle important dans nos expérimentations. L'existence d'un certain nombre de changements est connue, et leur nature est déductible des archives de la rédaction du TLF. Cependant, on ne dispose pas de mesure des évolutions sur la structure d'ensemble et on ne sait évaluer les biais qu'elles peuvent éventuellement introduire.

1.2 Du dictionnaire lexicographique au dictionnaire sémique

1.2.1 Fondements théoriques : des entrées aux sémèmes

Un principe fondateur de la rédaction est l'analyse sémique, comme l'affirme Imbs dans la Préface du TLF :

> « … l'information sémantique principale consiste dans la définition, qui est la forme lexicographique traditionnelle de l'analyse componentielle. La définition consiste en effet à rendre compte, sous la forme d'un énoncé analytique, des sèmes pertinents qui entrent dans la composition d'un sens. » (Imbs, 1971:XXXV)

La description en sèmes devait notamment structurer les définitions en fonction de sèmes au statut variable (sèmes génériques et sèmes spécifiques comme pendants de la structure des définitions en genre prochain et différences spécifiques) (Imbs, 1971:XXXV). Intégrer les sèmes comme bases de description est, certes, clairement affirmé, mais en pratique, cette volonté est loin d'avoir été systématiquement respectée (Radermacher, 2004:236).

Le potentiel des définitions comme réservoirs de sèmes est une hypothèse qui respecte la philosophie de rédaction et à laquelle nous adhérons. Cette hypothèse est défendue par (Valette, 2008), qui propose une méthode concrète pour extraire le contenu sémique des définitions du TLFi :

> « Pour la réalisation d'un[dictionnaire de traits sémantiques infralexicaux], nous avons exploité un dictionnaire de langue informatisé : le *Trésor de la Langue Française* (Dendien & Pierrel 2003), désormais *TLF*. (…). [La démarche reposait sur] une hypothèse minimaliste : une définition est une sémie mise en texte. Ainsi, les mots pleins d'une définition (substantifs, adjectifs, verbes et certains adverbes) sont, une fois lemmatisés, les sèmes potentiels qui constituent la sémie d'une unité lexicale en attente d'actualisation. »

[34] Communication personnelle de Pascale Bernard, lexicographe ayant participé à la rédaction du TLF.

Selon ce principe, une entrée est convertie en un sémème qui se présente sous forme de sacs de traits sémantiques. Les traits sémantiques proviennent de la lemmatisation des mots sémantiquement pleins des définitions.

Le format proposé pour la représentation du sémème est, certes, réducteur par rapport à la richesse des entrées et à leur potentiel sémique maximal. On perd notamment :

- La structure des entrées en acceptions. L'ordre des définitions, les divisions retenues et le niveau de chaque subdivision répondent à de grands principes d'organisation de l'information, bien que non systématiques (Henry, 1996:27-28). La méthode d'extraction de sémème n'intègre pas cette structure déterminante dans l'organisation du contenu sémantique.
- Les informations en termes de sèmes génériques et de sèmes spécifiques. Ces informations n'ont pu être récupérées car elles sont absentes de l'encodage du TLFi : elles ont certes guidé la rédaction sur le plan théorique, mais leur identification a posteriori, notamment lors de la transcription du TLF au format informatique, constituait une tâche trop complexe (Henry, 1990)[35]. Remarquons que des travaux sont actuellement en cours pour introduire des balisages des définitions en genre prochain et différences spécifiques, à travers le projet Definiens (Barque *et al*., 2010 ; Barque et Polguère, 2009).

Bien qu'il n'exploite pas toute la richesse de la ressource, le mode d'extraction sémique préconisé par (Valette, 2008) permet de rendre opérationnel le passage de la ressource lexicographique à une représentation sémique. De fait, les principes défendus dans ce cadre ont permis de mettre en place un projet interne au laboratoire, fondateur d'une approche textuelle du lexique. Sur la base de ces principes, un outil d'extraction sémique a pu être développé.

1.2.2 Une représentation outillée par une plateforme d'extraction sémique

(Grzesitchak, 2008)[36] a développé une plateforme, Semy, qui extrait les représentations sémiques du TLFi et les projette dans des ressources textuelles.

Le programme de Semy est articulé à la base de données SEMEME, version dérivée du TLFi. Cette base représente chaque entrée lexicographique dans un fichier XML, contenant les définitions, les domaines associés le cas échéant et la représentation convertie des définitions comme ensemble de traits sémantiques.

La plateforme Semy interface la base SEMEME avec des ressources textuelles. Au niveau des fonctionnalités, elle peut procéder à de l'annotation de corpus et générer des indicateurs statistiques sur les distributions sémique et lexicale en corpus. En entrée, Semy dispose de textes ou de corpus de textes. Elle conserve deux niveaux de structure : les fichiers (textes constitutifs du corpus) et les lignes (dans nos applications, les unités textuelles, à savoir les paragraphes constitutifs des textes). Elle procède à la lemmatisation des formes lexicales à l'aide de l'étiqueteur Treetagger (Schmid, 1995), à l'élimination de la ponctuation et des mots-outils, puis elle affecte à tout lemme un ensemble de traits sémantiquement pleins issus de la définition (en l'occurrence, les noms, verbes, adjectifs et adverbes). Au niveau du corpus, tout lemme se voit substituer l'ensemble de traits sémantiques associés, pas à pas. On obtient en

[35] « … il s'agit d'analyser soigneusement le contenu informatif du dictionnaire papier dans le but non seulement e repérer les données constitutives de ce dictionnaire, mais aussi d'en identifier la nature (…) – définition : le principe d'une décomposition est parfaitement défendable et même justifié, mais poserait de très sérieux problèmes…» (Henry, 1990:203)

[36] La plateforme Semy est en développement et n'est pas encore en libre accès. Pour d'autres présentations de Semy et de ses applications, on pourra se référer à (Grzesitchak *et al.*, 2007) et (Baider, 2007).

sortie une image sémique du corpus, avec un découpage en fichiers et lignes analogue au découpage initial. Chaque ligne est constituée d'un ensemble de traits sémantiques.

L'annotation réalisée par Semy est paramétrable. Elle peut être réduite ou élargie à d'autres catégories grammaticales que celles mentionnées précédemment. Il est également possible de restreindre l'annotation aux étiquettes de domaines. Ajoutons que Semy identifie certains traits sémantiques comme des métasèmes, tels que /avoir/, /chose/ ou /être/, sources de bruit dans les expériences et difficiles à interpréter. Il est alors possible d'exclure ces métasèmes de l'annotation. La plateforme peut également identifier des collocations et des locutions dans les textes, lorsque ceux-ci possèdent des équivalents dans le dictionnaire encodés comme syntitas[37] ; les syntitas peuvent alors définir des traits sémantiques.

1.3 Alternatives à la représentation du sens codé

Notre cadre de travail et, de ce fait, le choix de la ressource ont été définis à partir d'un certain nombre de contraintes, notamment le choix d'un dictionnaire de langue, la langue française comme langue de référence, l'accès libre à la ressource, un contenu à couverture large et une exploitabilité informatique (diffusion sur support électronique d'une version complète et non en cours de réalisation). À notre connaissance, aucune autre ressource existante ne vérifie l'ensemble de ces contraintes. Cependant, en relâchant les contraintes, d'autres ressources pourraient servir de support. Certaines des ressources alternatives ont des atouts qui font défaut au TLFi. Nous n'analyserons que quelques ressources, en nous focalisant seulement sur des aspects particuliers qui mériteraient de faire partie d'une ressource idéale.

1.3.1 Des ressources plus à jour

D'autres ressources en lexicographie française présentent un profil de même nature que le TLFi et auraient pu servir de supports de représentation du sens codé. Les dictionnaires du Larousse, du Robert et de Hachette constituent les grands concurrents nationaux. Tout comme pour le TLFi, des versions électroniques existent (pour une liste complète, *cf.* Gasiglia, 2009:257-261).

Le contenu de ces dictionnaires est plus à jour que celui du TLFi. En particulier, le *Nouveau Petit Robert* (NPR) et le *Petit Larousse* (PL) proposent tous les ans des amendements, avec l'introduction de nouveaux mots et de nouveaux sens, ainsi que des suppressions de mots. Cependant, ces mises à jour relèvent d'un "artisanat" (Rey, 2008) et comportent une part de subjectivité, sans garantie d'exhaustivité. La part subjective dans les modifications est sensible tant au niveau de divergences entre le *Nouveau Petit Robert* et le *Petit Larousse* qu'à travers certaines pratiques telles que les suppressions puis réintroductions à un petit nombre d'années d'intervalle. Pour une analyse détaillée des évolutions et des différences entre les deux dictionnaires, on pourra se reporter à (Martinez, 2009).

Ces dictionnaires ne sont pas institutionnels mais à vocation commerciale. De ce fait, ils ne sont pas en libre accès et restent soumis à des droits.

1.3.2 Plus petit vocabulaire pour les définitions, moins d'éparpillement

Un dictionnaire anglais pour apprenants, le *Longman Dictionary of Contemporary English* (LDOCE), utilise une liste restreinte de mots dans les définitions lexicographiques :

[37] Les syntitas sont les syntagmes du TLFi. Ils apparaissent en italique. Certains syntitas possèdent une définition, d'autres non. Lorsque les syntitas correspondent à des syntagmes définis, des traits sémantiques peuvent leur être affectés.

> « On sait que la bonne définition doit s'efforcer d'utiliser des mots plus "simples" que le mot à définir (…). Les auteurs du LDOCE sont les premiers lexicographes à fournir, dans le corps du dictionnaire, une liste de 2000 mots "simples" qui serviront à définir tous les autres » (Béjoint, 2009:136)

Dans notre cadre, cette approche revient à réduire la diversité du vocabulaire sémique. Ceci permettrait de réduire un problème résultant d'un double choix : le choix des rédacteurs du TLFi d'adopter un vocabulaire varié et celui que nous avons fait de considérer comme sèmes les noms, verbes, adjectifs et adverbes sous forme lemmatisée sans effectuer de regroupement. De ce fait, le vocabulaire sémique est diversifié et faiblement récurrent d'une définition à l'autre, ce qui va à l'encontre de la recherche d'isotopies et amplifie l'éparpillement des données. À l'inverse, un vocabulaire simple favorise la récurrence et réduit l'éparpillement.

Le dictionnaire mentionné appartient à la lexicographie anglaise. Il n'existe pas d'équivalent français à notre connaissance, c'est-à-dire de dictionnaire complet dont les définitions sont construites à partir d'un vocabulaire réduit.

1.3.3 Une microstructure en unités pleines et relations sémantiques

Au niveau de la microstructure, un autre format de représentation permettrait d'articuler plus facilement les définitions au formalisme utilisé pour analyser les ressources textuelles. En effet, il est extrêmement difficile de comparer directement les définitions rédigées aux usages discursifs. Pour récupérer le contenu sémantique d'une définition, il est nécessaire de retravailler au préalable sa structure, notamment pour récupérer des unités porteuses de contenu sémantique. Dans notre approche, cette reconfiguration des définitions se présente comme une déstructuration sous forme de sac de sèmes.

Le *MacMillan English Dictionary for Advanced Learners* (MEDAL), construit à partir du Sketch Engine (Kilgarriff et al., 2004), propose un formalisme complètement différent de celui habituellement présent dans les dictionnaires, mais qui est beaucoup plus proche des représentations obtenues à l'issue de traitements textuels en linguistique de corpus. Les entrées se présentent comme des "portraits de mots" (Béjoint, 2009:142-144). Un "portrait de mots" est un ensemble d'unités lexicales sémantiquement associées au mot-vedette. Cet ensemble est subdivisé en sous-ensembles, chacun étant relié au mot-vedette par une relation. À chaque élément est affectée une pondération. Cette représentation du sens repose sur le même principe que celui qui intervient dans la constitution du sémème, mais en outre, le type de relation est précisé et les sèmes ne sont pas tous à poids égal.

Le DiCo (Mel'čuk et Polguère, 2006) propose également un formalisme qui se distingue des définitions rédigées, avec des unités sémantiquement pleines et des relations sémantiques explicitées par les fonctions lexicales de la théorie Sens-Texte. Moins littéraire et plus mathématique que les définitions habituelles, ce formalisme ouvre sur une autre représentation du sens, susceptible de se transposer avec moins de perte et plus de richesse à des données textuelles, moyennant une approche en graphe plutôt qu'en sac de mots.

1.3.4 Des définitions hiérarchisées en fonction de l'usage

Les définitions du *TLFi* sont hiérarchisées, mais la hiérarchie n'est pas systématique, et elle n'est pas toujours fonction de l'importance relative des sens. Le *New Oxford Dictionary of English* (NODE) est organisé selon l'hypothèse que les unités polysémiques ont un ou des sens principaux dont les sens secondaires sont des dérivations propres à des contextes particuliers (Béjoint, 200:145). Dans les entrées, la hiérarchie des définitions s'efforce de respecter cette hiérarchie des sens. En pratique, la qualité des hiérarchies du NODE reste à vérifier, mais l'idée d'organiser systématiquement les définitions d'une entrée en fonction de leur importance dans l'usage comme potentiel de sens principal ou secondaire est à conserver.

Une telle hiérarchie permettrait d'affecter des poids aux différentes définitions, et ainsi d'avoir une représentation sémique modulée.

2.3.5 Une macrostructure intégrant réseaux et regroupements

Au niveau de la macrostructure, certaines ressources proposent des regroupements thématiques d'entrées ou elles ouvrent sur une organisation en réseaux des entrées. Ainsi, deux dictionnaires anglais destinés à des allophones, c'est-à-dire à des personnes pour lesquelles l'anglais est une langue étrangère, proposeraient des regroupements thématiques qualifiés d'entrées :

> « un traitement onomasiologique donn[e] accès à une pluralité de modes d'expression d'une même idée ou notion (…). Les traitements onomasiologiques (…) ont été adoptés pour le *Longman Language Activator* et le *Longman Essential Activator*, qui permettent de confronter commodément les valeurs des mots et expressions réunis au sein d'une même rubrique thématique » (Gasiglia, 2009:283)

Par ailleurs, l'amorce de réseaux lexicaux est présente à travers des liens hypertextuels dans *Robert électronique* (Brunet, 1996:144-150), bien que la construction du réseau ne soit pas aboutie, seulement amorcée. Les réseaux et regroupements sont aussi indirectement accessibles à partir du TLFi par l'intermédiaire du moteur de recherche, mais il s'agit d'un stade d'accessibilité inférieur d'un cran aux ressources évoquées.

1.3.6 Description précise et accessibilité des lexies complexes

Les dictionnaires de collocations, tels que le DiCo (Mel'Xuk et Polguère, 2006), permettent une extension de la nomenclature. Ils offrent la possibilité d'accorder un statut d'importance équivalente aux lexies simples et aux lexies complexes. Ces dictionnaires permettent un accès aux unités lexicales complexes et ils donnent une description précise de celles-ci. Au niveau du TLFi, des lexies complexes apparaissent dans les syntagmes définis, mais ceux-ci ne sont pas au même niveau que les autres unités. Leur accès est possible via le moteur de recherche complexe, mais il n'est pas aussi immédiat que dans les dictionnaires spécifiquement conçus pour les collocations. De plus, les syntagmes définis du TLFi ne sont pas nécessairement des collocations, ils n'ont pas été choisis en fonction d'un critère de compositionnalité ou de figement, mais ils correspondent à des unités associées à la vedette et que les lexicographes ont jugé pertinent de définir.

1.3.7 Bilan : ressource idéale et réalité du terrain

En combinant les caractéristiques d'autres ressources évoquées ci-dessus à celles du TLFi, on pourrait tendre vers une ressource plus performante. Le tableau ci-dessous propose un récapitulatif.

Caractéristique recherchée (✓)	Problème corrélé (la caractéristique recherchée n'est pas vérifiée) (×)	Dictionnaires[38]							
		TLFi	NPR	PL	LDOCE	DiCo	MEDAL	NODE	LLA ou LEA
Langue française	Autre langue	✓			×		×	×	×
Couverture large	Couverture réduite	✓				×			×
Ressource libre	Ressource commerciale	✓	×	×					
Mise à jour régulière (annuelle)	La rédaction du TLFi date de 20 à 50 ans	×	✓	✓					
Vocabulaire simple dans les définitions	Éparpillement (sparseness)	×			✓				
Formalisme en unités sémantiquement pleines et liens sémantiques entre unités	Microstructure difficile à projeter dans les textes	×				✓	✓		
Hiérarchie des définitions en fonction de l'importance de l'usage	Poids équivalent pour chaque définition	×						✓	
Réseau de définitions et regroupements thématiques	Macrostructure non exploitée	×							✓
Description précise et accessibilité des collocations	Représentation des lexies complexes et accessibilité de celles-ci	×				✓			

Tableau II.1.1 : Bilan des caractéristiques d'une ressource lexicographique plus performante

Les caractéristiques d'autres ressources lexicographiques contiennent donc des voies d'amélioration du TFLi. De plus, un certain nombre de travaux ouvrent des pistes prometteuses et permettraient d'améliorer la représentation du sens codé. Cependant, ces travaux ne sont pas achevés, du moins en lexicographie française : certains sont en cours de réalisation, tels que la version balisée du TLFi en genre prochain et différences spécifiques (projet Definiens ; Barque *et al.,* 2010) ou l'encodage des définitions du TLFi sous forme de fonctions lexicales, telles qu'elles apparaissent dans le DiCo (Mel'Xuk et Polguère, 2006) ou sont destinées à apparaître (Lux-Pogodalla et Polguère, 2011). D'autres n'ont pas fait l'objet de validations suffisamment convaincantes par la communauté scientifique, comme les travaux de (Gaume, 2006), dont les faiblesses sont pointées par (Loiseau *et al.,* 2010). D'autres encore présentent des faiblesses avérées, comme les regroupements morphologiques de (Ramdani, 2007) qui comportent notamment des regroupements de taille importante au contenu hétérogène (*cf.* (Reutenauer, 2009:20-21)) concernant certains effets indésirables).

Nous faisons l'hypothèse que si les applications développées fonctionnent avec la ressource imparfaite dont on dispose actuellement, elles seront plus performantes sur une ressource de plus grande qualité.

[38] Abréviations – TLFi : Trésor de la Langue Française informatisé ; NPR : Nouveau Petit Robert ; PL : Petit Larousse ; LDOCE : Longman Dictionary of Contemporary English ; MEDAL : MacMillan English Dictionary for Advanced Learners ; NODE : New Oxford Dictionary of English ; LLA : Longman Language Activator, LLE : Longman Essential Activator. Pour les dictionnaires autres que le TLFi, nous nous sommes limités à un ou deux points forts absents du TLFi, jugés importants et validés par la ressource considérée, et à une caractéristique nécessaire non vérifiée par la ressource considérée, mais validée par le TLFi.

2. Représentation du sens en discours : corpus textuels

2.1 Un corpus adapté à une approche diachronique

Le choix du corpus dépend doublement du facteur temps : en termes de durée ou d'étalement dans le temps, indissociables du phénomène étudié ; en termes de période ciblée, relative à la ressource de référence choisie, à savoir, dans notre cadre, le TLFi.

Sur le plan de la durée et de l'étalement dans le temps, le corpus peut raisonnablement s'étendre de quelques mois à environ une ou deux dizaines d'années. La durée de quelques mois semble un minimum pour disposer d'une quantité de données suffisante, relativement au nombre d'occurrences de la cible lexicale, et pour limiter les effets dus à des pics événementiels. Cette durée de quelques mois permettrait d'observer des changements sémantiques naissants, tels que le syntagme *tsunami nucléaire,* apparu avec la catastrophe nucléaire de Fukushima déclenchée par un séisme suivi d'un tsunami en mars 2011, et dont le devenir lexical n'est pas prédictible quatre mois après. Un étalement sur une à deux décennies reste adapté à une perspective de diachronie courte, il permet d'avoir un certain recul et d'observer un déroulement relativement complet de l'émergence d'un nouveau sens, comme dans l'analyse de (Viprey, 2010) : l'évolution de sens de *mutualisation* s'étale sur une quinzaine d'années et cette durée permet d'observer la diffusion du nouveau sens, une période de basculement dans cette diffusion et d'avoir un aperçu d'un début de stabilisation du nouveau sens. L'espacement d'une à deux décennies est d'un moindre intérêt s'il correspond à deux périodes disjointes plutôt qu'à un continuum temporel : la phase transitoire et les informations qui lui sont propres (en particulier, à travers le foisonnement néologique) risquent d'échapper à l'analyse. Ainsi, une analyse de *tsunami* en 2000 et en 2007 témoignera d'un accroissement des emplois, mais on perdra le pic événementiel de 2004-2005, qui a déclenché la diffusion de l'emploi métaphorique. La finalité du présent travail invite, à terme, à privilégier un étalement réduit (de quelques mois à un petit nombre d'années) : l'approche automatisée peut certes contribuer à fournir un observatoire, mais elle a aussi pour vocation d'aider à une détection et caractérisation précoces de l'émergence de nouveaux sens, et de favoriser une plus grande réactivité.

Au niveau des dates, le TLFi a été rédigé de 1960 à 1990, il semble donc raisonnable de travailler sur des corpus postérieurs à la fin de la rédaction, afin de pouvoir détecter et caractériser des sens effectivement nouveaux. Selon le principe d'homéostasie (même si certaines unités lexicales évoluent vite, le lexique dans son ensemble évolue lentement), on peut considérer que l'essentiel des définitions lexicographiques reste d'actualité plusieurs décennies après. Comme l'effet de masse joue un rôle prépondérant dans notre démarche, il nous paraît valide de recourir au contenu sémantique du TLFi pour un corpus débutant dans les années 2000, sauf pour des champs particuliers, ayant connu une évolution explosive dans les années 90, comme l'informatique.

Plusieurs découpages temporels du corpus sont envisageables. Les plus élémentaires ont servi de base à nos expériences (*cf.* chapitre III.1), d'autres demanderont à être approfondis dans des développements ultérieurs :

- **Une période unique, sans division temporelle :** il y a contraste direct entre les nouveaux emplois en corpus et le sens codé. Cela repose sur l'hypothèse que l'image du sens codé et celle issue des emplois en corpus sont comparables. Or l'image du sens codé est une image synthétique, déformée, d'usages antérieurs. Le parallèle entre anciens emplois discursifs et représentation du sens codé est une étape intermédiaire occultée : on n'observe pas la façon dont le sens codé se manifeste en discours dans les emplois

antérieurs au changement de sens. Cette démarche implique soit une précision dans les représentations des sens lexicographiques et discursifs ainsi que dans les techniques de rapprochement des représentations, soit une analyse manuelle et un effort interprétatif marqués. En cela, on opte pour un certain risque, car les différences et similitudes sont construites directement à partir de deux images de nature différente (sac de traits de la ressource lexicographique *vs* saillances dans le corpus). De plus, une telle approche ne permet pas d'observer la diffusion dans les discours : il n'est pas possible d'accéder à la diffusion domaniale reflétée par les empreintes de fréquence, ni de visualiser l'évolution par interpolation comme le font (Boussidan et Ploux, 2011)[39] ; de même, des phénomènes de stratification du sens au cours du temps ne sont accessibles. Mais savoir confronter le sens codé à un corpus synchronique répond à un principe d'économie : économie d'étapes et de comparaisons, économie en ressources textuelles (pas de sous-corpus propre au passé). D'un point de vue complémentaire, ceci permet d'avoir des méthodes de construction du sens même lorsque qu'on dispose de ressources textuelles incomplètes (synchroniques) ou imparfaites (non-respect d'une certaine homogénéité dans le temps).

- **Deux périodes disjointes :** cette configuration permet de réaliser plusieurs comparaisons successives, en ne faisant varier qu'une chose à la fois, soit le temps (une image textuelle est comparée à une image textuelle antérieure), soit le type de représentation (l'image textuelle des anciens emplois est comparée à l'image lexicographique de l'ancien sens). Ce cas de figure offre un contrôle intermédiaire et il apparaît ainsi comme plus fiable. Il permet d'obtenir deux images : une image du changement proprement discursive (elle caractérise la variation des emplois discursifs dans le temps et elle est établie à partir de ressources de même nature) ; une image issue de la mise en relation de la représentation discursive des anciens emplois et de la représentation du sens codé. Le principe suivant s'applique ensuite : le nouveau sens codé est à l'ancien ce que les nouveaux emplois sont aux anciens emplois. L'existence de deux périodes disjointes permet d'exhiber des sauts ou des ruptures. Elle ne permet pas de voir la diffusion progressive du nouveau sens : les pics événementiels, une évolution stratifiée du sens, ou encore l'état de stabilisation ne sont pas accessibles à travers cette dichotomie. Cependant, elle est à même de refléter l'essentiel, c'est-à-dire les contrastes ou les similitudes. Des raffinements peuvent par la suite venir nuancer l'image obtenue et la corréler à une dynamique temporelle, mais ne sont pas indispensables dans un premier temps.
- **Plusieurs périodes disjointes :** elles permettent de préciser les étapes de l'évolution de sens. Elles peuvent notamment mettre en évidence la diffusion dans les discours, faire émerger une construction du nouveau sens par strates comme dans l'analyse de (Lecolle, 2007b), refléter des pics événementiels ou encore préciser l'évolution d'un foisonnement néologique.
- **Une période continue :** cette configuration présente les mêmes avantages que précédemment. Elle offre une souplesse supplémentaire dans le découpage en périodes. Elle permet des partitions, mais elle offre aussi des possibilités de lissage ou d'intégration (fenêtres temporelles glissantes). Elle permet aussi des élargissements de périodes parfois nécessaires pour disposer de données en quantité suffisante pour permettre des traitements de type statistique. Ce cas de figure semble plus adapté à une approche incrémentale,

[39] (Boussidan et Ploux, 2011) cherchent à observer des glissements sémantiques à partir de corpus découpés en tranches de temps. A chaque période est effectué un calcul d'affinités sémantiques, suivi d'une génération de cliques, c'est-à-dire de regroupements sémantiques. Les représentations successives sont mises en relation : une interpolation est générée entre les différents espaces sémantiques associés aux cliques à chaque période. L'interpolation n'est possible que parce que l'on dispose de représentations de même nature à au moins deux moments différents.

comme pour la base de données utilisée dans la Wortwarte (Lemnitzer et Ule, 2011), plateforme de détection de néologismes formels[40]. Ce découpage est potentiellement à même de montrer le détail de l'évolution. Il offre notamment la possibilité de rectifier des modèles d'interpolation. La multiplication des découpages pose toutefois le problème de la multiplication des traitements. Cette configuration présente un intérêt soit pour affiner une image, soit lorsque les données augmentent en flux continu.

En bref, pour des explorations préliminaires ou lorsqu'il n'est pas possible de disposer de données textuelles diachroniques, on se restreindra à une seule période, propre aux nouveaux emplois. Un corpus en deux périodes permet de construire une image du sens par contraste et s'inscrit plus dans une démarche différentielle (on n'est pas sûr de savoir transposer ou projeter correctement le sens codé en contexte dans l'absolu, mais une mise en correspondance du sens codé et en discours dans le passé peut, par analogie, éclairer sur la correspondance présente). Pour affiner l'image en exploitant la dynamique de l'évolution, on aura intérêt à recourir à une période continue ou à un ensemble de périodes disjointes.

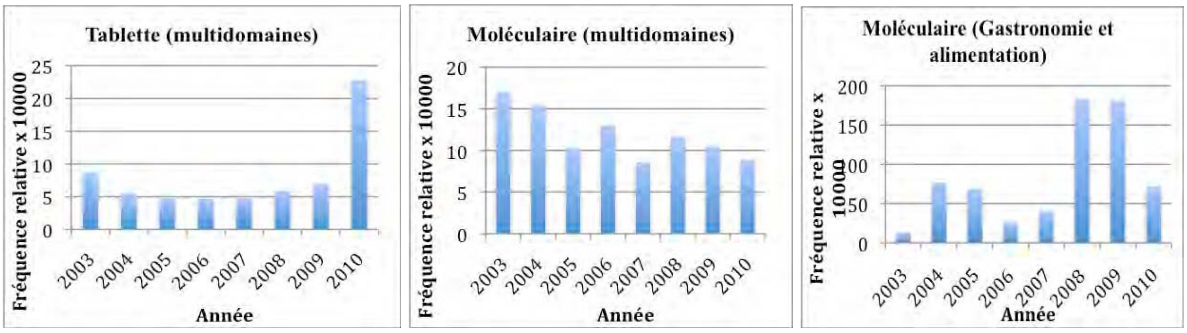
2.2 Un corpus adapté aux changements domaniaux

Dans le chapitre I.2, nous avons choisi un positionnement en français standard. Ce positionnement apparaît à travers le choix du TLFi comme ressource de référence et d'un corpus en adéquation avec ce choix, en l'occurrence un corpus journaliste (*cf.* chapitre I.1, 1.2.2).

Les changements domaniaux sont fondamentaux pour les néologies sémantiques ciblées. Pour les appréhender, il paraît préférable de disposer d'une multiplicité de domaines dans le corpus d'étude. Dans le cas d'une domanialisation, la diffusion peut certes être observée de façon restreinte, exclusivement au sein du nouveau domaine, mais la diffusion relative aux autres domaines d'emploi permet d'enrichir l'analyse. Ceci permet de voir le lien entre nouveau sens et ancien sens, ainsi que l'ampleur de la diffusion au sein du nouveau domaine. Ainsi, si l'on compare l'évolution de *tablette* (nouvel emploi lié aux nouvelles technologies) et *moléculaire* (nouvel emploi en gastronomie), on constate que l'emploi de *moléculaire* reste marginal par rapport à l'ensemble des emplois de cette unité lexicale, tandis que *tablette* connaît une diffusion massive. L'évolution temporelle des fréquences montre un accroissement des emplois pour *tablette* tous domaines confondus (figure II.1.3.a), mais pas pour *moléculaire* (figure II.1.3.b), alors que l'accroissement d'emploi est bel et bien apparent pour *moléculaire* lorsqu'on se restreint au domaine GASTRONOMIE ET ALIMENTATION (figure II.1.3.c)[41].

[40] La Wortwarte récupère quotidiennement les parutions d'un ensemble de journaux allemands. Les documents sont appelés à partir de la base utilisée pour la constitution d'un corpus de référence de l'allemand (projet "deutsche Referenzkorpus", DeReKo), dépouillés et éliminés une fois les néologismes extraits (Lemnitzer et Ule, 2011). Dans le projet DeReKo (Kupietz *et al.*, 2010), le corpus est constitué de façon incrémentale. Il comptait en mars 2011 plus de 4,1 milliards de mots et il augmente de 300 millions de mots par année. La constitution de ce corpus est régie par des licences aux conditions strictes (en particulier, accès contrôlé, par l'intermédiaire d'outils seulement, sans possibilité de reconstituer les données sources). Elle s'effectue techniquement grâce à la connaissance de la structure des métadonnées de chaque source, qui permet de récupérer les données dans un format de représentation inspiré des normes TEI. Le corpus est annoté en morphosyntaxe à partir de plusieurs étiqueteurs.

[41] Les fréquences relatives sont obtenues sur un corpus issu de la base de données journalistique Factiva. Les sources retenues sont les journaux français *Le Figaro, Libération, La Tribune, Ouest France, Les Echos* et *L'Expansion*, sur la période de 2003 à 2010. Le profil multidomaine est établi à partir des sujets (assimilés à des domaines) INFORMATIONS ECONOMIQUES, ARTS ET SPECTACLES, ENVIRONNEMENT, MODE DE VIE, POLITIQUE / RELATIONS INTERNATIONALES, SANTE, SCIENCE ET TECHNOLOGIE et SOCIETE / COMMUNAUTE / TRAVAIL. Le domaine GASTRONOMIE ET ALIMENTATION est un sous-domaine de MODE DE VIE. Les fréquences relatives sont



Figures 4.2.a), b) et c) : Évolution temporelle des fréquences relatives de (a) tablette et (b) moléculaire sur un ensemble diversifié de domaines, ainsi que (c) de moléculaire dans le domaine GASTRONOMIE ET ALIMENTATION.

De même, si le corpus ne recouvre qu'un domaine d'emploi, il n'est pas possible d'extrapoler à propos des autres domaines d'emploi : on ne peut savoir, par exemple, si un sens obtenu par métaphore correspond à une métaphorisation généralisée (le sens métaphorique se diffuse dans de nombreux domaines), ou s'il s'agit d'une métaphore spécifique à un domaine donné, qui correspond à une redomanialisation. Ainsi, pour *tsunami*, l'émergence de *tsunami financier* dans un corpus sur la crise financière ne permet pas de conclure si le sens métaphorique s'impose spécifiquement en économie ou s'il relève d'une diffusion plus générale, transposable à tous les domaines.

Un corpus thématique est donc pertinent pour observer l'émergence d'un nouveau sens. Cependant, l'analyse sur ce type de corpus reste partielle, elle risque de fournir une image du sens non pas erronée, mais incomplète. Le caractère complet ou incomplet de l'image fournie par un corpus thématique dépend de la finalité de l'allocation de signifié :

- s'il s'agit d'observer la dynamique lexicale d'un domaine particulier et de mettre à jour le vocabulaire qui peut y être employé, la restriction du corpus à un domaine est pertinente, elle suffit. Étant donné l'objectif, l'image du sens ne nécessitera pas d'être complétée ;
- s'il s'agit de mettre à jour une ressource de référence du français standard, il convient de compléter l'approche focalisée sur un domaine donné par une approche plus générale, afin de mettre en perspective les évolutions identifiées et de les observer par rapport à un ensemble de domaines.

2.3 Les unités : nature et loi de comportement

2.3.1 Nature des unités

Les unités en corpus ne sont pas données, elles sont construites et résultent de choix linguistiques ou pratiques (coût en temps, en difficulté de traitement ou en outils d'analyse pour segmenter, lemmatiser, etc.) (Lebart et Salem, 1994:33-42). Les formes graphiques sont les unités les plus faciles à identifier automatiquement, mais leur légitimité sémantique est critiquable : des unités sémantiquement distinctes telles que les homographes peuvent être confondues en une seule forme graphique ; des unités très proches sémantiquement, telles que les formes fléchies d'un même verbe, se retrouvent dissociées ; des unités complexes sont décomposées en unités simples qui, ainsi isolées, sont vecteurs d'un sens différent de celui de l'unité complexe. Des prétraitements permettent d'extraire des unités répondant à des critères linguistiques, par regroupement ou ajout d'information, mais ces unités sont parfois plus

calculées en divisant le nombre de documents où apparaît l'unité observée (pour l'année considérée et tous domaines confondus) par le total nombre de documents (pour l'année considérée et tous domaines confondus). Les valeurs obtenues sont multipliées par 10000 pour des raisons de lisibilité d'échelle.

difficiles à identifier et les procédures de prétraitement peuvent introduire des informations erronées ou contestables. Par exemple, on peut rencontrer des erreurs de lemmatisation, des regroupements morphologiques qui lissent des différences sémantiques réelles, ou encore des annotations sémantiques définies manuellement qui posent le problème de la subjectivité et de la reproductibilité du traitement. Le bénéfice de prétraitements reste discutable par rapport aux nouvelles erreurs et incertitudes qu'ils introduisent. En particulier, la lemmatisation a fait l'objet de nombreux débats, avec ses partisans et ses détracteurs : (Mellet, 2003) constate une amélioration sensible des résultats quantitatifs grâce à la lemmatisation ; à l'inverse, (Lemaire, 2008) y voit une perte d'informations essentielles propres aux formes fléchies ; d'autres auteurs écartent la question de la suprématie de l'une ou l'autre approche, mais de façon différente : (Brunet, 2000) souligne le faible impact de la lemmatisation sur les résultats statistiques et insiste ainsi sur la convergence de résultats, tandis que (Bécue-Bertaut, 2003) recommande une approche combinée et met ainsi en relief leur complémentarité.

Dans notre cadre, l'objectif est de pouvoir mettre en correspondance les unités du corpus et les entrées du TLFi, afin d'accéder à une représentation du contenu sémantique des unités provenant du corpus. Pour des raisons d'efficacité, le choix des unités est conditionné à la fois par la structure du TLFi et par les outils d'exploitation existants, c'est-à-dire la plateforme SEMY. On cherchera donc à se ramener à des unités lemmatisées simples. Celles-ci sont obtenues par segmentation en formes graphiques, lemmatisation au moyen de l'étiqueteur Treetagger, intégré dans Semy, puis élimination des mots-outils. Les fonctionnalités de Semy offrent la possibilité de travailler sur des collocations textuelles, qu'il aurait été possible de mettre en relation avec les syntitas du TLFi, mais cette option a été écartée pour plusieurs raisons : parce que les syntitas du TLFi regroupent indifféremment des syntagmes de nature différente, qu'ils soient figés ou non syntaxiquement, compositionnels ou non ; parce que la reconnaissance des collocations dans les textes est une opération délicate, au niveau de leur repérage et de leur mise en relation avec la base de syntitas du TLFi. Intégrer les collocations au processus aurait nécessité un investissement conséquent sur le plan théorique et sur le plan des outils de traitements, et dont la portée dépasse l'objet du présent travail.

2.3.2 Loi de comportement

Les unités observées en linguistique, qu'elles soient des formes fléchies, lemmatisées ou autres, répondent à des lois de comportement distinctes de celles qui s'appliquent aux observables d'autres champs disciplinaires (entités de la physique ou chimie par exemple). Ainsi, la loi de Zipf caractérise les unités linguistiques. Il s'agit d'une loi empirique systématiquement vérifiée (Pincemin, 1999:436) qui donne une approximation grossière de la distribution des unités lexicales. Elle affirme que, si les unités lexicales sont classées par fréquence décroissante, le rang est proportionnel à la fréquence de l'unité lexicale. La loi de Zipf est étroitement liée au problème de la dispersion des données : la distribution du vocabulaire comporte un grand nombre d'unités rares, notamment d'hapax, et un petit nombre d'unités très fréquentes. Par conséquent, une grande partie du vocabulaire reste inaccessible à une approche fondée sur des saillances, de type statistique : les unités rares sont filtrées. Or, dans le cadre de la théorie de l'information, les événements rares sont considérés comme porteurs de plus d'information. Deux questions se posent : la perte d'informations sémantiques a-t-elle des répercussions importantes ? Comment y remédier ?

(Manning et Schütze, 2003:199) étudient un modèle approprié pour des données dispersées et ils affirment que l'élimination des unités rares correspond d'une part à un gain considérable d'espace mémoire lors de traitements, d'autre part à une faible dégradation de la qualité du modèle, car les unités rares représentent une proportion considérable du vocabulaire, mais une petite partie des occurrences :

> « Because of the Zipfian distribution of words, cutting out low frequency items will greatly reduce the parameter space (and the Memory requirements of the system being built), while not appreciably affecting the model quality (hapax legomena often constitute half of the types, but only a fraction of the tokens). »

L'élimination systématique des unités rares n'est pas une position unanimement partagée. Ainsi, (Drouin *et al.*, 2006) choisissent, dans leur étude, de conserver un certain nombre d'unités rares, illégitimes par rapport aux traitements statistiques mais jugées sémantiquement pertinentes par des experts.

Pour éviter l'élimination des unités rares, il serait possible de travailler directement sur le contenu sémantique des unités, c'est-à-dire directement à partir d'une annotation en traits sémantiques plutôt que sur les unités lexicales. Cette solution a été testée, nous verrons qu'elle ne fait que repousser et amplifier le problème. Une explication est que la dispersion des données s'applique également au TLFi, donc aux traits sémantiques qui en sont extraits. De plus, la politique de rédaction du dictionnaire accentue l'éparpillement du vocabulaire : les rédacteurs ont cherché à utiliser un vocabulaire riche, précis, et tenté d'éviter les redondances. D'autres projets lexicographiques sembleraient plus adaptés pour réduire l'effet de dispersion. Ainsi, le projet RLF (Lux-Pogodalla et Polguère, 2011) s'inscrit dans une optique inverse, avec l'utilisation d'un vocabulaire simple, donc moins susceptible d'accroître la dispersion et plus favorable à une approche informatisée destinée à faire émerger des redondances et régularités.

2.4 Un espace structuré

Le corpus constitue un espace structuré, dans lequel on peut définir une topographie textuelle (Mellet et Barthélémy, 2009). Les structures qui nous intéressent sont :

- La ***structure spatiale*** correspondant à des unités textuelles de granularité décroissante : le texte, plus précisément l'article pour les corpus journalistiques comme unité globale ; le paragraphe comme unité locale, notamment pour délimiter l'environnement cooccurrentiel ; le syntagme, également associé au local, pour accéder aux concurrents (appartenance au paradigme associé à la cible lexicale).
- La ***structure temporelle***, pour permettre une approche diachronique et une comparaison entre anciens et nouveaux emplois.
- Une ***structure thématique***, ou plus exactement domaniale, pour accéder aux phénomènes domaniaux associés à la néologie sémantique et pour réduire l'ambiguïté sémantique.

À titre d'illustration, analysons par rapport aux trois structures ci-dessus l'exemple qui suit sur l'emploi de *toxique* en contexte de crise financière :

> « Plus de trois heures de débats enflammés ont précédé quarante minutes de vote chaotique, durant lequel des élus criaient le cours du Dow Jones. Des représentants ont comparé leur décision à l'octroi de pouvoirs de guerre ou à l'"impeachment" du président. Jeb Hensarling, républicain du Texas, a jugé qu'elle mettrait la nation "sur la pente glissante du socialisme". Son collègue démocrate Lloyd Dogget a reproché aux négociateurs de "n'avoir jamais envisagé sérieusement une alternative" au rachat des **créances "toxiques"** de Wall Street. » (*Le Figaro*, 30/09/2008)

- ***Au niveau de la structure spatiale***, cet exemple s'insère dans un corpus découpé en articles journalistiques, certains provenant du *Monde Diplomatique* sur la période 1994-1998, certains provenant du *Figaro* ou de *l'Humanité* sur la période de septembre 2008 à février 2009. *Toxique* appartient à l'article intitulé "Pourquoi le Congrès a sabordé le plan Paulson" (unité globale). Les articles comportent une sous-structure, ils sont segmentés en

paragraphes. Le paragraphe présenté ci-dessus constitue un espace inclus dans l'article qui définit de façon locale l'environnement de *toxique*. Parmi les syntagmes selon lesquels se décompose le paragraphe, le syntagme *créances toxiques* définit l'environnement de *toxique* de façon encore plus locale.

- ***Au niveau de la structure temporelle***, le corpus est structuré en deux périodes : une période de référence pour d'anciens emplois, à savoir entre 1994 et 1998 (sous-corpus du *Monde Diplomatique*) ; une période caractéristique de nouveaux emplois, fin 2008 début 2009. L'exemple mentionné, du 30/09/2008, provient de la nouvelle période.
- ***Au niveau de la structure thématique***, le corpus (sous-corpus des anciens emplois et sous-corpus des nouveaux emplois) est constitué d'articles appartenant à des domaines variés. Le sous-corpus de fin 2008-début 2009 (nouveaux emplois) est thématique, il porte sur la crise financière et les articles appartiennent au domaine économique ou financier. Le sous-corpus de 1994 à 1998 (anciens emplois) comporte également des articles économiques ou financiers, mais les articles se répartissent également selon d'autres domaines (politique, environnement, etc.). L'exemple est issu du domaine économique et sera confronté à des emplois antérieurs issus d'autres sous-ensembles de la structure en domaines, par exemple, des articles appartenant au domaine de l'environnement.

Bien que la notion de structure puisse évoquer l'idée d'un espace figé, donné a priori, un espace en expansion reste compatible avec une structure, pourvu que les caractéristiques de l'espace se prolongent aux parties incrémentales : la structure se propage alors de proche en proche. Ainsi, le corpus peut être un espace fixe ou évolutif. Par corpus évolutif, on entend un corpus qui s'accroît progressivement au cours du temps, enrichi par un flux de données continu. À titre d'exemple, l'Observatoire de Néologie (Cabré *et al.*, 2003)[42], la plateforme Télanaute (Issac et Ouenniche, 2010)[43] ou encore la Wortwarte (Lemnitzer et Ule, 2011) génèrent des corpus de façon dynamique, augmentés en permanence par incrémentation. Nos corpus d'étude seront des corpus fixes, destinés à mettre en place des protocoles expérimentaux, mais à terme, il conviendra d'adapter les propositions faites à un corpus dynamique, pour répondre à la problématique de veille qui sous-tend notre démarche.

2.5 Application : les corpus utilisés comme supports d'expériences

2.5.1 Corpus 'Crise financière', terrain d'observation de nouveaux sens, et corpus de comparaison du *Monde Diplomatique*

Le corpus 'Crise financière' est un corpus thématique sur la crise économique et financière amorcée en 2008. Il s'étend de septembre 2008 à février 2009. Il est issu du discours journalistique et comporte 1 587 articles de presse, soit près d'un million d'occurrences de formes et un vocabulaire d'environ 35 000 formes. Les articles sont issus de deux quotidiens

[42] L'Observatoire de Néologie télécharge des données journaux de plusieurs façons : à partir de nouvelles délivrées par les éditeurs par exemple sous forme de CD, par téléchargement page à page, par un programme qui parcourt et télécharge des pages HTML obtenues par le web en fonction du nom du journal, par récupération automatique de nouvelles envoyées par courrier électronique après abonnement. Les données détectées sont converties du HTML vers un format SGML standard.

[43] La plateforme récupère des données du web de façon incrémentale à partir d'un robot (crawler). Elle parcourt des pages préalablement listées, susceptibles d'être mises à jour par les administrateurs du site (par exemple, la page d'accueil du journal *le Monde* www.lemonde.fr) et définies par certaines contraintes sur l'URL ou le contenu. Elle récupère la page, en extrait les informations respectant les contraintes imposées par l'utilisateur, elle effectue divers traitements (filtrage, segmentation, etc), définis au préalable, et stocke la page avec une indication sur la date de récupération (Issac et Ouenniche, 2010).

français aux lignes éditoriales contrastées, *Le Figaro* et *L'Humanité*. La sélection des articles s'est faite manuellement, sur critère de pertinence avec la thématique choisie, par parcours des archives et, afin d'équilibrer en volume les deux journaux, par filtrage des articles selon le mot-clé *crise* dans le cas du *Figaro*.

Dans les expériences réalisées, ce corpus n'a pas fait l'objet d'un découpage en périodes : l'ensemble du corpus a été traité comme une période unique, correspondant au nouveau sens. Deux ressources de comparaison ont servi à étudier de nouveaux sens. D'une part, le corpus a été soumis à des confrontations directes avec le TLFi, sans comparaison intermédiaire avec une base textuelle équivalente. D'autre part, il a été contrasté avec une ressource textuelle antérieure. Cette ressource est un corpus de presse, il correspond aux archives du mensuel français *Le Monde Diplomatique* de 1994 à 1998.

Le corpus 'Crise financière' tout comme le corpus issu du *Monde Diplomatique* disposent d'une structure en articles et en paragraphes.

2.5.2 Corpus 'Outreau'

Ce corpus porte sur l'affaire judiciaire d'Outreau. Il est constitué d'articles de presse parus entre novembre 2001 et avril 2006, sélectionnés sur critère de présence du nom *Outreau* et nous a été gracieusement fourni par sa conceptrice. Il a été initialement constitué dans le cadre de l'étude linguistique de la polysignifiance du nom propre *Outreau* (Lecolle, 2007). L'allocation de signifié sera étudiée relativement à l'unité lexicale *Outreau*.

Selon (Lecolle, 2007 et 2009), l'évolution diachronique du sens d'*Outreau* peut s'observer à travers un découpage en cinq périodes-clés, correspondant à des temps forts dans la succession des événements concernant l'affaire d'Outreau :

- 1) 2001-2002 : découverte d'un réseau pédophile à Outreau, arrestations
- 2) mai-juin 2004 : procès de Saint-Omer
- 3) 1-2/07/2004 : attente du verdict de Saint-Omer
- 4) 3-8/07/2004 : verdict du procès
- 5) 2/12/2005 à avril 2006 : procès en appel à Paris ; suite et conséquences (commission d'enquête parlementaire).

La structure temporelle du corpus utilisée dans nos expériences respecte le découpage proposé ci-dessus. À cette structure temporelle s'ajoute un découpage en articles et en paragraphes.

2.5.3 Corpus 'Factiva'

Le corpus 'Factiva' est un corpus multidomaines. Il est issu de la base de données d'actualité internationale *Factiva*. Cette base de données, constituée de plus de 10 000 sources, notamment journalistiques (presse nationale aussi bien qu'internationale), a été restreinte à une sélection de journaux francophones et de thématiques. Les sources sélectionnées sont les journaux *Libération, Le Figaro, L'Humanité, Ouest-France, La Tribune, Les Échos* et *L'Expansion*. Les thématiques (définies comme *sujets* dans le moteur de recherche de Factiva, et que nous qualifierons par la suite de domaines) sont les suivantes : informations économiques, arts et spectacles, environnement, mode de vie, politique / relations internationales, santé, science et technologie, société / communauté / travail.

Le corpus compte 1,2 million d'articles. Sa restriction aux deux années 2004 et 2010, base d'une série d'expériences, est d'environ 300 000 articles. L'unité textuelle est l'article et sert également d'unité de décompte (par exemple, la taille d'un sous-corpus ou encore les

occurrences d'une cible lexicale seront évaluées en nombre d'articles), contrairement aux autres corpus où l'unité textuelle est le paragraphe et l'unité de décompte est la forme lexicale ou le sème (par exemple, un sous-corpus de taille 20 000 est constitué de 20 000 occurrences de formes ou sèmes).

Le corpus sera utilisé dans les expériences pour observer l'évolution domaniale. Il ne sera pas analysé à travers des cooccurrences lexicales ou infra-lexicales, mais seulement supra-lexicales (domaines) . Le découpage en domaines est fondamental. De même, le découpage en périodes aura un rôle important pour étudier la diffusion progressive de nouveaux emplois.

3. Une articulation non immédiate des deux représentations

Les représentations issues de la ressource lexicographique et les ressources textuelles ne sont pas directement comparables : elles peuvent poser des problèmes de recouvrement de vocabulaire (3.1) ; dans l'une, les unités sont hors contexte, dans l'autre, en contexte, ce qui pose le problème de l'ambiguïté dans le basculement de la première vers la deuxième (3.2) ; elles sont issues de structures d'ensemble différentes, chacune avec ses caractéristiques propres (3.3). Nous préciserons notre démarche pour établir des correspondances (3.4).

3.1 Recouvrements imparfaits entre le vocabulaire textuel et les entrées du dictionnaire

Certaines formes du corpus n'ont pas d'entrée correspondante dans le dictionnaire, il existe des lacunes. Comme évoqué, ces lacunes peuvent être dues à l'évolution du lexique (le verbe *titriser* est absent de la version actuelle du TLFi ; il devrait normalement être intégré au *Supplément du TLFi*), à des contraintes éditoriales (place limitée entraînant une suppression ou une réduction des entrées, *cf*. (Martinez, 2009:48-60)) ou à des choix linguistiques de délimitation du vocabulaire (par exemple, le choix entre vocabulaire spécialisé et de langue générale ; ainsi, pour les figures acrobatiques, le substantif *flip-flap* a une entrée qui lui est propre et *roue* possède une définition associée au domaine technique de la gymnastique, tandis que *salto* n'a pas d'entrée associée, alors qu'il s'agit d'une figure tout aussi centrale que les précédentes dans la discipline) ;

Les données les plus immédiatement accessibles en corpus sont les formes lexicales alors que les lemmes sont les points d'entrée du TLFi. La lemmatisation du corpus dépend d'un outil, en l'occurrence l'étiqueteur Treetagger, qui peut d'une part commettre des erreurs d'étiquetage (par exemple, dans notre corpus sur la crise financière, l'adjectif *japonaises* a été identifié comme un verbe au présent, et lemmatisé en *japonaiser*), d'autre part affecter des étiquettes pertinentes, mais qui divergent de celles du TLFi (par exemple, dans le corpus de (Lecollle, 2007b) sur Outreau, l'adjectif *pédophile* étiqueté comme adjectif par Treetagger n'avait pas d'équivalent dans le TFLi, où *pédophile* est répertorié comme substantif).

3.2 Écart entre des unités lexicales contextuelles et des entrées décontextualisées

Les entités lexicales présentes dans le corpus apparaissent de façon contextualisée et, de ce fait, sont non ambiguës pour un interprétant : guidé par le contexte, ce dernier associe spontanément un sens pertinent aux unités (Rastier et Valette, 2009). Dans notre cadre, le sens n'est pas affecté par un interprétant, mais il est recherché là où il est codé, c'est-à-dire dans un dictionnaire. Or les entrées de dictionnaire sont le lieu de représentation de la polysémie : une entrée est une synthèse d'emplois, elle est censée contenir les éléments d'information capables d'éclairer sur le sens d'un vocable quel que soit son contexte d'emploi. Les entrées ne sont pas

liées à un contexte d'emploi particulier, elles sont portables d'un contexte à l'autre, autrement dit, elles sont par essence décontextualisées.

Pour relier les unités lexicales présentes en corpus avec le sens codé, il faut dans un premier temps extraire des unités, donc perdre une partie du contexte plus ou moins importante, dans un second temps les associer à une entrée par l'intermédiaire d'un mot-vedette, enfin, naviguer dans la masse d'information contenue dans l'entrée, dont la totalité du contenu risque d'introduire de l'ambiguïté, afin de sélectionner le contenu pertinent. De façon schématique, l'accès aux unités de sens pertinentes peut se décomposer en une série d'opérations élémentaires :

- *point de départ* : unités lexicales en contexte ;
- découpage du contexte en unités dans un formalisme qui permet de l'associer à une entrée de dictionnaire ;
- extraction des unités du contexte (où une unité se décompose en un vocable formellement identique à un mot-vedette du dictionnaire plus des informations additionnelles extraites du contexte d'emploi) ;
- appariement entre unité issue du corpus et dictionnaire : substitution du contenu de l'entrée au vocable ;
- conversion des informations additionnelles en clés de navigation dans le contenu de l'entrée ;
- application des clés de navigation au contenu de l'entrée ;
- *sortie :* éléments de sens pertinents associés à chaque unité.

La mise en relation entre unités lexicales contextualisées et entrées lexicographiques par essence décontextualisées est loin d'être immédiate : le type d'unités extraites, le choix des informations additionnelles, leur conversion en clé de navigation dans les entrées de dictionnaire et l'application de ces clés au contenu sont à construire. En l'absence de clés de sélection, ou lorsqu'elles ne sont pas suffisamment filtrantes, le contenu sémantique affecté peut introduire une ambiguïté interprétative.

Dans un traitement automatique, les unités les plus faciles à extraire (techniquement parlant) ont plus de risque de devenir ambiguës que des unités moins évidentes à obtenir. Ainsi, il est plus simple d'obtenir automatiquement des mots-formes que des unités enrichies par une analyse morpho-syntaxique (étiquettes grammaticales par exemple) ou syntactico-sémantiques (différenciation entre homonymes, intégration d'informations à même d'exclure certaines ambiguïtés sémantiques). Supposons qu'on dispose d'un formalisme minimal pour associer les sorties du corpus aux entrées du dictionnaire (représentation sous forme lemmatisée, c'est-à-dire sous une forme conventionnelle avec une étiquette grammaticale, sans information additionnelle donnant des indices sur le contexte d'emploi). Ce formalisme minimal risque de rapatrier un contenu sémantique non pertinent, notamment dans le cas d'homonymes. À l'inverse, choisir d'intégrer des informations de désambiguïsation demande un formalisme enrichi pour représenter les unités du corpus. Il y a de fortes chances pour que ce formalisme ne puisse être mis en relation directe avec les représentations du dictionnaire.

Considérons un exemple extrait de nos analyses (*cf.* chapitre III.1) : dans notre corpus sur la crise financière, le substantif *titres* est une unité lexicale sélectionnée pour éclairer le sens du candidat à la néosémie *toxique*. La lemmatisation et la recherche de l'entrée associée sont relativement immédiates, mais le contenu sémantique associé à *titre* génère du bruit dès lors qu'on consulte la ressource lexicographique sans ajouter d'informations relatives à l'objet textuel : *titre* évoque aussi bien l'idée de distinction honorifique, étrangère au corpus, que

l'idée de produit financier. À l'inverse, une représentation enrichie en information textuelle de *titres,* avec ajout d'une étiquette de domaine (économie par exemple) ou association à un ensemble d'unités distributionnellement proches (comme *créances, crédits, actifs*), fournit des clés de navigation dans l'entrée de dictionnaire associée, mais la récupération du contenu sémantique pertinent n'est pas immédiate. Ainsi, la correspondance entre sorties du corpus et entrées du dictionnaire requiert un formalisme et des procédures adaptés, conditionnés par un arbitrage entre qualité et coût : coût pour extraire l'information ad hoc du corpus, et coût pour mettre en place une procédure de navigation intelligente dans la ressource lexicographique à l'aide de cette information supplémentaire.

3.3 Structure du corpus vs structure du dictionnaire

Le sens en corpus dépend de la structure du corpus, entre autres de l'environnement cooccurrentiel, de thèmes ou domaines, ou encore de la syntaxe. Le sens d'une entrée lexicographique est précisé par la structure interne du dictionnaire, en particulier avec une décomposition en définitions, notamment en fonction de domaines. Pour mettre en parallèle les structures dictionnairiques et textuelles, il est nécessaire de disposer de représentations qui entretiennent des similitudes et des corrélations. Par exemple, il semble indispensable de pouvoir mettre en relation des informations globales provenant du corpus avec des informations macrosémantiques ou mésosémantiques dans le dictionnaire, comme les indications de domaines.

3.4 Méthodes d'articulation mises en œuvre

Nous proposons deux façons d'établir des liens entre corpus et dictionnaire :

- projeter l'information du dictionnaire (*i.e.* les traits sémantiques) sur le corpus, puis se servir de la structure du corpus pour moduler et structurer l'information sémantique ;
- extraire des unités lexicales du corpus, puis naviguer dans la ressource lexicographique pour en extraire des traits sémantiques communs ou pour utiliser les traits sémantiques communs comme clés de structuration de l'information lexicale.

Ces deux démarches ne sont pas exclusives, elles offrent des regards complémentaires et méritent d'être combinées. Nous détaillerons la procédure adoptée au (chapitre III.1).

4. Bilan

Aussi bien au niveau du dictionnaire qu'au niveau des données textuelles, l'analyse des ressources et leur mise en relation nécessitent de manipuler de grandes quantités de données et de recourir à des techniques d'analyse des données. Nous présentons dans le chapitre qui suit quelques outils mathématiques et informatiques pour permettre un traitement effectif des données.

Chapitre II.2
Des mesures statistiques aux structures sémantiques

L'allocation de signifié à partir des ressources décrites au chapitre précédent nécessite des techniques adaptées aux bases de données importantes. Ces techniques peuvent être issues de différents courants, relativement voisins : textométrie, recherche d'information, fouille de données ou encore TAL. Notre approche s'inscrit dans le cadre de la textométrie (Pincemin, 2009 ; Equipe TXM, 2011). Issue d'un courant français, la statistique linguistique initiée par (Muller, 1968), la textométrie accorde une part importante aux fondements aussi bien linguistiques que mathématiques, ainsi qu'à l'articulation entre théorique et empirique. Elle cherche à exploiter la dimension textuelle : d'une part, elle favorise le retour au texte et le contrôle manuel, dans le même esprit que l'analyse du discours ; d'autre part, elle accorde une place non négligeable aux paliers de textualité larges, au-delà du syntagme, ce qui la distingue notamment du TAL. Enfin, elle privilégie les approches contrastives : son objectif est de « mettre en évidence les spécificités et les contrastes significatifs. » (Equipe TXM, 2011). Notre approche rejoint la perspective de la textométrie puisque nous cherchons à caractériser une rupture sémantique en exhibant des unités propres aux nouveaux emplois, qui reflètent des contrastes avec des emplois antérieurs.

Deux types de structures peuvent être associés aux ressources : une structure d'ensemble, témoin de l'organisation générale, systémique ; une structure focalisée, qui reflète comment s'organise le contenu des ressources relativement à la cible lexicale. Ces structures seront les supports des modèles mathématiques présentés dans ce chapitre. Elles seront assimilées à des espaces mathématiques, dont les éléments seront l'équivalent des unités lexicales ou sémiques. De même que des distances peuvent être calculées entre éléments selon leur distribution dans l'espace mathématique, de même, l'affinité sémantique d'unités lexicales ou sémiques découlera de leur répartition dans ces structures. Disposer de plusieurs structures revient à se positionner dans plusieurs espaces mathématiques, ou encore à redéfinir la configuration des éléments dans l'espace mathématique. L'évaluation des affinités sémantiques sera établie en fonction des différentes structures, elle repose ainsi sur un jeu de structures, ou encore un jeu d'espaces.

Dans un premier temps, on définira la structure associée à chaque ressource (section 1). Puis on précisera comment obtenir des pondérations à partir de ces structures pour en extraire des unités remarquables (section 2). On abordera ensuite la façon de combiner l'information issue de plusieurs jeux de contrastes (section 3), pour en dégager des structures sémantiques

(section 4). Enfin se posera la question de l'accès aux informations issues des traitements quantitatifs, à travers les représentations visuelles (section 5) et la limite de validité des résultats (section 6).

1. Définir l'espace mathématique associé aux ressources

Le corpus et la ressource lexicographique correspondent à deux espaces distincts, dotés chacun d'une structure propre. On abordera d'abord la description de l'espace associé au corpus, puis celle de l'espace associé à la ressource lexicographique.

1.1 Espace associé au corpus

1.1.1 Formalisation de la structure

L'espace E associé au corpus peut se structurer en différents sous-espaces en fonction de différents paramètres et de la présence de la cible lexicale.

Les *paramètres* sont les suivants :

- ***Les domaines D_i :*** le corpus se divise en m domaines, m≥1. Les domaines ne forment pas nécessairement une partition. En particulier, il peut y avoir des chevauchements.
- ***Le temps t_j*** : le corpus est découpé en 1, 2 ou p tranches de temps ou périodes. Dans notre cadre, les périodes sont disjointes.
- ***Le palier textuel, P***, qui permet de décomposer le corpus en K unités textuelles $u_k(P)$, $1 \leq k \leq K$: il correspond approximativement à une longueur spatiale, dont les limites dépendent de règles linguistiques. Il peut prendre trois valeurs dans notre cadre, correspondant à trois types d'unités textuelles : le document (ou article, étant donné que nos corpus sont des corpus de presse) ; le paragraphe ; le syntagme. Ces trois paliers entretiennent des relations d'inclusion. Les documents et les paragraphes constituent une partition du corpus : $E = \bigcup_{k=1}^{K} u_k(P)$, où P = palier du document ou P = palier du paragraphe.

Ces paramètres divisent l'espace textuel E en sous-espaces E(D,t,P).

La présence de la cible permet de décomposer en deux l'espace du corpus, ou éventuellement un sous-espace du corpus (par exemple, le sous-espace associé à un domaine fixé). Autrement dit, dans une perspective ensembliste, on obtient deux sous-ensembles : le voisinage de la cible associé aux paramètres choisis et son complémentaire. Le voisinage est défini par :

- la présence de la cible lexicale, qui peut se traduire comme la valeur du booléen "la cible appartient au sous-ensemble" :
- un paramètre propre au voisinage, l'ordre du voisinage o, où $o \in \{1 ; 2\}$; les cooccurrents sont définis dans un voisinage d'ordre 1, les concurrents dans un voisinage d'ordre 2 (ils sont obtenus à partir des cooccurrents des cooccurrents) . Nous avons choisi de définir une contrainte sur le palier selon l'ordre du voisinage : le palier est le paragraphe pour le voisinage d'ordre 1 et le syntagme pour le voisinage d'ordre 2. Autrement dit, les cooccurrents d'une cible lexicale sont des unités qui apparaissent dans le même paragraphe que celle-ci ; les concurrents sont obtenus en recherchant d'abord des unités qui cooccurrent avec la cible au sein d'un syntagme, puis les cooccurrents de ces unités, recherchés à leur tour au sein de syntagmes contenant la cible.

La *structure* de l'espace se définit par la combinaison des structures introduites par chacun des paramètres :

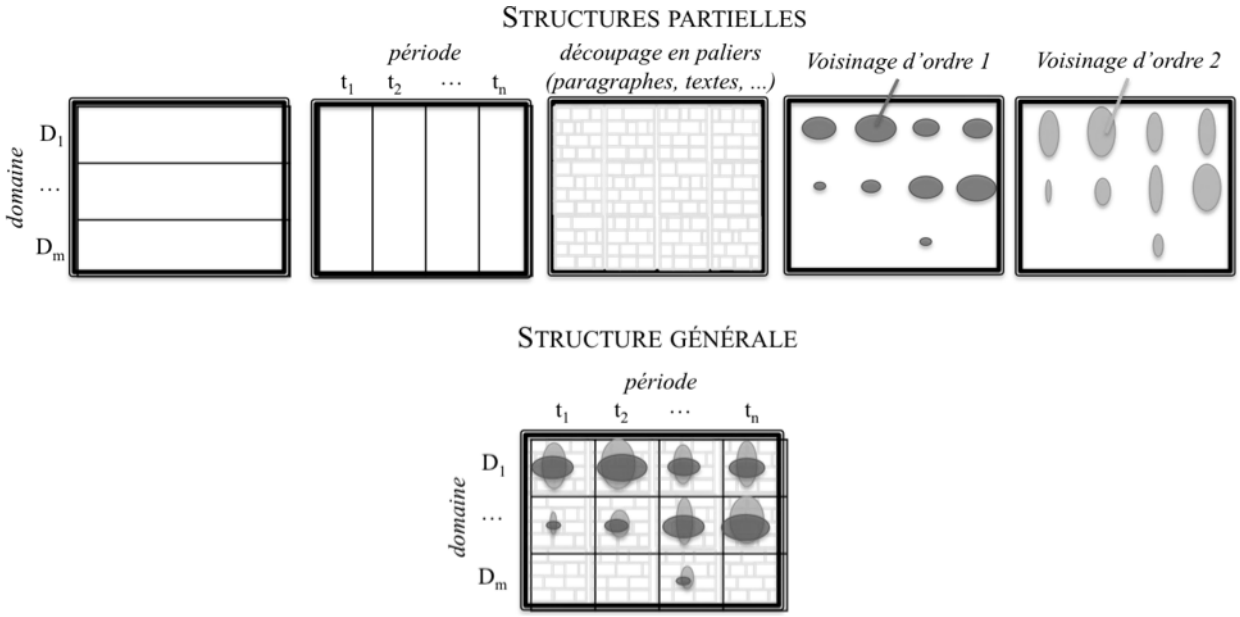


Figure II.2.1 : Structure de l'espace textuel

1.1.2 Relations et propriétés

Au sein de chaque sous-espace défini par un ensemble de paramètres, la structure interne s'efface : les sous-ensembles sont alors de type "sac de" (sac d'unités lexicales ou sac de traits sémantiques), avec des éléments indifférenciés a priori au sein du sous-ensemble ou de son complémentaire. La relation qui prime est la ***relation d'appartenance*** au sous-ensemble : appartenance de la cible à un sous-espace doté d'une certaine étiquette de domaine, appartenance d'unités lexicales ou sémiques au sous-ensemble des voisinages de la cible, ou inversement à son complémentaire, etc.

La construction des voisinages d'ordre 1 ou 2 de la cible lexicale (cooccurrents ou concurrents) amène à s'interroger sur deux propriétés de la relation "est dans le voisinage de" : la ***réflexivité*** et la ***transitivité***. La cible fait-elle partie de son propre voisinage ? Par ailleurs, dans quelle mesure une unité appartenant au voisinage d'une unité voisine de la cible est-elle voisine d'ordre 2 de la cible ?

Au niveau de la ***réflexivité***, le voisinage de la cible peut être envisagé de façon stricte (la cible est exclue, ou plus exactement éliminée du corpus après génération du voisinage) ou large. Si la cible est incluse, il y a déséquilibre par construction, avec une saillance maximale pour la cible elle-même dans son voisinage. L'inclusion ou l'exclusion de la cible dépendent de ce qu'on cherche à observer.

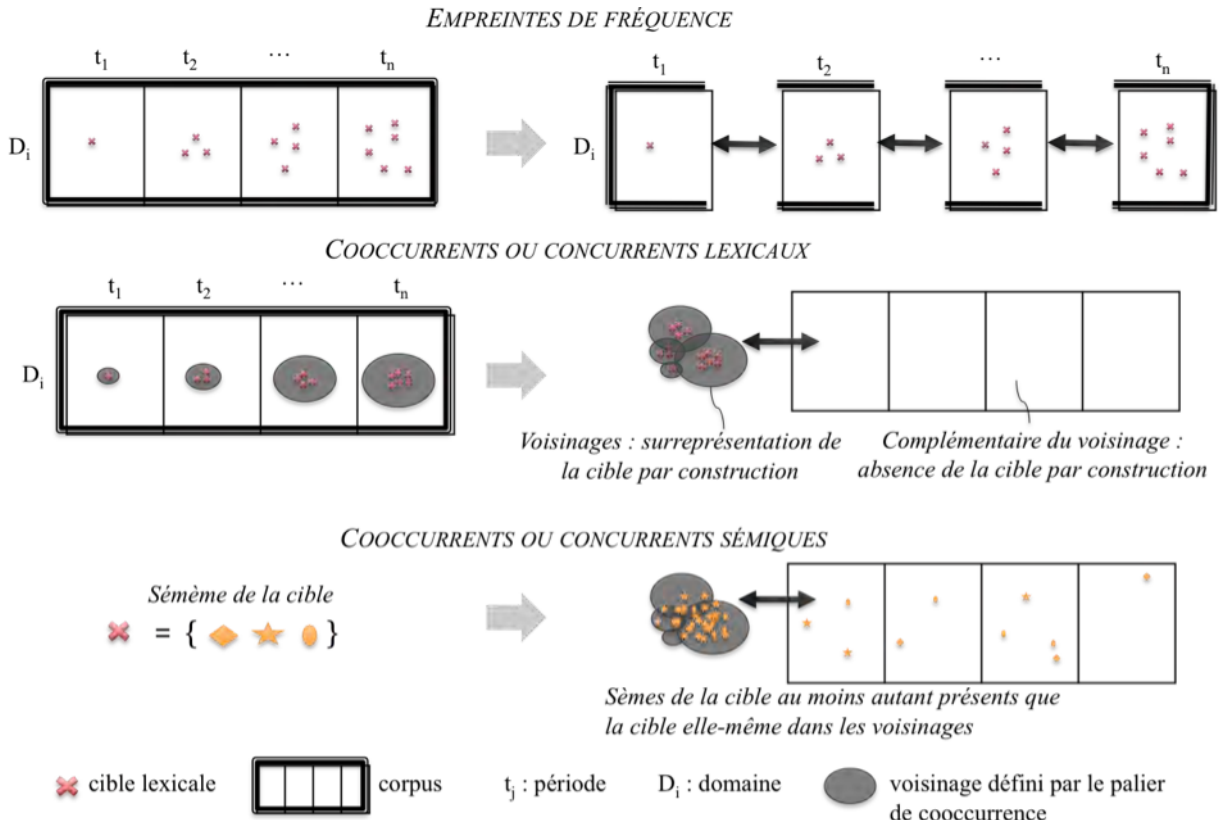


Figure II.2.2 : Inclusion ou exclusion de la cible de l'ensemble des observables selon l'axe d'analyse

Pour les empreintes de fréquence, on cherche à évaluer la diffusion de la cible dans le temps et éventuellement par rapport aux domaines ; la cible est alors l'observable, il est hors de question de l'exclure. Lorsque l'espace est structuré relativement à la cible, autrement dit, lorsqu'on observe le voisinage de ses cooccurrents ou de ses concurrents, la situation n'est pas la même : la cible est le pôle, elle sert de clé de structuration de l'espace, et les observables sont les unités lexicales ou sémiques du voisinage, non la cible elle-même. Lorsque les observables sont des unités lexicales (donc en l'absence d'annotation sémantique), conserver la cible n'est pas problématique : d'une part, elle est l'unique entité à créer le biais et elle est facile à identifier ; d'autre part, l'évaluation de sa saillance peut servir de point de comparaison aux autres valeurs de saillance calculées sur le voisinage. En revanche, pour une analyse sémique, qui comporte une annotation sémique du corpus, il semble préférable d'exclure la cible de son voisinage. En effet, celle-ci est remplacée par son sémème, donc l'ensemble des sèmes de sa définition. La saillance de la cible est due au fait que toutes ses occurrences sont dans le voisinage, par construction même des voisinages ; cette saillance par construction aura tendance à se transmettre à ses sèmes, car il y aura au moins autant d'occurrences de chacun des sèmes que d'occurrences de la cible. Cet effet risque de brouiller les informations apportées par le voisinage strict, à savoir la récurrence de ces sèmes dans le voisinage et, de ce fait, leur modulation.

La question de la ***transitivité*** se pose pour la construction du paradigme des concurrents. Les concurrents sont issus des cooccurrents de cooccurrents. Les cooccurrents d'une cible lexicale éclairent sur son sens, mais en est-il de même pour les cooccurrents des cooccurrents ? Autrement dit, en considérant que cooccurrence et clés d'accès au sens sont liées, la contribution à l'accès au sens est-elle une propriété transitive ?

La propriété de transitivité est loin d'être respectée en linguistique. Par exemple, la synonymie n'est pas transitive (*crise* est synonyme de *secousse*, lui-même synonyme de *spasme*, et pourtant, *crise* n'est pas synonyme de *spasme*). Au niveau des affinités sémantiques portées par la cooccurrence, un certain nombre de travaux vont dans le sens d'une forme de transitivité en terme d'affinités sémantiques pour la cooccurrence d'ordre 2 : (Sjöblom, 2007) utilise la fonction d'*associations privilégiées* du logiciel *Hyperbase* (Brunet, 2011), qui génère un réseau à partir des cooccurrents d'ordre 1 et 2 d'un mot-pôle, et elle voit dans les résultats des "réseaux isotopiques" capables de mettre à jour des isotopies thématiques.

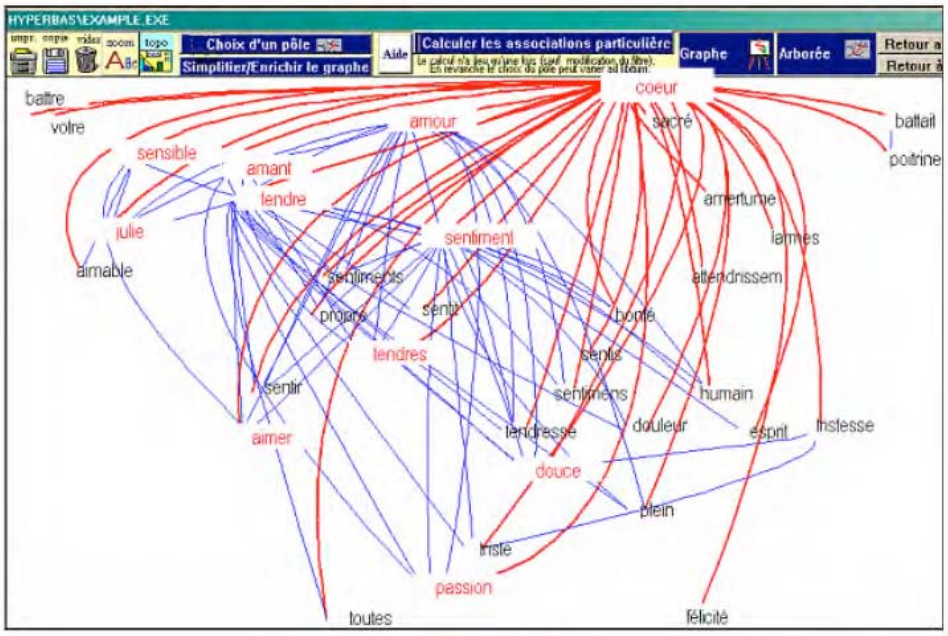


Figure II.2.3 : Exemple d'associations privilégiées extrait du manuel Hyperbase de Brunet ; associations privilégiées construites autour de cœur dans un corpus d'écrits de Balzac

De même, (Mayaffre, 2008b) considère que ce traitement des cooccurrences d'ordre 1 et 2 fait émerger des "faisceaux isotopiques non triviaux" : cinq isotopies, renvoyant à l'histoire, à du pathétique, à l'autorité, à la famille ou société et à la religion se forment et se combinent autour du mot *patrie* dans un corpus de discours de Nicolas Sarkozy. En revanche, (Heiden et Tournier, 2001) énoncent des conclusions plus prudentes sur les liens sémantiques obtenus à partir de cooccurrents d'ordre 2 : dans leur étude, ils utilisent des *lexicogrammes récursifs*[44] pour analyser le *je* et le *moi* de Chirac et Jospin, candidats adverses aux présidentielles de 1995, et y voient moins des informations sur le sens des mots-pôles que des indications sur les stratégies des candidats pour se distinguer l'un de l'autre et se positionner par rapport aux Français.

Nous sommes d'avis que la cooccurrence d'ordre 2 permet de faire émerger des affinités sémantiques qui n'apparaissent pas avec la cooccurrence simple, donc qu'il existe une relation proche de la transitivité : tout comme les cooccurrents d'ordre 1, les cooccurrents d'ordre 2 peuvent participer à l'accès au sens. Mais les liens d'affinités sémantiques ne s'obtiennent pas directement et nécessitent de faire intervenir d'autres relations, par exemple lexicographiques.

[44] Les *lexicogrammes* sont une synthèse des cooccurrents situés à droite et à gauche d'une cible lexicale, utilisée comme pivot. Ils se présentent sous forme d'une liste hiérarchique de cooccurrents établie sur un critère probabiliste (spécificités de Lafon) et dotée de valeurs complémentaires, telles que la distance du cooccurrent à la cible. Les *lexicogrammes récursifs* sont obtenus par récurrence : la méthode des lexicogrammes est appliquée à la cible, puis aux cooccurrents de la cible, qui deviennent alors les pivots, et ainsi de suite. Ceci permet de générer des réseaux de cooccurrents (Heiden, 2004).

La cooccurrence d'ordre 2 brute, sans traitement, tend à accroître la dispersion de l'information sémantique ainsi qu'à diversifier la nature des liens sémantiques, et ce d'autant plus que le palier de cooccurrence est large : il existera des cooccurrents d'ordre 2 qui n'informeront pas sur le contenu sémantique de la cible et y seront même totalement étrangers.

Dans le cadre d'une approche sémique, l'annotation sémique telle que nous la pratiquons tend à accroître la dispersion de l'information ; la cooccurrence d'ordre 2 amplifie encore cette dispersion, ce qui exige des traitements plus poussés pour filtrer le contenu de façon pertinente. Pour la construction du paradigme de concurrents (cooccurrents d'ordre 2 destinés à constituer un paradigme de substitution), on ne recherchera pas des affinités sémantiques obtenues à partir de paliers textuels de l'ordre du paragraphe, mais à partir de paliers textuels plus réduits (syntagmes). D'une certaine façon, on fait l'hypothèse que la transitivité de l'affinité sémantique a plus de chance de se réaliser à partir de cooccurrence syntagmatique : il faudra probablement moins de filtres ou des filtres moins sévères pour éliminer les cooccurrents d'ordre 2 non pertinents.

1.1.3 De partitions en deux sous-ensembles à une structure complexe

Pour établir des affinités sémantiques, le point de départ est toujours une décomposition simple de l'espace textuel, c'est-à-dire qu'au niveau élémentaire, on se ramène à des comparaisons d'un sous-ensemble par rapport à un ensemble. Ce type de structure renvoie au schéma d'urne, à la base des modèles probabilistes sur lesquels on reviendra dans les paragraphes qui suivent. Cependant, notre approche ne se réduit pas à ce schéma d'urne[45]. D'une part, la constitution des sous-ensembles amène à quitter la linéarité textuelle. Par exemple, le sous-espace des voisinages de la cible est obtenu par réunion de paragraphes dispersés, qui ne se succèdent pas linéairement dans les textes. D'autre part, on ne joue pas sur une unique division de l'espace, mais sur une multiplicité de dichotomies, en fonction des paramètres (domaines, période, etc.). Le jeu sur les paramètres (restriction de l'espace à un sous-espace par fixation d'un paramètre, réunion de sous-espaces associés à un paramètre, etc.) permet d'introduire des combinaisons d' "urnes" ou de sous-espaces. Ces combinaisons ouvrent sur une plus grande complexité spatiale.

1.1.4 Les observables

Pour les empreintes de fréquence, l'objectif est d'établir comment se diffuse la cible lexicale dans les différents domaines et d'en déduire quels domaines peuvent être considérés comme nouveaux domaines d'emploi. Pour observer la diffusion, l'attention se porte sur la cible lexicale. Pour qualifier la diffusion, on considère ce par rapport à quoi la diffusion a lieu. Comme la diffusion est établie relativement aux domaines, l'attention se focalise sur les étiquettes de domaines, autrement dit, sur ce qui qualifie les sous-espaces. D'une certaine façon, les domaines ont un double rôle :

- ils constituent des clés de structuration de l'espace textuel (le sous-corpus du domaine D est constitué de toutes les unités textuelles qui comportent l'étiquette de domaine D) ;
- ils constituent des observables dès lors que l'espace textuel est analysé en fonction de la présence de la cible (on étudie la répartition des sous-corpus des domaines par rapport au sous-corpus des voisinages).

[45] En cela, nous rejoignons les recommandations de (Mayaffre, 2007). Celui-ci invite à dépasser le schéma d'urne pour voir dans les textes des structures plus complexes, témoins d'une topologie textuelle : il propose de privilégier une cartographie textuelle, qui joue sur une structure complexe et qui ne se limite ni à la linéarité, ni aux saillances statistiques.

Les empreintes de fréquence sont celles de la cible, puisqu'on décompte ses occurrences, mais elles sont établies relativement aux domaines : la fréquence de la cible est fonction des domaines. Les empreintes de fréquence établissent ainsi un lien entre cible lexicale et domaines : l'élément particulier que constitue la cible est décrit relativement aux sous-espaces associés aux domaines.

Dans les autres cas, les observables sont les éléments de l'espace associé au corpus : soit les unités lexicales, en l'absence d'annotation sémique ; soit les traits sémantiques issus de l'annotation.

1.2 Espace associé au dictionnaire

1.2.1 Formalisation de la structure

La structure du dictionnaire repose sur une division en entrées. Celles-ci constituent l'unité de division de référence et sont le pendant des unités textuelles qui structurent le corpus (documents pour les empreintes de fréquence, paragraphes pour les cooccurrents). Les entrées sont elles-mêmes subdivisées en définitions. L'entrée est la structure de base qui permet de déterminer le contenu sémantique, tout comme le voisinage est fondamental pour les affinités en corpus.

On dispose de deux types d'observables de nature différente : les mots-vedettes, qui constituent le point de raccord avec les unités lexicales (une unité lexicale du corpus sera directement associée à un mot-vedette) ; les éléments de définition (traits sémantiques), qui sont des traits sémantiques domaniaux ou des traits sémantiques définitoires, autres que les domaines.

La structure associée au dictionnaire a un niveau de complexité moindre que la structure associée au corpus.

Au niveau de la macrostructure, les entrées sont considérées comme indépendantes dans notre cadre expérimental. Elles ne se caractérisent pas par des emboîtements ou par un ordonnancement linéaire, contrairement, par exemple, aux paragraphes constitutifs d'un document. Le dictionnaire est considéré comme l'image d'un état synchronique, il n'y a donc pas de suite ordonnée dans l'espace lexicographique. Ajoutons que les domaines ne servent pas à définir la macrostructure, autrement dit, ils ne sont pas des étiquettes qui définissent des ensembles, mais ils sont considérés comme des éléments.

Au niveau de la microstructure, la complexité reste réduite, car on n'exploite pas toute la richesse du dictionnaire. Le seul paramètre structurel dont on dispose est la subdivision en définitions d'une entrée[46]. À cela peut s'ajouter un lien de dépendance entre domaine et sème, c'est-à-dire l'apparition d'un sème non domanial dans une définition caractérisée par un sème domanial. Rappelons que les bases et outils dont on dispose pour accéder au contenu ont conditionné la définition de la microstructure du dictionnaire. Ainsi, la base SEMEME perd une partie de l'information (exemples, synonymes, etc.), de même que la plateforme SEMY (perte de la structure hiérarchique des définitions).

La structure relative à la cible lexicale est basique : le sous-espace associé est l'entrée correspondante, son complémentaire est le reste du dictionnaire. Cette division structurelle permet d'établir les frontières entre anciennes unités de sens et nouvelles unités de sens, elle

[46] Dans le cadre expérimental, on ne fera pas usage de cette structuration interne, plus délicate car elle nécessite d'une part de mettre en place de la désambiguïsation en corpus, d'autre part, de sélectionner la définition appropriée au sens désambiguïsé. La seule sous-structure utilisée est la réunion des définitions associée à une entrée.

permet d'accéder à la dichotomie enrichissement / reconfiguration. Les frontières ainsi définies sont nettes. Une structure avec des frontières floues serait plus complexe à établir, mais elle offrirait des potentialités plus intéressantes, sur lesquelles nous reviendrons ultérieurement (*cf.* chapitre III.2, 3.2,).

1.2.2 Relations et propriétés

De même que pour le corpus, l'organisation interne des unités structurantes du dictionnaire, c'est-à-dire des entrées, est une organisation ensembliste, de type "sac de" : les éléments d'une entrée constituent des sacs de traits sémantiques. Ils sont traités de façon indifférenciée, seule leur présence compte. Le découpage en entrées définit des relations entre unités de nature différente, les mots-vedettes d'une part, les traits sémantiques d'autre part.

La relation "est sème de" entre un trait sémantique et un mot-vedette correspond à une relation d'appartenance entre une unité de la microstructure (sème) et le sous-ensemble caractéristique du mot-vedette (entrée). Le critère de présence ou absence revient à une caractérisation booléenne de l'appartenance à un ensemble.

La relation "est rattaché au domaine" entre un sème non domanial et un sème domanial est une ***relation de dépendance***, caractérisée par le fait qu'un sème appartient à une définition étiquetée par un certain domaine. Dans nos expériences, nous n'exploiterons que de façon limitée cette relation (chapitre III.1, 4.1.4.c), même si une telle relation mériterait d'être utilisée de façon beaucoup plus large dans un cadre applicatif. L'organisation retenue est donc loin d'être optimale : un sémème se devrait d'être un ensemble structuré. La distinction des définitions est plus prometteuse, mais cela introduit une complexité dans les traitements, que nous n'avons pas mise en œuvre.

Les entrées possèdent une propriété qui s'apparente à de la ***réflexivité***. Normalement, la réflexivité n'est possible que pour une relation définie d'un espace dans lui-même. Ici, les relations sont entre l'espace des mots-vedettes et l'espace des sèmes. Cependant, un mot-vedette possède un analogue sémique, représenté par la même chaîne graphique que celle associée au mot-vedette (par exemple, le sème /toxique/ pour le mot-vedette *toxique*). Cet analogue est considéré comme élément du sémème, d'où une propriété proche de la réflexivité.

1.2.3 Une structure trop simple ?

Le sens codé est délimité par entrée, avec une frontière nette. De ce fait, la macrostructure est morcelée, constituée de sous-ensembles disjoints. On n'introduit pas a priori la possibilité de regrouper des sous-ensembles ou de jouer sur des relations d'inclusion. Plusieurs raisons justifient le choix de ne pas complexifier la macrostructure :

- ***Travail de synthèse du lexicographe :*** le lexicographe cherche à établir une description synthétique du sens de l'unité lexicale, à la fois complète et compacte. Ceci ne signifie pas qu'en-dehors du sous-espace de l'entrée, les éléments ne sont pas définitoires, mais que le sous-espace contient les éléments nécessaires et suffisants à la définition.
- ***Méconnaissance des caractéristiques de la macrostructure :*** le TLFi possède certainement une macrostructure cachée, qui permettrait de dégager des groupes sémantiques cohérents (thématiques par exemple), ou encore des affinités plus ou moins grandes entre entrées. Cependant, cette macrostructure n'a pas encore été dégagée. Des travaux (Loiseau *et al.*, 2010 ; Valette *et al.* 2006) ont eu pour objectif de dégager des formes d'organisation du contenu lexicographique et de fournir des clés pour comprendre la macrostructure du TLFi, mais ces travaux ne sont pas un stade suffisamment avancé pour permettre une réutilisation dans notre cadre expérimental.

Au niveau de la microstructure, c'est-à-dire de la structure interne à une entrée, la structure en "sac de traits" correspond à une exploitation insuffisante de la richesse du contenu, mais qui s'explique triplement :

- ***Absence d'encodage de certaines informations :*** un certain nombre d'informations n'est pas traduit en termes mathématiques ni encodé, comme la subdivision en genre prochain et différences spécifiques ou les formulations qui relèvent du métalangage ;
- ***Intuition de lois de comportement, mais pas de formalisation de référence :*** certains traits, par exemple les unités en tête de définition, seraient plus pertinents pour qualifier le sens que des traits qui apparaissent ultérieurement. Cette idée a été reprise dans les travaux de (Muller *et al.*, 2006:68), qui a initialisé le poids des liens entre mot-vedette et éléments de définition en fonction de la position dans la définition : le 1er élément recevait un poids de 10, le suivant de 9, et ainsi de suite par pas de 1 décroissant, puis stabilisation à 1 pour les éléments restants. Mais la pondération choisie comporte une part d'arbitraire qui n'est pas moindre que la nôtre, et qui, de plus, n'a pas de garantie d'être de même pertinence pour toutes les définitions, comme en témoigne l'exemple de *pollen* et *toxique* :

	pollen	***toxique***
Déf	Poussière très fine (généralement jaune) produite dans les loges des anthères et dont chaque grain microscopique est un utricule ou petit sac membraneux contenant le fluide fécondant	Produit d'origine animale, végétale ou minérale qui provoque l'intoxication, la destruction d'un organisme vivant
10	poussière (subst)	produit (subst.)
9	fin (adj)	origine (subst)
8	jaune (adj)	animal (adj)
7	produire (v)	végétal (adj)
6	loge (subst)	minéral (adj)
5	anthère (subst)	provoquer (v)
4	grain (subst)	intoxication (subst)
3	microscopique (adj)	destruction (subst)
2	être (v)	organisme (subst)
1	utricule (subst), petit (adj), sac (subst), membraneux (adj), contenir (v), fluide (subst), féconder (v)	vivre (v)

Tableau II.2.4 : Poids aux traits sémantiques selon la position dans les définitions de pollen et toxique

Pour les noms, verbes et adjectifs de *pollen,* des éléments sémantiquement pleins et d'usage courant (/poussière/, /fin/, /jaune/) ressortent en tête de définition et se voient affecter un poids considérable, tandis qu'un poids plus faible échoit à des éléments plus techniques, moins pertinents en français standard et plus adaptés à un champ de spécialité (/anthère/, /utricule/ par exemple). A l'inverse, pour *toxique,* le caractère néfaste porté principalement par /intoxication/ et /destruction/ se fait nettement supplanter par des traits peu informatifs, tels que /produit/ et /origine/, qui apparaissent en tête de définition.

Une autre loi à explorer est que l'affinité entre mot-vedette et unité de l'entrée dépend du type de relation lexicale en jeu (synonymie, métonymie, etc.). On reviendra sur cette question au (chapitre III.2, 3.1.3).

- ***Complexité des traitements induits :*** plus la structure est complexe, plus il est difficile de la réutiliser dans des traitements. Par exemple, la subdivision en entrées nécessite d'introduire des procédures de désambiguïsation formelle avant d'effectuer l'annotation sémique, pour tenir compte de l'homonymie encodée dans le dictionnaire. L'introduction de représentations formelles telles que les fonctions lexicales nécessite de mettre en place

de traitements spécifiques, qui rattachent ce formalisme aux emplois en corpus. Ceci implique notamment de retravailler la représentation des textes.

Ajoutons que la différence de statut entre mot-vedette et éléments de définition (analogue des unités lexicales vs sème) est aussi réducteur par rapport à l'ensemble des relations qu'il serait possible de construire. À travers cette distinction, indirectement, on privilégie une représentation du sens atomique, de type contenu / contenant, plutôt qu'une approche différentielle, issue d'une représentation connexionniste et privilégiant les relations de dépendance. Des alternatives existent, mais on souhaite maintenir une distinction de nature entre unités lexicales et traits sémantiques. Par exemple, le sémème pourrait se construire par un lien inversé : les sèmes pourraient être choisis parmi les mots-vedettes vérifiant la relation "est défini par" (la cible serait alors recherchée dans les définitions) au lieu de "est dans la définition de". D'autres types de relations sont évoqués par (Loiseau *et al.*, 2010), du type "cooccurre dans la même définition que" ou "cooccurre dans les exemples". Nos choix expérimentaux sont conditionnés par la distinction unité lexicale / trait sémantique, ainsi que par les contraintes techniques (existence d'outils et de modules permettant d'extraire le type d'unité choisi).

1.3 Espace discret ou continu : données discrètes, espace de modélisation quelconque

Dans une approche numérique du sens, on peut s'interroger sur le caractère discret ou continu des représentations choisies. La question porte sur ce dont on dispose, ce qu'on cherche à obtenir, les modèles applicables et les implications théoriques que cela peut avoir.

Les données dont on dispose en entrée constituent un ensemble discret. La segmentation des textes, la déstructuration des définitions et la description des sous-espaces comme sacs de traits sémantiques ou sacs d'unités lexicales orientent vers une représentation discrète des données en entrée. Le choix du discret apparaît aussi à travers le point de départ usuel des traitements mathématiques en textométrie :

> « l'individu constitue l'unité classique des statisticiens, alors que l'occurrence des mots est celle des lexicomètres » (Lebart, 2004)

Dans notre cadre, l'idée de mots est élargie aux sèmes, mais le constat énoncé reste le même. Les traitements résultent de décomptes d'occurrence. Les décomptes d'occurrence prennent des valeurs entières, l'ensemble de départ est l'ensemble des entiers naturels, c'est-à-dire un ensemble discret.

Les objets recherchés relèvent à la fois du discret et du continu. Les idées d'isotopies et de faisceaux d'isotopies, de contrastes entre fonds et formes sémantiques, de propagation sémique ou de diffusion des néologismes orientent vers du continu. À l'inverse, les notions de molécules sémiques ou encore de sémème comme ensemble de sèmes décrivent les phénomènes sémantiques comme quelque chose de discret. On semble donc à l'intersection du discret et du continu, ou plutôt à cheval sur deux types de représentations, discrète et continue.

Derrière ces considérations se dessine en filigrane la question de la nature du sens : est-elle discrète ou continue ? Nous ne nous prononcerons pas par rapport à cette question, qui dépasse le cadre de nos préoccupations et rejoignons sur ce point (Venant, 2010) : elle découple nature du sens et représentation du sens (la continuité de l'un n'implique pas celle de l'autre, et réciproquement) et invite à se concentrer sur la question de la modélisation.

1.4 D'un espace à l'autre

L'accès au nouveau sens passe par des interactions entre les deux espaces, lexicographique et textuel. La sortie est un sens codé amendé : à l'issue des étapes successives, on se ramène à la ressource lexicographique. En amont, on peut dans un premier temps effectuer des traitements en corpus et se focaliser sur les unités lexicales, qui seront ensuite projetées sur l'espace lexicographique et qui serviront à générer un nouveau sous-espace. On peut également procéder à une projection de sèmes dans l'espace textuel, y effectuer les traitements, puis repositionner les résultats des traitements dans l'espace lexicographique.

Le schéma suivant récapitule les différentes étapes dans le cas où l'on procède à une annotation sémique du corpus avant toute autre opération:

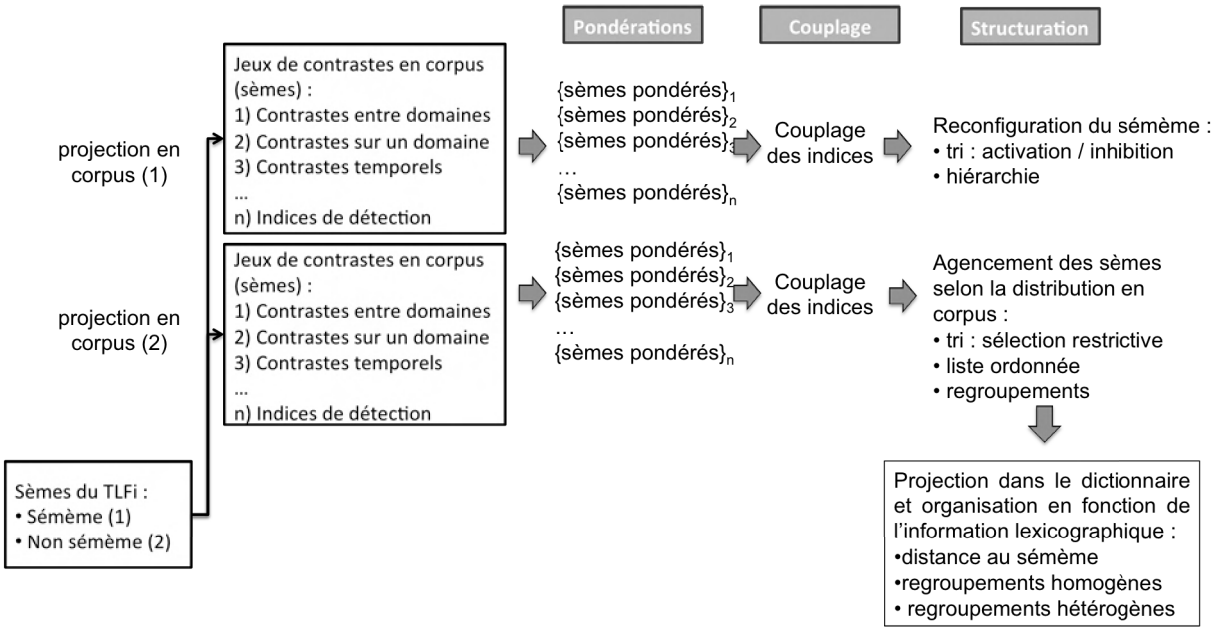


Figure II.2.5 : Étapes de traitement avec une annotation préalable du corpus en traits sémantiques

Dans le cas où l'on n'annote pas le corpus et où on se ramène dans un second temps seulement à la ressource lexicographique, les étapes sont les suivantes :

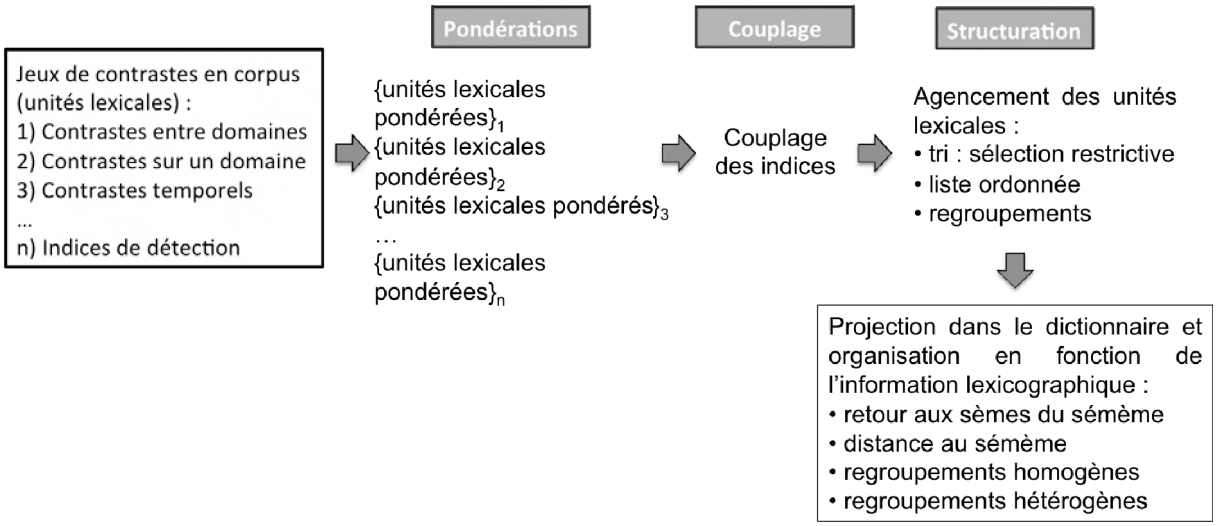


Figure II.2.6 : Étapes de traitement sans annotation préalable du corpus en traits sémantiques

Dans chaque cas, on retrouve les mêmes étapes : affectation d'ensembles de pondérations à travers des jeux de contrastes ; mise en relation des ensembles de pondérations (couplages) ; extraction de structures à partir des pondérations ; projections entre espaces.

2. Extraire des unités saillantes

2.1 Affecter des pondérations

2.1.1 Sélection d'un modèle : distribution et modèle probabiliste

a- Hypothèse de distributionnalité

La recherche d'unités en affinité sémantique avec la cible et l'évaluation quantitative de cette affinité sémantique s'appuient sur les principes énoncés par Firth et Harris :

> « You shall judge a word by the company it keeps. » (Firth, 1957:11)

> « Dans la mesure où la structure formelle (distributionnelle) peut se découvrir dans le discours, elle est d'une certaine manière liée à ce qui est dit. (…) Si on veut parler de la langue comme existant, d'une certaine manière, sur deux plans – le plan de la forme et le plan du sens –, il est du moins possible de dire que les structures des deux plans ne sont pas identiques, bien qu'elles présentent à certains égards des ressemblances. » (Harris et Balagna, 1970)

Autrement dit, des mots cooccurrents, c'est-à-dire partageant les mêmes contextes ou encore qui présentent les mêmes distributions dans l'espace associé aux ressources linguistiques, seront supposés de sens proche. Les affinités sémantiques sont mathématisables : la structure des ressources linguistiques fait pendant à une structure mathématique. Cela ne signifie pas que le sens est de nature mathématique, mais que les distributions mathématiques, issues des distributions en contexte d'unités lexicales, fournissent des descriptions susceptibles de faire ressortir des liens sémantiques. Cette hypothèse a fondé la plupart des approches du contenu informationnel qui utilisent une représentation en sac de mots, notamment les approches textométriques et bon nombre d'approches reposant sur des graphes.

L'hypothèse de distributionnalité sera appliquée au corpus : sa structure formelle permettra de moduler les liens entre unités (lexicales ou sémiques) et de faire émerger des saillances ou des associations privilégiées.

Au niveau du dictionnaire, la structure associée aux sens codés ne permet pas a priori de jouer sur des distributions : les affinités sémantiques entre unités lexicales et sèmes associés sont considérées comme données, on ne cherche pas à les amender. L'analyse sémique peut se faire par projection des sèmes en corpus, puis analyse distributionnelle en corpus. Cependant, on peut également introduire une nouvelle structure dans le dictionnaire (regroupements ou connexions d'entrées lexicales identifiées comme liées au nouveau sens à partir des informations du corpus). Cette nouvelle structure ouvre sur une approche distributionnelle dans l'espace lexicographique.

b- Critère de récurrence

La recherche du nouveau sens s'appuie sur le principe d'isotopie, donc de récurrence sémique. Ce choix implique d'introduire des critères sur la répétition d'un sème, ou éventuellement sur la répétition d'une classe associée à un certain sème.

Au niveau quantitatif, ce critère se traduira par des seuils de fréquence minimale. Certes, dans la théorie de l'information, on considère que les événements rares apportent une grande quantité d'information. Mais une unité rare ne permet d'éclairer que ponctuellement le sens

d'un mot ; c'est la répétition qui permet au nouveau sens de se fixer et de s'implanter. De ce fait, les hapax sémiques seront exclus, non pas parce qu'ils n'apportent aucune information, mais parce qu'ils ne répondent pas à notre perspective.

À défaut de récurrence proprement dite (le trait sémantique est repris à l'identique), une forme de récurrence indirecte est également possible : on considère alors non pas le trait sémantique seul, mais une classe d'éléments à laquelle il est associé. Le critère de récurrence peut s'appliquer à cette classe, à partir de la récurrence de ses éléments. Cette approche plus souple de la récurrence a été adoptée entre autres pour faire face à la dispersion des données sémiques induites.

c- Unités linguistiques et objets mathématiques : des booléens, des réels ou des variables aléatoires ?

De nombreux modèles mathématiques ont été utilisés en textométrie, recherche d'information et autres champs disciplinaires connexes. Les trois modèles classiques en recherche d'information sont le modèle booléen, le modèle vectoriel et le modèle probabiliste (Tannier, 2006).

Le modèle booléen s'appuie sur des représentations binaires (0/1), c'est-à-dire en termes de présence/absence au sein d'ensembles ou sous-ensembles (par exemple, présence d'une forme lexicale dans un document, présence d'un sème dans une entrée de dictionnaire). Il ouvre sur des opérations ensemblistes. On y recourt partiellement dans notre cadre : la représentation des sémèmes issue du dictionnaire s'appuie sur un formalisme booléen.

Le modèle vectoriel décrit les phénomènes à partir des vecteurs réels, dans un espace dont les dimensions sont souvent les documents. Ce modèle (Vector Space Model) est issu des travaux de (Salton *et al.*, 1975). Il a notamment servi dans le cadre d'appariement de documents et de requêtes. Les coefficients dépendent du nombre d'occurrences des éléments. Chaque élément est considéré comme un vecteur dans un espace multidimensionnel (espace des documents dans le cadre évoqué précédemment) et implique généralement un positionnement dans un espace euclidien. Les métriques appliquées ne sont pas sans poser quelques problèmes : en sémantique, les propriétés des mesures, notamment des distances, sont loin d'être respectées (par exemple, l'inégalité triangulaire n'est pas vérifiée, ce qui a amené à introduire un certain nombre de mesures de *dissimilarités*, qui se distinguent des *distances* par le fait qu'elles ne respectent pas l'inégalité triangulaire).

Le modèle probabiliste revient à considérer les occurrences comme des événements (réalisations de variables aléatoires, celles-ci étant associées aux unités lexicales, aux sèmes, aux documents ou tout autre objet linguistique). Dans notre cadre, les unités qui tiennent lieu d'individus ou d'observations sont les unités lexicales ou les sèmes. Les périodes de temps ou encore les domaines peuvent se concevoir comme des modalités des variables associées aux unités lexicales ou sémiques. Les modèles probabilistes introduisent la notion d'incertitude, absente du cadre théorique des modèles vectoriels.

Les différents modèles ont des bases théoriques distinctes, mais ils utilisent plusieurs outils et méthodes entre lesquels on peut établir formellement un parallèle. Les résultats quantitatifs peuvent donc être relativement proches, mais leur interprétation est susceptible de différer.

2.1.2 Validité des modèles probabilistes

a- Détournement théorique et légitimité empirique des modèles probabilistes

Un modèle aléatoire pour un phénomène non aléatoire. Dans les modèles statistiques, les données linguistiques sont considérées comme des échantillons aléatoires. Les variables (unités lexicales, sèmes ou domaines dans notre cas) sont supposées indépendantes.

L'hypothèse d'indépendance en statistique se traduit comme suit :

P(X|Y)=P(X) ou P(Y)=0, où X et Y sont des variables aléatoires

Dans notre cadre, ceci pourrait par exemple se traduire par le fait que l'occurrence d'une unité lexicale (resp. sème) x ne dépend pas de l'occurrence des autres unités lexicales y (resp. sèmes) : quelles que soient les unités présentes dans son voisinage, x a autant de chance d'apparaître.

L'hypothèse d'indépendance (hypothèse nulle de bon nombre de tests statistiques) n'est jamais vraie : par essence, le langage n'est pas un phénomène aléatoire. La présence d'une unité lexicale conditionne celle des unités voisines, aussi bien pour des raisons de combinatoire sémantico-syntaxique que pour des raisons de cohérence thématique. (Brunet, 1984) invite à relativiser les effets dus aux dépendances syntaxiques : répétée sur des segments nombreux et larges, la déstructuration des dépendances syntaxiques (génération de structures de type 'sac de') ouvre sur des combinaisons multiples d'unités et crée un brassage. On peut supposer que, en regroupant l'ensemble des occurrences, il y a relâchement des liens syntactico-sémantiques et plus d'indépendance a priori que pour une occurrence isolée. En revanche, il est plus délicat d'écarter les liens stylistiques et thématiques entre unités (Brunet, 1984:256). Ces considérations touchent la question des isotopies : le principe d'isotopies locales et de formes sémantiques (tendance pour des unités associées à une thématique à se concentrer dans une zone, avec un effet grappe) va à l'encontre de l'hypothèse d'indépendance. Dans la mesure où nous considérons les isotopies comme fondamentales, cela revient à rejeter la validité théorique du modèle d'urne. En effet, dans le modèle d'urne, les boules qui constituent l'urne sont indépendantes, indifférenciées au toucher et par toute autre propriété susceptible de distinguer les boules dans un tirage en aveugle ; dans une urne respectant le principe d'isotopies, les propriétés des boules seraient tout autre, avec par exemple des boules reliées en chapelets, ou encore des boules aimantées, formant des amas par répulsion ou attraction. Dans la même veine, (Lafon, 1981) a mis en évidence une distribution des éléments du discours en *rafales*, c'est-à-dire de façon non régulière, avec des concentrations d'occurrences localisées et irrégulièrement espacées. Ce phénomène de rafales témoigne de l'invalidité d'un modèle de distribution aléatoire. Dès lors qu'on dispose de suffisamment de données ce qui est souvent le cas en linguistique de corpus, il y aura rejet systématique de l'hypothèse nulle. Cette même idée apparaît dans la littérature anglo-américaine sous le nom de *clumpiness* (littéralement, la tendance à apparaître en bouquets ou en touffes) qui indique qu'un mot a tendance à ne pas apparaître isolément, mais à être repris (Katz, 1996 ; Kilgarriff, 2001). À travers des expériences, (Church et Gale, 2000) mesurent cet effet de répétition et cherchent à le préciser. Ils constatent d'une part que cet effet est très marqué, d'autre part que les mots affectés par la *clumpiness* sont principalement des mots très informatifs (les *content words*, c'est-à-dire des mots sémantiquement pleins par opposition aux mots-outils, vecteurs de thématiques ou assimilables à des mots-clés)[47].

[47] Dans leurs expériences, (Church and Gale, 2000) estiment l'effet de répétition à partir d'une partition obtenue par fractionnement d'unités textuelles : un sous-corpus est constitué de demi-documents (*history corpus*), l'autre des moitiés de documents complémentaires (*test corpus*). L'idée est que si un mot apparaît dans une unité textuelle, il aura tendance à être repris dans cette unité textuelle. Les auteurs montrent que la probabilité de

Un pis-aller ? Les modèles probabilistes construits sur les lois de la statistique classique sont des modèles qui ont été analysés théoriquement, éprouvés empiriquement et rendus largement accessibles à travers un certain nombre de supports techniques. On dispose donc d'un regard critique et de moyens de mise en œuvre (par exemple, à travers des logiciels tels que Hyperbase (Brunet, 2011), Lexico3 (Salem *et al.*, 2003), Weblex (Heiden, 2000) ou encore TXM (Equipe TXM, 2011). À notre connaissance, il n'existe pas de modèle vierge de toute critique, aussi bien en théorie qu'en pratique. On rejoint en cela l'idée de (Muller, 1964), reprise notamment par (Labbé et Labbé, 2001), selon laquelle, à défaut de modèle parfaitement fondé sur le plan théorique et opérationnel, on se contente de ce dont on dispose à l'heure actuelle. Des alternatives aux modèles probabilistes existent, mais, par-delà leur intérêt théorique, ils se heurtent à des difficultés de mise en œuvre pratique et leur supériorité par rapport aux modèles probabilistes n'a pas été démontrée. Si cet argument laisse un sentiment de semi-échec et d'insatisfaction, il a le mérite de mettre en avant un principe de réalité et le fait que la communauté linguistique a conscience des limites des outils utilisés.

Des succès empiriques. La validité théorique du modèle probabiliste est contestable, mais ces modèles ont permis d'obtenir des résultats probants. Lorsque (Kilgarriff, 2005:264) propose un regard critique sur la validité des présupposés théoriques des modèles probabilistes, il défend le recours aux méthodes probabilistes malgré leurs failles théoriques selon l'argument que ces méthodes ont permis de grands progrès à travers des approches linguistiques empiriques. De même, pour (Pincemin, 1999), l'adéquation avec la réalité textuelle et l'efficacité expérimentale autorisent à sortir des limites de validité théorique d'un modèle : outrepasser le cadre théorique est possible dès lors que les formules appliquées sont articulées à une représentation des textes, qu'elles s'avèrent valides en pratique, que l'utilisateur est conscient des limites théoriques et que les fondements théoriques à l'origine des formules sont considérés comme des hypothèses empiriques[48].

Un détournement approuvé. Les modèles statistiques, en particulier les tests d'hypothèse qui sous-tendent la plupart des méthodes communément utilisées, peuvent être détournés de leur vocation initiale et être articulés à une perspective relativement proche. La question n'est pas de savoir ***si*** on peut rejeter l'hypothèse nulle (on sait qu'il n'y a pas indépendance), mais ***ce qui*** contribue à s'écarter le plus d'une distribution aléatoire (1), et ***dans quelle mesure*** (2) :

- (1) Comme nous l'avons déjà évoqué, le langage n'est pas un phénomène aléatoire, notamment du fait de l'existence de thèmes ou d'isotopies locales. Plus largement, l'hypothèse distributionnelle de Harris revient à dire que les unités sémantiquement liées le sont aussi distributionnellement, donc qu'elles tendent à éloigner d'un modèle aléatoire.

retrouver un mot ayant une probabilité p d'occurrence se rapproche de p/2 (surtout pour les *content words*), alors que, sous hypothèse d'indépendance, cette probabilité devrait être de p^2, soit une valeur nettement inférieure étant donné que $p << 1$.

[48] « Ce qui prime ici, dans la recherche de formules pour décrire les textes et les unités linguistiques qu'ils comportent, ce n'est pas l'exactitude d'un modèle, au sens où il est établi et démontré suivant un raisonnement formel. Ce qui importe avant tout, c'est l'adéquation des formules à la représentation des phénomènes textuels. Pour les formules heuristiques, proposées comme expérimentalement efficaces, le but est de les expliquer (…). Pour les formules issues d'un modèle théorique, il convient d'expliciter les hypothèses sous-jacentes, notamment pour vérifier dans quelle mesure elles concordent avec la représentation que l'on se fait de la réalité textuelle et linguistique. (…) Une application de la formule hors du cadre prévu ne peut plus se référer au modèle théorique initial ; en revanche, elle peut encore être analysée d'un point de vue heuristique. » (Pincemin, 1999:429-430)

On peut donc supposer que ce qui amène à rejeter l'hypothèse d'indépendance correspond à des éléments linguistiques en affinité thématique, ou plus largement en affinité sémantique, et de façon encore plus générale en affinité linguistique.

- (2) À travers le choix d'un seuil, les tests statistiques nous disent qu'on peut rejeter l'hypothèse nulle avec un certain niveau de confiance. Le raisonnement suivant sous-tend l'usage des statistiques tel qu'il est souvent pratiqué : plus le niveau de confiance est élevé pour rejeter l'hypothèse d'indépendance, plus le lien de dépendance est supposé important. Un haut niveau de confiance sera considéré comme le reflet d'une forte dépendance. On peut donc supposer que le degré d'affinité distributionnelle (donc sémantique, selon l'hypothèse harrissienne) peut s'obtenir à partir du niveau de confiance. Il s'agit, certes, d'un abus théorique : le niveau de confiance qui permet de rejeter l'hypothèse nulle n'est pas une quantification du degré d'affinité sémantique. Certains auteurs proposent d'ailleurs des alternatives aux tests statistiques, qui sont des méthodes statistiques directement destinées à évaluer le degré d'association, au lieu d'assimiler le degré de certitude de l'existence d'une association à un degré d'association. Ces méthodes sont des estimations de l'*effet de taille* (*effect size* ; Cohen, 2003:5) . Elles mesurent l'importance du lien entre deux variables en retranchant les effets dus à la taille de l'échantillon (c'est-à-dire du corpus ou du sous-corpus) intégrés dans les valeurs des tests d'hypothèse. Taille d'effet et tests d'hypothèse sont liés par la relation suivante :

  test d'hypothèse = effet de taille x taille de l'échantillon (Rosenthal, 1994:232).

  Une autre alternative aux tests statistiques est celle des mesures d'entropie au sens de la théorie de l'information. Il nous paraît préférable de ne pas exclure les méthodes fondées sur les tests statistiques, dont les résultats ont connu des validations expérimentales et qui sont largement répandues dans la communauté.

b- Limites quantitatives : taille des données et événements rares

Les modèles statistiques sont adaptés pour des grandes quantités de données, mais on peut s'interroger sur leur limite de validité à deux points de vue : en fonction de la quantité de données, en particulier lorsque celle-ci tend à se réduire, et en fonction du type d'événements observé, en particulier quand il s'agit d'événements rares.

b1) Influence de la taille sur les modèles

Plus la quantité de données est importante, plus le modèle statistique est fiable. La taille des données est celle du corpus d'étude, c'est-à-dire le volume total des données disponibles. Cependant, d'autres paramètres que la taille du corpus sont susceptibles de jouer sur les limites de validité du modèle. La plupart des modèles statistiques s'appuient sur une table de contingence où interviennent quatre paramètres :

- (1) la taille totale du corpus, qui correspond à la taille de la population ;
- (2) la taille du sous-corpus, qui correspond à la taille de l'échantillon ;
- (3) le nombre total d'occurrences de l'unité observée dans le corpus, qui correspond au nombre de succès dans la population ;
- (4) son nombre d'occurrences dans le sous-corpus, qui correspond au nombre de succès dans l'échantillon.

Si les paramètres (2) et (3) (taille du sous-corpus et nombre total d'occurrences de l'unité observée) prennent des valeurs faibles, ils amènent à se placer dans des conditions aux limites, celles des événements rares et des petits échantillons.

b2) Linguistique de corpus et émergence d'un nouveau sens : quelle quantité de données ?

Même si on dispose initialement d'un grand corpus, le phénomène ciblé, à savoir l'émergence d'un nouveau sens, peut amener à des observations particulières, qui demandent de réduire la taille des données. Le jeu de contrastes est un jeu sur des paramètres pour focaliser l'attention sur un sous-corpus. Ceci implique d'observer un corpus qui n'est qu'une fraction du corpus initial : on se limite à un domaine donné, à une période, à une réunion de voisinages (paragraphes). De plus, on souhaite détecter et caractériser le plus tôt possible l'émergence du nouveau sens, donc on cible un nombre d'occurrences des nouveaux emplois d'autant plus réduit. Le choix d'un certain type de contraste et d'une observation de la cible sur un espace réduit se traduit soit par une taille de population réduite (taille du corpus pour le type de contraste choisi), soit par un échantillon réduit (taille du sous-corpus des voisinages propres au nouveau sens). Les phénomènes que nous souhaitons observer nous amènent à nous situer hors du cadre où l'efficacité des modèles statistiques est incontestable[49], et là où des erreurs ou des imprécisions du modèle sont susceptibles de se manifester.

Actuellement, on peut disposer de corpus de taille considérable. Plusieurs projets actuels visent à constituer des corpus de grande ampleur, tels que le corpus français frWaC (Ferraresi *et al.*, 2010)[50] ou encore le corpus de référence de l'allemand DeReKo (Kupietz *et al.*, 2010)). Il ne nous semble cependant pas indispensable de disposer de corpus de taille démesurée, de plusieurs milliards de mots. Par contre, il est pertinent de disposer de suffisamment d'informations relatives à la cible lexicale : même si le corpus est plus réduit dans son ensemble, les contextes associés aux nouveaux emplois de la cible doivent être suffisamment nombreux.

Dans notre cadre expérimental, les corpus utilisés resteront de petite taille (un million de mots ou taille inférieure). Les tests réalisés se veulent des explorations de pistes rudimentaires, à petite échelle. Les résultats de ces tests sont destinés à impulser des projets plus ambitieux sur de plus grands corpus, ou à l'inverse à éviter de s'engager dans des voies hasardeuses.

b3) Événements rares ou petits sous-corpus : un certain nombre de modèles expliquent mal les faibles fréquences

Dans des corpus de taille moyenne, un certain nombre de modèles statistiques n'expliquent pas bien les événements rares ou les événements observés relativement à de petits sous-corpus (Dunning, 1993). Ainsi, pour les modèles construits sur des tests d'hypothèse asymptotiques, notamment ceux qui reposent sur une hypothèse de distribution normale (tels que le t-score ou l'écart-réduit *z*) , la significativité a tendance à être surestimée pour les basses fréquences : la statistique des tests d'hypothèse converge asymptotiquement vers une loi de probabilité qui

[49] (Dunning, 1993) distingue trois catégories d'approches statistiques : des mesures sur de très grands volumes de données, des mesures sur des volumes relativement petits, pas de mesures. Il cite comme approche de référence sur de très grands volumes de données une étude d'IBM, effectuée sur un corpus de près d'un milliard de mots, pour laquelle il estime que la qualité de leurs résultats statistiques est incontestable. Comme approche sur corpus de taille réduite, il mentionne une étude sur les associations prédicat-argument réalisée sur un corpus de 15 millions de mots. Il estime que, sur de tels volumes de données, les méthodes statistiques posent problème.

[50] Le corpus frWaC est un corpus de français tiré du Web. Il s'agit d'un corpus de grande taille (1,6 milliard de mots), constitué en parallèle d'un corpus de même nature en anglais et qui a été construit afin d'obtenir des corpus de référence parallèles. Les documents constitutifs ont été obtenus à partir de requêtes effectuées sur 1800 paires de mots à partir d'un moteur de recherche. Les mots qui ont servi aux requêtes proviennent des mots de fréquence moyenne du *Monde Diplomatique* (1980-2000) ou de mots d'un français fondamental. Le corpus a été nettoyé et étiqueté, mais son contenu, encore mal connu, doit encore faire l'objet d'investigations et de validations.

sert de modèle de référence pour calculer l'affinité sémantique ; or lorsqu'on sort du cadre asymptotique, ce qui est précisément le cas des événements rares, la convergence est moins bonne, les probabilités estimées à partir des tests sont supérieures aux probabilités exactes, donc des événements qui ont peu de chance de se produire paraîtront d'autant plus remarquables.

À l'inverse, les modèles issus des tests exacts, dont le test de Fisher, expliquent mieux les événements rares ou liés à de petits sous-corpus. Cependant, d'autres phénomènes peuvent survenir. Pour de faibles fréquences, l'échelle des valeurs de significativité sera plus tassée. De ce fait, une absence d'occurrence pourra être considérée non pas comme négativement significative (c'est-à-dire comme une sous-représentation), mais comme non significative. Le nombre d'événements sera insuffisant pour qu'une absence d'occurrence soit considérée comme remarquable. De même, le modèle hypergéométrique, à la base du test de Fisher, n'est pas symétrique : le nombre d'événements qui apparaissent comme sous-représentés est moins important que le nombre d'événements surreprésentés. On tend vers de plus en plus de symétrie dès que le sous-corpus considéré ou la fréquence en corpus sont grands, par convergence de la loi hypergéométrique vers la loi normale. À l'inverse, la dissymétrie est d'autant plus marquée pour les événements rares. Une absence d'occurrence considérée comme non significative ou un déséquilibre entre valeurs positivement et négativement significatives ne nous paraît pas un défaut du modèle en soi. Il s'agit plus d'un comportement du modèle qu'il convient de garder à l'esprit pour ne pas aboutir à des interprétations abusives, comme le soulignent (Labbé et Labbé, 2001:6).

2.1.3 Indices : interprétation en terme d'affinité sémantique, comparaison et indice retenu

a- Des indices pour quantifier l'affinité sémantique ou la saillance dans le voisinage de la cible

Les pondérations peuvent être affectées à deux types de relations : avant tout, à des relations lexicales ou sémiques relatives à la cible lexicale ; ensuite, à des relations entre unités lexicales ou sémiques distinctes de la cible lexicale.

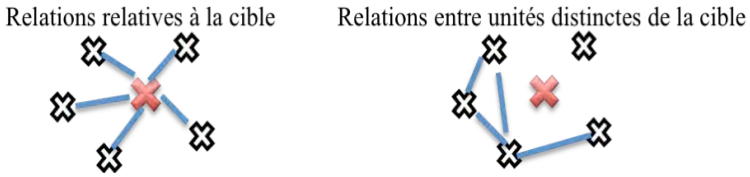


Figure II.2.7 : Définition des relations sémantiques relativement ou non à une cible lexicale

Pondérations relatives à la cible. Pour extraire l'information pertinente du voisinage de la cible, on peut mobiliser deux notions : on cherche d'une part à établir des affinités entre unités linguistiques (cible – autres unités lexicales ; cible – sèmes ; sèmes du sémème – sèmes) ; d'autre part, la cible lexicale est étroitement associée à un espace (les cotextes ou l'entrée de dictionnaire) et l'étude des autres unités est relative à cet espace. Les pondérations relatives à la cible peuvent donc refléter soit le degré d'affinité sémantique avec la cible, soit la saillance dans le sous-espace propre à la cible. Ces deux perspectives se rejoignent, sans être totalement identiques : dans le premier cas, on se positionne par rapport à une unité (la cible), dans le second cas, par rapport à un espace.

Ces deux perspectives rejoignent deux notions communes en linguistique de corpus : le score d'association et la *keyness* (c'est-à-dire le degré d'importance d'une unité en tant que mot-clé, ou plutôt unité-clé). Les scores d'association permettent de quantifier la force du lien entre deux unités, ils témoignent d'une affinité sémantique entre unités. La *keyness* permet de

quantifier dans quelle mesure l'unité linguistique est caractéristique d'une unité textuelle (document par exemple), elle est le reflet d'une saillance d'une unité par rapport à un espace[51].

Ce qu'on cherche à établir, à savoir un lien entre unités linguistiques, amène à privilégier l'association, mais l'asymétrie de rôle des unités (la cible lexicale a un rôle de pôle, les autres unités gravitent en quelque sorte autour de ce pôle) oriente vers la *keyness*.

Pondérations entre unités lexicales ou sémiques distinctes de la cible. Pour ce type de pondération, la notion appropriée est celle d'association, puisque les unités comparées ont des rôles symétriques. Ce type de pondération peut servir dans un second temps pour constituer des regroupements parmi les unités lexicales ou sémiques caractéristiques de la cible lexicale.

On abordera donc la question des pondérations à travers l'étude de scores d'association et de scores associés aux mots-clés (*keyness*).

b- Base de calcul : structurer les occurrences sous forme de table de contingence

Les techniques mathématiques applicables pour les scores d'association et de *keyness* sont similaires, les changements interviennent au niveau des objets auxquels s'appliquent les techniques. Les pondérations sont généralement calculées à partir de la même présentation initiale des données, sous forme d'une table contingence.

La table de contingence découle assez immédiatement d'une certaine structuration de l'espace textuel en fonction de l'unité et de l'espace observés (pour la *keyness*), ou en fonction d'un couple d'unités observées (pour l'association). La structuration amène à dégager quatre sous-ensembles, comme cela apparaît dans le schéma suivant :

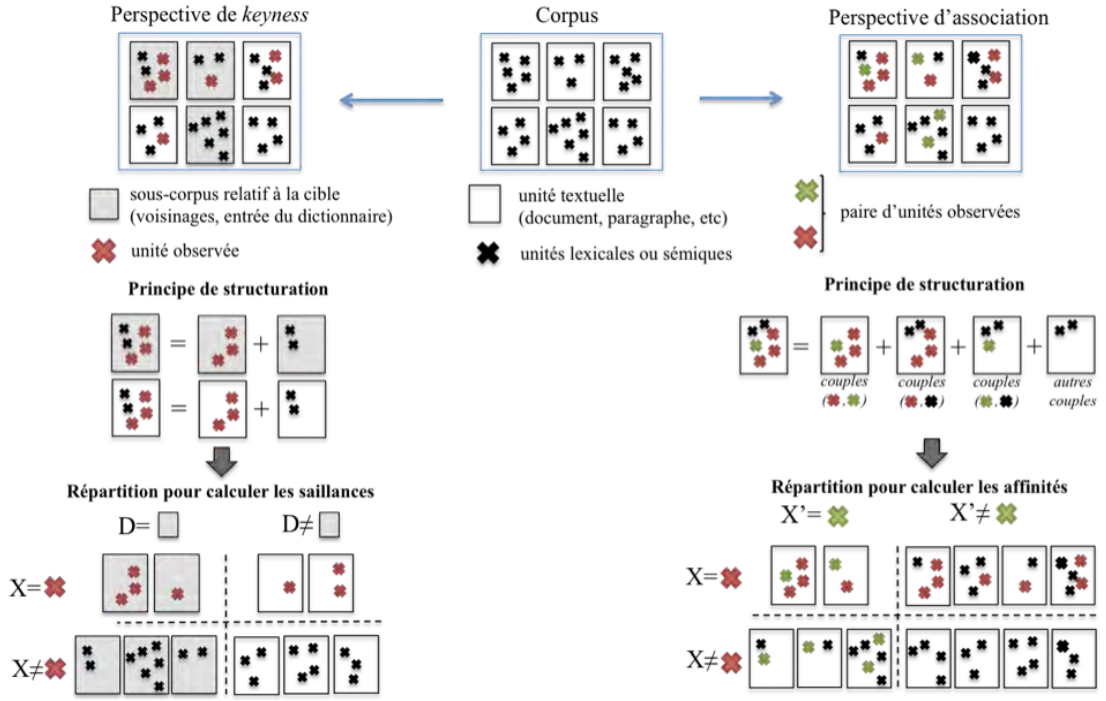


Figure II.2.8 : Structuration du corpus à l'origine des tables de contingence pour la keyness et l'association

[51] Le terme exact en français pour cette notion serait celui de *spécificités*. Nous préférons le terme anglais *keyness*, car nous parlerons par la suite d'un indice précis, issu de la loi hypergéométrique, qualifié de *spécificités de Lafon* et, en abrégé, de *spécificités*. Le terme anglais permet ainsi de limiter la confusion. Il est notamment employé par (Scott et Tribble, 2006).

La partition de l'espace textuel se fait en fonction de variables aléatoires associées à un sous-corpus ou à une unité lexicale ou sémique. On notera D la variable aléatoire associée à un sous-corpus (ensemble de documents, paragraphes, etc.) ; X et X' les deux variables aléatoires associées à des unités élémentaires (lexicales ou sémiques).

À chaque sous-espace est affectée une valeur, qui sera une entrée de la table de contingence. Les entrées de la table de contingence sont des nombres d'occurrences ou de cooccurrences, qu'on désigne généralement comme les *valeurs observées*, et notées O_{ij} (O_{11}, O_{12}, O_{21}, O_{22}). Le nombre d'occurrences d'une unité est relatif soit à un espace d'appartenance (le décompte est celui des occurrences dans cet espace), soit à une unité de même nature (il s'agit alors d'un décompte de cooccurrences).

La table de contingence réalisée à partir des occurrences observées a la configuration suivante :

- pour la *keyness*

	D=d	D≠d	
X=x	O_{11}	O_{12}	$O_{1.}=O_{11}+O_{12}$
X≠x	O_{21}	O_{22}	$O_{2.}=O_{21}+O_{22}$
	$O_{.1}=O_{11}+O_{21}$	$O_{2.}=O_{12}+O_{22}$	N

- pour les scores d'association

	X'=x'	X'≠x'	
X=x	O_{11}	O_{12}	$O_{1.}=O_{11}+O_{12}$
X≠x	O_{21}	O_{22}	$O_{2.}=O_{21}+O_{22}$
	$O_{.1}=O_{11}+O_{21}$	$O_{2.}=O_{12}+O_{22}$	N

Les deux tables se ressemblent, mais la nature de ce qui est décompté diffère ($O_{11}+O_{21}$ représente dans le premier cas la taille du document, c'est-à-dire le nombre d'éléments qu'il contient ; dans le second cas, il s'agit du nombre d'occurrences total de x' dans le corpus). Les tables de contingence contiennent des valeurs observées, c'est-à-dire obtenues à partir des occurrences effectives en corpus. Pour le calcul d'un certain nombre d'indices, les valeurs observées sont confrontées à des valeurs théoriques (ou attendues) E_{ij}, dont les valeurs sont obtenues sous hypothèse d'indépendance : $E_{ij} = \frac{O_{i\bullet} \times O_{\bullet j}}{N}$. À chaque valeur observée correspond une valeur théorique, ce qui permet d'obtenir les tables de valeurs attendues suivantes :

- pour la *keyness*

	D=d	D≠d
X=x	E_{11}	E_{12}
X≠x	E_{21}	E_{22}

- pour l'association

	X'=x'	X'≠x'
X=x	E_{11}	E_{12}
X≠x	E_{21}	E_{22}

c- Une multitude d'indices à disposition

Il existe un grand nombre d'indices (scores d'association ou de *keyness*). (Pecina et Schlesinger, 2006) recensent ainsi 82 mesures d'association utilisées pour les collocations de bigrams, qui sont ensuite comparées en termes de précision moyenne. Leur perspective ne rejoint pas tout à fait la nôtre : les mesures d'association attribuent un rôle symétrique aux deux unités considérées et le palier contextuel est extrêmement réduit (fenêtre de deux mots, puisqu'il s'agit de bigrams). Cependant, l'ensemble des indices est potentiellement applicable à notre cadre d'étude.

Les indices relèvent de familles différentes : ils se distinguent par leurs fondements théoriques, les champs disciplinaires où ils sont appliqués, leur degré de popularité. Deux

familles fondamentales émergent de façon récurrente (*cf.* (Manning et Schütze, 2003:162-183), (Evert, 2005:75-91), (Pecina et Schlesinger, 2006)) : les indices issus de tests d'hypothèse et ceux associés à la notion d'entropie et dérivés de la théorie de l'information.

Les indices associés aux tests d'hypothèse fournissent des mesures qui évaluent le degré de certitude d'une non-association entre variables. Largement diffusés en sciences sociales, ils comportent deux classes principales : les tests d'hypothèse asymptotiques, qui découlent d'une hypothèse de distribution normale (t-score, écart-réduit, χ^2, rapport de log-vraisemblance (*log-likelihood ratio*)) ; les tests d'hypothèse exacts (test exact de Fisher, Poisson, modèle binomial).

L'indice principal associé à la notion d'entropie est l'*information mutuelle* (Shannon, 1948 ; Church et Hanks, 1989 pour une application aux données textuelles de référence dans la communauté anglophone). L'information mutuelle mesure l'information que chaque variable fournit sur l'autre variable. Autrement dit, l'information mutuelle représente la quantité d'information qu'apporte une unité sur son cooccurrent (ou un sous-corpus sur l'occurrence d'une unité).

De nombreux autres indices existent. Parmi ceux-ci, on compte ceux qui mesurent l'*effect size*, parmi lesquels on compte le coefficient de Jaccard, le coefficient de Dice ou la moyenne harmonique, et qui font intervenir des rapports ou différences entre valeurs observées et valeurs attendues et dont les résultats ne sont normalement pas fonction de la fréquence absolue. On pourra trouver une description plus détaillée chez (Manning et Schütze, 2003:299) et (Evert, 2005:84-88). Des mesures heuristiques sont également utilisées. Celles-ci reprennent d'autres mesures, en les combinant ou en y ajoutant des paramètres (ajout d'un exposant dans l'information mutuelle par exemple, qui est un paramètre réglable ; *cf.* Daille 1994).

Nous proposons de nous limiter à un petit nombre d'indices, répandus dans les études textométriques, et dont des supports logiciels assurent une large diffusion.

d- Comparaison de quelques indices

Nous proposons d'étudier les indices suivants :

- l'écart-réduit z ;
- le t-score ;
- l'indice du χ^2 ;
- le G^2 (ou log-likelihood ratio, ou encore rapport de log-vraisemblance) ;
- les spécificités de Lafon ;
- l'information mutuelle.

Les quatre premiers indices appartiennent à la famille des tests d'hypothèse asymptotiques. Les spécificités de Lafon, qui sont étroitement associées au test de Fisher et proviennent de la loi hypergéométrique, font partie des tests exacts[52]. L'information mutuelle est l'indice-type issu de la théorie de l'information.

[52] Point sur le vocabulaire : ***spécificités, keyness, spécificités de Lafon*** et ***test exact de Fisher***
Pour dénommer certaines mesures, nous nous sommes heurtés à des ambiguïtés terminologiques liées d'une part à des hiatus entre littérature française et littérature anglophone, d'autre part, à des différences entre des notions précises et leur appellation dans l'usage.
Les ***spécificités de Lafon*** désignent un indice construit sur la loi hypergéométrique. Les ***spécificités*** stricto sensu et les ***spécificités de Lafon*** (ou plus exactement l'indice issu de la loi hypergéométrique) ne sont pas des notions identiques : les spécificités de Lafon sont un cas particulier des spécificités stricto sensu, dont l'acception est

Ces indices peuvent être qualifiés d'indices à large diffusion : au-delà de leur intérêt théorique, ils constituent des outils classiques en linguistique, récurrents dans les études lexicométriques, et leur présence dans des travaux futurs reste assurée, car leur intégration dans différents logiciels de textométrie favorise leur diffusion. Le tableau qui suit présente les indices utilisés dans un ensemble de logiciels de textométrie ou d'exploration lexicale de corpus : Alceste, AntConc, CQPWeb, XAIRA, WordSmith, Dtm-Vic, SATO, Hyperbase, Lexico3, TXM[53].

	Alceste	Antconc	CQPWeb	XAIRA	WordSmith	Dtm-Vic	SATO	Hyperbase	Lexico 3	TXM
écart-réduit (z-score)			+	+	+		+	+		
t-score		+								
χ^2	+				+	+	+	+	+	+
G^2 (log-likelihood ratio)		+	+		+					
Spécificités						+		+	+	+
Information mutuelle		+	+	+	+					

Tableau II.2.9 : Indices intégrés à différents logiciels de textométrie

Les indices se distinguent à plusieurs niveaux, qui dépendent les uns des autres et qu'on détaillera successivement, après avoir donné au lecteur une vue d'ensemble. Sur le plan théorique, il y a des différences au niveau des formules qui les définissent ; au niveau des modèles théoriques qui sous-tendent ces formules ; au niveau des limites d'adéquation entre formules et modèles théoriques. Sur le plan applicatif, les différences théoriques ont des

---

beaucoup plus large. Les ***spécificités*** au sens large correspondent au degré de saillance d'une unité par rapport à un sous-corpus, sans qu'il y ait dépendance à une loi précise a priori. Des logiciels de textométrie tels que Lexico3 ou Hyperbase utilisent un indice fondé sur la loi hypergéométrique, provenant des ***spécificités de Lafon*** et souvent désigné de façon abrégée par "spécificités". La popularité de ces logiciels a favorisé la diffusion du terme *spécificités* au sein de la communauté francophone d'analyse des données textuelles pour désigner l'indice de significativité issu de la loi hypergéométrique. Pour limiter la confusion, nous éviterons de parler de *spécificités* pour le sens large, *i.e.* pour qualifier le degré de saillance par rapport à un sous-corpus, et nous privilégierons le terme anglais ***keyness*****,** qui renvoie à la même notion. Nous réserverons le terme *spécificités* (*spécificités* tout court ou *spécificités de Lafon*) à l'indice issu de la loi hypergéométrique.
Par ailleurs, les ***spécificités de Lafon*** entretiennent un rapport analogue avec le ***test de Fisher*** à celui que l'indice du χ^2 entretient avec le test du χ^2. Dans la littérature anglophone que nous avons parcourue, il est plutôt fait référence à Fisher pour désigner cet indice (test de Fisher, mesure de Fisher (Evert, 2000, p. 80 par ex)).
Pour rester cohérent avec le terme en usage en français, nous emploierons ***spécificités*** plutôt que ***mesure de Fisher*** pour parler de l'indice issu de la loi hypergéométrique, en précisant éventuellement l'existence d'un lien avec le test de Fisher pour maintenir une passerelle avec la littérature anglophone.

[53] Logiciels accessibles via les pages suivantes :
- Alceste (Reinert, 2002) : http://www.image-zafar.com/index_alceste.htm ;
- AntConc (Anthony, 2005) : http://www.antlab.sci.waseda.ac.jp/software.html ;
- CQPWeb (Hardie, 2009) : http://cqpweb.lancs.ac.uk/ ;
- XAIRA : http://xaira.sourceforge.net/ ;
- WordSmith (Scott, 2001) : http://www.lexically.net/wordsmith/ ;
- Dtm-Vic (Lebart et Piron, 2011) : http://www.dtmvic.com/ ;
- SATO (Daoust, 2011) : http://www.ling.uqam.ca/sato/ ;
- Hyperbase (Brunet, 2011) : http://ancilla.unice.fr/~brunet/pub/hyperbase.html ;
- Lexico3 (Salem *et al.*, 2003) : http://www.tal.univ-paris3.fr/lexico/ ;
- TXM (Equipe TXM, 2011) : http://textometrie.ens-lyon.fr/.

répercussions d'importance variable au niveau des fréquences (notamment les faibles fréquences, sur lesquelles on s'attardera) et au niveau des résultats d'ensemble.

Les formules associées aux différents indices sont présentées ci-dessous. Il existe des formules plus générales associées à certains indices (par exemple, l'information mutuelle peut s'exprimer de façon plus générale à l'aide de probabilités *p(x,y), p(x), p(y)* obtenues par d'autres moyens), mais d'une part, celles qui sont présentées sont couramment utilisées, d'autre part, elles sont toutes en adéquation avec la table de contingence précédente, et donc aisément comparables les unes avec les autres.

Point de vue théorique : formules et lois associées. **Les** valeurs des indices probabilistes sont considérées comme des réalisations de variables aléatoires. Ces variables aléatoires tendent vers des lois de probabilité dans le cas des tests asymptotiques (écart-réduit, t-score, χ^2, G^2), ou elles suivent des lois de probabilité dans le cas des tests exacts. Les formules, les variables aléatoires types associées (dans le cas des tests exacts) et les lois de probabilité correspondantes sont résumées dans le tableau ci-dessous.

Indice	***Formule***	***Type de test et variable aléatoire-type***	***Loi de probabilité***
Ecart-réduit (z-score)	$\frac{O_{11}-E_{11}}{\sqrt{E_{11}}}$	*Test d'égalité des moyennes* $\frac{\overline{X}-m}{\frac{\sigma}{\sqrt{n}}}$ *où* $\overline{X} \sim N\left(m, \frac{\sigma}{\sqrt{n}}\right)$ $\overline{X}$ *est la moyenne d'échantillon. La variance* σ *est supposée connue*	Loi normale (loi continue)
t-score	$\frac{O_{11}-E_{11}}{\sqrt{O_{11}}}$	*Test d'égalité des moyennes* $\frac{\overline{X}-m}{S}$ *où* $\overline{X} \sim N(0,1)$, $S \sim \chi^2_N$ *La variance est inconnue*	Loi de Student à N degrés de liberté (loi continue)
χ^2	$\sum_{i,j} \frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}}$	$N\sum_{i,j}\left(\frac{p_{ij}-p_{i\bullet}p_{\bullet j}}{p_{i\bullet}p_{\bullet j}}\right)^2$	Loi du χ^2 (loi continue)
G^2	$2\sum_{i,j} O_{ij}\log\left(\frac{O_{ij}}{E_{ij}}\right)$	$-2\sum_{i,j}\log\left(\frac{L_0^{(ij)}\left(X_1,...,X_{O_{ij}}\right)}{L_1^{(ij)}\left(X_1,...,X_{O_{ij}}\right)}\right)$, *où* $\left(X_1,...,X_{O_{ij}}\right)$ *est un échantillon de même loi. On teste* $p_{ij}=p_{ij}^0=\frac{E_{ij}}{N}$ *contre* $p_{ij}=p_{ij}^1=\frac{O_{ij}}{N}$	Loi du χ^2 (loi continue)
Spécificités	$\sum_{k=O_{11}}^{O_{11}+O_{12}} \frac{\binom{O_{11}+O_{12}}{k}\binom{N-(O_{11}+O_{12})}{O_{11}+O_{21}-k}}{\binom{N}{O_{11}+O_{21}}}$	(test exact, la formule correspond exactement à l'expression de *p(X≥k)*)	Loi hypergéométrique (loi discrète)
Information mutuelle	$\log_2\left(\frac{O_{11}}{E_{11}}\right)$	$\log_2\left(\frac{p(x,y)}{p(x)p(y)}\right)$	/

Tableau II.2.10 : Formules et lois associées aux indices retenus

Les formules du tableau ne sont pas complètement en adéquation avec les formules implémentées dans les logiciels de textométrie, qui procèdent parfois à des adaptations. On reviendra plus en détail sur ce point pour le calcul des spécificités de Lafon (paragraphe 2.1.3.e).

Par ailleurs, les formules de l'écart-réduit, du t-score et de l'information mutuelle permettent de distinguer surreprésentation et sous-représentation, car elles sont signées, autrement dit, elles retournent des valeurs positives en cas de surreprésentation et négatives sinon. Les formules du G^2, du χ^2 et des spécificités (test exact de Fisher) évaluent seulement le degré de saillance sans distinguer les saillances positives (par surreprésentation) et négatives (par sous-représentation). On aura donc tendance à les ajuster comme suit :

- pour le G^2 et le χ^2, les indices sont multipliés par $sgn(O_{11}\text{-}E_{11})$, c'est-à-dire que le signe est celui de la différence $(O_{11}\text{-}E_{11})$;
- pour les spécificités, le signe dépend de la valeur modale m : il est positif si O_{11} est supérieur à m, négatif sinon. m est définie par $\frac{(f+1)(t+1)}{T+2} - 1 \leq m \leq \frac{(f+1)(t+1)}{T+2}$ où $\begin{cases} f = O_{11} + O_{12} \\ t = O_{11} + O_{21} \end{cases}$ et correspond dans la plupart des cas à la partie entière de $\frac{(f+1)(t+1)}{N+2}$ (Lafon, 1984:59).

Limites de validité théorique. Les indices des tests asymptotiques (écart-réduit, t-score, χ^2 et G^2) connaissent des limites de validité théorique de par leur nature même : ils ne retournent pas les probabilités exactes, mais des valeurs supposées proches, sous hypothèse que l'indice peut être assimilé à une loi donnée. Les limites théoriques se situent à deux niveaux : au niveau de la validité des hypothèses qui amènent à associer un indice à une loi ; au niveau des limites de la convergence asymptotique (l'écart avec la loi de probabilité est limité pour de grands échantillons mais il tend à s'accroître sur de petits échantillons). De façon plus détaillée :

- *Écart-réduit z* : pour associer légitimement l'écart-réduit à la loi normale, la formule doit avoir pour dénominateur la variance (celle de la variable aléatoire dont O_{11} est une réalisation et qu'on considère comme égale à $\sqrt{E_{11}}$). Or cette variance n'est généralement pas connue (Manning et Schütze, 2003:188). (Lafon, 1984:109) oppose au doute théorique un contre-argument de validité empirique de l'écart-réduit, sur la base de simulations. Ces simulations donnent des résultats particulièrement concluants sur les fréquences élevées, moins concluants sur les faibles fréquences.
- *T-score* : dans le cadre de l'étude des cooccurrences, la statistique du test t ne respecte pas certaines hypothèses qui permettraient de l'associer à la loi correspondante (loi de Student). La variable aléatoire associée à l'indice est obtenue par quotient de deux variables aléatoires. La variable au dénominateur, associée à $\sqrt{O_{11}}$, doit normalement être la somme de plusieurs variables aléatoires suivant une loi normale et indépendantes (une variable pour chaque occurrence du corpus, susceptible de prendre les valeurs 0 ou 1). Le problème théorique se situe au niveau des variables associées aux occurrences : celles-ci, binaires, sont difficilement assimilables à des variables normales. Pour plus de détails, nous invitons le lecteur à se reporter à (Evert, 2005:82). Ce dernier préconise de considérer le test t comme une heuristique de l'écart-réduit *z* plutôt que de l'associer à la loi de Student.

- χ^2 *:* la statistique du χ^2 rejoint asymptotiquement la loi de probabilité associée, à savoir la loi du χ^2, mais elle s'en écarte pour les faibles fréquences. Les limites couramment évoquées dans la littérature sont des fréquences théoriques de 5 ($E_{ij} \leq 5$) ou une taille du corpus de 20 au minimum.
- G^2 *(LLR)* : les limites sont du même ordre que pour l'indice du χ^2. Tout comme ce dernier, le G^2 converge asymptotiquement vers une loi du χ^2, mais il s'en écarte pour les faibles fréquences. L'écart est toutefois moins marqué que pour le χ^2 (Dunning, 1993).
- *Spécificités (Fisher)* : le test de Fisher est un test exact, associé à une loi discrète, la loi hypergéométrique. En tant qu'indice associé à un test exact, les spécificités ne se heurtent pas aux écueils rencontrés par les autres indices.

Les spécificités, c'est-à-dire l'indice associé au test de Fisher, nous serviront de référence par la suite, aussi bien pour comparer les divers indices que dans notre cadre expérimental. Nous revenons plus en détail sur les spécificités à l'issue de ce paragraphe comparatif.

Divergences au niveau des basses fréquences. Les indices ont tendance à converger entre eux au niveau des hautes fréquences, mais leur comportement diffère au niveau des basses fréquences et s'écartent des résultats des spécificités issues de la loi hypergéométrique.

Plus précisément, le χ^2, l'écart-réduit, le G^2 et l'information mutuelle surestiment les faibles fréquences : la significativité de l'occurrence ou de la cooccurrence d'une unité sera jugée plus importante que ce qu'aurait permis d'établir une probabilité exacte. Ce phénomène est particulièrement marqué pour l'information mutuelle :

> « None of the measures we have seen [, *ie* t-score, χ^2 and likelihood ratios] works well for low frequency events. But there is evidence that sparseness is a particularly difficult problem for mutual information » (Manning et Schütze, 2003:181)

La surestimation est plus importante pour le χ^2 que pour le G^2 (Dunning, 1993). Le G^2 est relativement proche du test de Fisher (Pedersen, 1996).

À l'inverse, le t-score sous-estime les faibles fréquences (Evert, 2005:112).

Des divergences modérées en pratique. Les lois de comportement, les formules et les tendances au niveau des basses fréquences diffèrent, mais dans un cadre expérimental tel que le nôtre, les résultats convergent pour la plupart des indices, moyennant quelques précautions.

La première précaution concerne l'interprétation des valeurs retournée par les indices. Il convient de préférer les rangs obtenus à partir des indices plutôt que de considérer les échelles de valeurs. En effet, les échelles de valeurs diffèrent d'un indice à l'autre, à la fois au niveau des valeurs et au niveau des accroissements de valeurs affectés à deux unités (pas d'accroissements proportionnels entre indices). Dans l'exemple ci-dessous, on voit que, pour tout indice, le classement des trois unités est le même : *réseaux* < *police* < *pédophile*. Cependant, les valeurs prises et la façon dont elles s'accroissent varient d'un indice à l'autre.

	O_{11}	$O_{1.}$	$O_{.1}$	N	écart-réduit	t-score	chi^2	G^2	spécif.	IM
réseaux	8	32	40474	395330	2,6	1,7	7,6	8,2	2,0	1,29
police	29	114	40474	395330	5,1	3,2	28,7	30,7	5,7	1,31
pédophile	73	213	40474	395330	11,0	6,0	134,0	128,7	22	1,74

Tableau II.2.11 : Comparaison des valeurs prises par les indices pour trois unités lexicales

Si les rangs sont utilisés, les indices permettent d'obtenir une hiérarchie qui varie peu d'un indice à l'autre :

> « the t test and other statistical tests are most useful for *ranking* collocations. The level of significances itself is less useful. » (Manning et Schütze, 2003:166)

De fait, dans des articles de référence comparant différents indices (Pedersen, 1996 ; Dunning, 1993), les auteurs se ramènent à une comparaison de rangs plutôt qu'une comparaison de valeurs. Nous proposons en Annexe 1 une illustration : à l'exception de l'information mutuelle, les rangs varient peu et les changements concernent surtout des permutations sur un petit nombre de rang.

La deuxième précaution se situe au niveau des faibles fréquences. Nous entendons par là les hapax principalement, les fréquences de 2 ou 3, et éventuellement 4. Si on ne les exclut pas, les hapax risquent d'émerger massivement en tête de liste et ainsi d'éclipser des unités saillantes qui présentent un minimum de récurrence, comme cela semble émerger dans les résultats expérimentaux de (Dunning, 1993). Les critiques à l'égard des tests portent principalement sur les faibles fréquences.

Si on respecte ces précautions (comparaison de rangs et exclusion des fréquences trop faibles), les résultats retournés par les différents indices sont comparables. Il convient toutefois de considérer l'information mutuelle à part, dont les bases théoriques diffèrent et dont les résultats sont susceptibles de s'écarter de ceux des autres indices, comme illustré dans l'expérience en Annexe 1. Les autres indices convergent. Selon le degré de convergence, il est possible d'effectuer des distinctions ou des rapprochements plus précis entre indices. Il y a une convergence particulièrement forte entre χ^2 et écart-réduit. Le G^2 est l'indice qui se rapproche le plus de Fisher, même s'il reste très proche du χ^2 et de l'écart-réduit. Le t-score tend à se distinguer des autres indices.

e- Indice retenu dans le cadre expérimental : les spécificités de Lafon

Les spécificités de Lafon sont liées au test de Fisher et proviennent de la loi hypergéométrique. Leur implémentation dans les différents logiciels varie légèrement, notamment au niveau des approximations retenues, mais la base est la même.

Plusieurs auteurs s'accordent pour dire que les indices issus du test de Fisher ont des fondements théoriques plus solides que d'autres indices :

> « In principle, all these tests should compute (more or less) the same p-values, although there are certain differences between asymptotic tests and exact tests (especially Fisher's test, which is based on the conditional distribution for fixed row and column sums). These differences have been discussed at length in mathematical statistics. After decades of controversy, most experts seem to agree now that Fisher's test produces the most meaningful p-values (*cf.* Yates 1984). » (Evert, 2005:110)

La qualité des spécificités (test de Fisher) a été mise en évidence par (Pedersen, 1996). (Brunet, 2007) souligne également son efficacité rapport à d'autres indices :

> « On a expérimenté deux autres indices, *le Rapport de Vraisemblance* de Dunning et *l'Information mutuelle* de Church, tous les deux issus de la formule de Jaccard et utilisant les mêmes ingrédients (…). Or si l'on note bien une convergence étroite pour donner une valeur théorique à la cooccurrence de deux mots, la mesure est incertaine dès qu'il faut apprécier les écarts. On s'en est donc tenu à la méthode hypergéométrique qui n'est pas la plus économique mais qui reste la plus sûre. » (Brunet, 2007:10)

Il utilise cette mesure pour calculer les *associations*, réseau de cooccurrents d'un mot-pôle construit à partir des indices de significativité issus du calcul de spécificités. (Evert, 2005)

souligne également la supériorité du test de Fisher par rapport à la plupart des autres indices utilisés, notamment ceux issus des tests statistiques :

> « After decades of controversy, most experts seem to agree now that Fisher's test produces the most meaningful p-values (cf. Yates 1984). We can thus take the Fisher association measure as a reference point for the significance of association group. » (Evert, 2005:110)

L'importance et, indirectement, la qualité de cet indice sont sensibles non seulement à travers les analyses comparatives d'indices évoquées précédemment, mais aussi à travers des choix applicatifs en lexicométrie française (intégration croissante dans les logiciels : différentes études et réflexions amènent Brunet à l'ajouter dans un second temps à son logiciel ; le choix s'est également porté sur cet indice pour le logiciel du projet Textométrie, élaboré après confrontation d'un ensemble de logiciels de textométrie existants). Les spécificités de Lafon (ou son analogue, l'indice désigné en abrégé par test exact de Fisher) constituent un outil de qualité pour mesurer l'affinité sémantique. On s'appuiera sur cet indice dans nos propres expériences. Plus précisément, nous nous appuierons sur les spécificités telles qu'implémentées dans le logiciel Lexico3 (Salem *et al*, 2003).

La critique principale à l'encontre de cet indice est son coût calculatoire, déjà évoqué par (Lafon, 1980:128), mais l'évolution des supports informatiques et l'utilisation d'approximations permettent de réduire ce problème.

Au niveau des calculs, les spécificités de Lafon peuvent être positives ou négatives, selon que le nombre d'occurrences O_{11} est supérieur ou inférieur à la valeur médiane *m*. On fait l'hypothèse que $t = O_{11} + O_{21}$ (taille du sous-corpus dans une approche de type *keyness*) est supérieure à $f = O_{11} + O_{12}$, nombre d'occurrences total de l'unité observée. Si $O_{11} \geq m$, la spécificité est positive et se déduit de la probabilité d'avoir au moins O_{11} occurrences : $p(X \geq O_{11}) = \sum_{k=O_{11}}^{f} \frac{\binom{f}{k}\binom{N-f}{t-k}}{\binom{N}{t}}$. Si $O_{11} \leq m$, la spécificité est négative et se déduit de la probabilité d'avoir au plus O_{11} occurrences : $p(X \leq O_{11}) = \sum_{k=0}^{O_{11}} \frac{\binom{f}{k}\binom{N-f}{t-k}}{\binom{N}{t}}$.

$p(X \geq O_{11})$ et $p(X \leq O_{11})$ sont obtenues par une formule de récurrence exacte à partir de $p(X = O_{11})$. Les approximations de calcul se situent au niveau des factorielles qui interviennent dans $p(X = O_{11})$ (par exemple, $\binom{f}{O_{11}} = \frac{f!}{O_{11}!(f-O_{11})!}$). Elles sont obtenues à l'aide d'une formule de Stirling dont la précision est améliorée par un terme en $e^{1/12n}$ par rapport à la formule utilisée classiquement : $n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}$. L'écart entre la valeur exacte de $p(X = O_{11})$ et la valeur approchée par des formules de Stirling $p'(X = O_{11})$ dépend

de O_{11} et de *f*, mais, même pour des valeurs très faibles, il faut aller au-delà de trois chiffres significatifs pour constater des différences[54] (*cf.* Annexe 2).

Dans les logiciels de textométrie, les spécificités retournées s'appuient sur l'approximation préconisée par Lafon. Elles ne sont pas strictement égales aux probabilités calculées, mais elles les convertissent en une autre valeur par l'application d'une fonction à celles-ci. Elles permettent généralement d'inverser l'échelle des probabilités : pour les probabilités, comprises entre 0 et 1, plus la valeur est petite, plus l'événement est significatif ; pour les spécificités des logiciels, plus la valeur est grande, plus l'événement est significatif. Les spécificités des logiciels ne sont pas limitées à l'intervalle [0;1], elles peuvent prendre des valeurs positives et négatives, supérieures à 1 en valeur absolue. Lexico3 retourne des valeurs entières, qui correspondent à l'exposant de $p(X \geq O_{11})$ ou $p(X \leq O_{11})$. Dtm-Vic utilise la valeur-test[55]. Hyperbase utilise une fonction dont nous ne connaissons pas le détail, mais qui est destinée à retourner une échelle de valeurs comparable à celles retournées par l'écart-réduit. Ceci facilite l'interprétation des résultats : lorsque l'utilisateur passe de valeurs retournées par l'écart-réduit à celles retournées par un calcul de spécificités, il reste familiarisé avec la gamme de valeurs[56].

f- Quelques considérations sur la fréquence

La fréquence occupe une position particulière dans le champ des indices, où elle est à la fois concurrente, base de calcul et critère d'évaluation des résultats, autrement dit, elle peut être utilisée elle-même comme indice, elle sert de base de calcul aux autres indices et elle est utilisée comme critère récurrent pour discriminer et évaluer les résultats.

La fréquence peut être utilisée comme indice pour évaluer la saillance ou l'affinité sémantique. Elle s'exprime de deux façons, sous forme de fréquence absolue (nombre d'occurrences) ou de fréquence relative (proportion relativement à un corpus). Un paramètre est nécessaire dans le premier cas, deux dans le second (nombre d'occurrences et taille du corpus), contrairement aux indices présentés qui s'appuient sur quatre paramètres (nombre d'occurrences dans le sous-corpus, nombre d'occurrences dans le corps, taille du sous-corpus et taille du corpus). Pour des corpus de même taille, fréquence relative et fréquence absolue sont équivalentes. Pour des corpus de taille disparates, la fréquence relative correspond à une normalisation de la fréquence absolue et rend les corpus comparables.

L'intérêt des autres indices est qu'ils intègrent plus d'informations, puisqu'ils dépendent de quatre paramètres. D'une certaine façon, ils synthétisent l'information que délivre un couple de fréquences (dans le corpus et dans le sous-corpus), tout en intégrant une forme de normalisation par rapport à la taille des corpus et sous-corpus. Ces indices sont plus complexes à calculer et apparemment plus complets, mais leur supériorité n'est pas démontrée. Ainsi, (Daille, 1994b) compare différents scores pour l'extraction de termes sous forme d'unités complexes, parmi lesquels les fréquences et des scores d'association tels que l'information mutuelle. Les résultats sont évalués par confrontation des termes obtenus selon

[54] c'est-à-dire qu'il faut aller au-delà de deux chiffres après la virgule selon la notation scientifique, de la forme $a \times 10^{-b}$, où *a* est un nombre décimal de l'intervalle [1;10[et *b* un nombre entier

[55] Une façon d'appréhender le rapport entre probabilité et valeur-test peut se faire de façon graphique, à partir de la représentation d'une gaussienne. La probabilité correspond à l'aire sous la courbe d'une gaussienne au-delà de deux bornes symétriques (c'est-à-dire sur les deux extrémités). La valeur-test est, en valeur absolue, l'ordonnée de ces bornes. Pour une description plus détaillée et précise, on pourra se reporter à (Lebart, 2003:201-202 ; Lebart et Salem, 1994:133,180-184).

[56] Communication personnelle d'E. Brunet.

chaque score à une liste de référence par domaine. La fréquence permet d'obtenir les meilleurs résultats dans le cadre de son expérience :

> « Frequency is the most significant score to detect terms of a technical domain. This result contradicts numerous results of lexical resources which claim that association criteria are more significant than frequency. » (Daille, 1994b)

Les indices statistiques ont donné des résultats concluants dans d'autres cadres, l'idée n'est pas de nier leur intérêt mais de souligner la complémentarité de la fréquence à leur égard : la fréquence donne un autre regard sur les résultats qui n'est pas nécessairement plus mauvais, et donc qu'il convient de ne pas écarter.

Par ailleurs, dans la démarche d'utilisation et d'analyse d'indices statistiques, deux tendances contraires s'affirment : d'une part, on cherche à s'affranchir de la fréquence, d'autre part, on y revient toujours. Le simple recours à des indices autres que la fréquence témoigne de la volonté de s'en affranchir. À l'inverse, les résultats associés aux indices sont souvent analysés en termes de fréquence, à travers la distinction entre haute fréquence et basse fréquence, à travers une comparaison des indices à des rapports de proportionnalité, à travers la distinction des familles d'indices selon qu'elles sont ou non rattachées au critère d'*effect size,* donc de fréquence relative (un critère de fréquence est alors utilisé implicitement, puisqu'on cherche à éliminer la part de significativité due à des effets d'échelle). De même, le filtrage des résultats s'appuie généralement sur deux seuils : un seuil de significativité, obtenu à partir des indices, et un seuil de fréquence. La fréquence reste une référence implicite, un critère intuitif auquel on se ramène systématiquement.

Quelle place accorder alors à la fréquence lorsque d'autres indices sont utilisés ? Certains auteurs préconisent de comparer des résultats pour des fréquences ou des tailles de corpus comparables, tels que (Labbé et Labbé, 2001) :

> « Étant donné l'influence considérable de la fréquence sur la probabilité, il n'est pas souhaitable de comparer, du point de vue de leur spécificité, des mots de fréquences trop différentes. Cela devrait conduire à faire apparaître clairement la fréquence aux côtés de la spécificité et à découper les tableaux en plusieurs compartiments (pour le moins : fréquences fortes, moyennes et faibles). » (Labbé et Labbé, 2001:17)

Mais cette position ne revient-elle pas à rejeter ce qui distingue les indices des fréquences absolues ou relatives ? Autrement dit, il semblerait que l'on perde une grande partie de l'intérêt des indices si on fuit les disproportions : dans ce cas, que gagne-t-on par rapport aux fréquences relatives ? Choisir un indice statistique, n'est-ce pas prendre le risque de s'affranchir de ce que donne à voir la fréquence ?

Sans aller jusqu'à ne considérer que des groupes de fréquence homogène, il semble judicieux de garder les fréquences en complément des indices statistiques, et sur ce point, nous rejoignons (Labbé et Labbé, 2001). Dans la mesure où fréquence et autres indices statistiques ne donnent pas les mêmes résultats, ils offrent un point de vue sur les données ayant chacun leur intérêt, qui méritent d'être tous deux pris en compte et mis en regard.

Enfin, il semble important d'avoir une idée de l'influence de la fréquence sur les indices statistiques. Cette influence n'a pas été étudiée en détail dans le présent travail, mais elle peut s'obtenir en analysant des variations de fonctions où la variable est la fréquence dans le corpus ($x = O_{11} + O_{12}$) et où certains paramètres sont fixés : la fréquence dans le sous-corpus ($O_{11} = constante$) ou encore la proportion d'occurrences dans le sous-corpus ($\frac{O_{11}}{O_{11} + O_{12}} = constante$).

2.2 Établir des critères de sélection

2.2.1 Choix de seuils fréquentiels et statistiques

Les pondérations affectées aux unités peuvent servir de critère pour sélectionner un ensemble d'unités susceptibles d'éclairer sur le nouveau sens. La sélection d'unité nécessite de déterminer des seuils de sélection. Ces seuils s'appliquent à la fréquence et à la significativité.

a- Seuils de fréquence

Les seuils de fréquence sont très variables en pratique. Par exemple, (Rossignol et Sébillot, 2002) recherchent des mots-clés thématiques à partir de seuils de fréquence relativement élevés (60 occurrences pour un corpus constitué de 8000 paragraphes provenant des archives du *Monde diplomatique*). À l'inverse, (Drouin *et al.*, 2006) descendent jusqu'à un seuil de 2, mais ils précisent les limites de validité statistique du modèle et la nécessité d'un contrôle par des experts au niveau des unités très peu fréquentes.

La diversité des seuils est conditionnée par plusieurs paramètres : l'objectif de l'étude (une recherche thématique peut imposer des seuils plus élevés qu'une recherche centrée sur des événements rares), la taille du corpus, la part accordée à une validité théorique ou empirique (modèles statistiques vs pertinence évaluée, par exemple, par des experts).

De façon générale, il n'existe pas de critère absolu pour établir un seuil de fréquence. Le choix d'un seuil relève d'un certain arbitraire. Cependant, un certain nombre d'arguments permettent de guider ce choix.

Limite de validité des modèles statistiques. La plupart des modèles statistiques ne sont plus valides pour les faibles fréquences. Selon (Evert, 2005:166-167), les modèles statistiques ne sont plus applicables pour des fréquences de 1 ou 2, tandis qu'un seuil de 5 suffit à garantir leur validité. La validité des modèles statistiques croît avec la fréquence.

Un seuil relatif à la taille du corpus. (Picton, 2009:116) souligne la difficulté d'appliquer des seuils élevés en pratique : les grands corpus permettent de choisir des seuils élevés, mais il est fréquent de ne disposer, en pratique, que de corpus de taille réduite. Cette question n'est pas sans intérêt dans notre cadre : dans la mesure où on vise une détection précoce, on souhaite pouvoir accéder au nouveau sens alors que les nouveaux emplois n'ont pas achevé leur diffusion. Le risque est alors de ne disposer que d'un nombre réduit de données. À cela s'ajoute le fait qu'on souhaite jouer sur différents contrastes, notamment à travers la constitution de sous-corpus domaniaux. Un corpus de grande taille risque ainsi de se fractionner en sous-corpus qui serviront de base de calcul et dont la taille sera beaucoup plus réduite. Dans le cadre du modèle hypergéométrique à la base des spécificités de Lafon, (Labbé et Labbé, 2001) suggèrent le critère de "seuil d'absence spécifique", inspiré de (Salem, 1987), selon lequel la fréquence seuil est celle pour laquelle une occurrence dans la plus petite partie corpus corresponde à une surreprésentation :

> « Le choix des formes soumises au calcul [de spécificité] doit donc se porter sur celles dont la fréquence totale est au moins égale à une valeur telle qu'une absence, dans la plus petite des parties du corpus, aboutisse à une spécificité négative » (Labbé et Labbé, 2001:6)

Récurrence sémique. La recherche du nouveau sens s'appuie sur la recherche d'isotopies locales. Les conséquences sont doubles : au niveau lexical, on peut se permettre d'accepter des fréquences faibles, car des formes lexicales peu présentes pourront être porteuses d'un sème qui sera associé à d'autres formes lexicales, et ainsi, cette forme peu fréquente pourra alimenter la récurrence sémique ; au niveau sémique, c'est-à-dire après annotation, la recherche d'une récurrence sémique implique d'imposer des seuils plus élevés, sauf si des

regroupements de sèmes (par exemple sur critère morphologique) sont prévus a posteriori. En effet, dans le cas de regroupements de sèmes, des sèmes peu fréquents peuvent s'associer pour former un groupe de taille non négligeable. Il y aurait alors intérêt à appliquer un seuil de fréquence à l'issue du regroupement, portant sur la fréquence cumulée.

b- Seuils de spécificités

Tout comme pour les fréquences, il n'existe pas de seuil absolu.

Seuil plus bas que les seuils de probabilité classiques. Le seuil de probabilité est souvent fixé à 5% dans les modèles statistiques, ce qui correspond à un seuil de spécificité de 2 d'après les spécificités de Lafon telles qu'implémentées dans le logiciel Lexico3. Or ce seuil, adapté à l'étude de données dans divers champs disciplinaires, est certainement moins pertinent pour des données linguistiques. On utilise, certes, le schéma d'urne et l'hypothèse d'indépendance comme base de modélisation, mais les données linguistiques ne se comportent pas selon ce schéma. De ce fait, un grand nombre d'unités présentera un comportement statistiquement remarquable. Le choix d'un seuil plus sélectif paraît donc pertinent pour de telles données. Au niveau des travaux existants, le choix du seuil de spécificité par défaut du logiciel Lexico3 en témoigne : il est fixé à 5, ce qui correspond à une probabilité de 1/100 000. Cependant, d'autres paramètres, tels que la taille du corpus ou sous-corpus considéré, interviennent également dans la valeur du seuil et viennent relativiser ces considérations.

Seuils en fonction de la taille du sous-corpus des voisinages. (Labbé et Labbé, 2001) ont montré que la taille des sous-corpus influe sur la gamme de valeurs prises par les spécificités. Toutes choses égales par ailleurs, un sous-corpus plus grand favorisera des spécificités plus marquées (plus élevées pour les spécificités positives, et inversement pour les spécificités négatives, avec un plus grand nombre d'éléments considérés comme spécifiques). Au niveau des logiciels, Lexico3 propose un seuil de spécificité par défaut qui est constant, quelle que soit la taille du sous-corpus (même si l'utilisateur reste libre de régler ce seuil). À l'inverse, dans Hyperbase, le seuil par défaut est variable, il est ajusté de façon à ce que la taille de la liste d'unités spécifiques soit comprise dans un certain intervalle et dépend donc de la taille du sous-corpus, puisque celle-ci influe sur la gamme de spécificités, donc sur le nombre d'unités au-delà d'un certain seuil de spécificité :

> « Le seuil minimal de cet indice est établi par défaut à une valeur convenable vu la taille du corpus. » (Brunet, 2007:10)

Seuils en fonction de la fréquence. (Labbé et Labbé, 2001) ont montré que les fréquences influaient sur la valeur des spécificités (des fréquences élevées permettront une gamme de spécificités plus étendues, avec des valeurs potentiellement plus élevées) et ils recommandent de n'étudier les spécificités que pour des ensembles d'unités de fréquences comparables. Il serait certainement judicieux de fixer des seuils de spécificités en fonction de la fréquence, mais si l'idée n'est pas inintéressante en théorie, elle nous semble relativement délicate à mettre en pratique.

Seuils bas pour brasser large. Le choix du seuil de spécificité peut également être fonction de la taille de la liste d'unités qu'on souhaite sélectionner. Supposons que la taille du corpus et du sous-corpus considérés permettent une gamme de spécificités étendue a priori. Notre approche nous pousse à sélectionner un seuil de spécificité plutôt bas, que les observables en corpus soient les formes lexicales ou les sèmes projetés en corpus. Un seuil plus bas permettra de récupérer une liste d'unités spécifiques plus importante. Si les observables en corpus sont les unités lexicales, une étude distributionnelle des sèmes interne au dictionnaire aura plus de chance d'être statistiquement valide si la taille des données est plus grande (celle-ci est fonction du nombre d'unités lexicales retenues, correspondant au nombre d'entrées lexicales

parcourues). Si les observables en corpus sont les sèmes issus d'une annotation sémique. En effet, on n'est pas sûr de la qualité des entités issues de la procédure d'annotation sémique : celle-ci fournit des candidats-sèmes, dont la pertinence n'est pas garantie et reste à valider. Si un candidat-sème saillant (spécificité supérieure au seuil) est proche d'un autre candidat-sème également saillant, on peut supposer que ce candidat-sème a une certaine pertinence car il est indirectement repris.

2.2.2 Règles au service de l'interprétation

Le choix de seuils fournit un critère de sélection en fonction du degré de saillance ou de présence des unités. L'ajout de règles permet d'introduire des critères destinés à faciliter une interprétation ou une réutilisation des résultats.

Les règles peuvent être quantitatives ou qualitatives. Parmi les règles quantitatives, on signalera le critère de la taille de la liste retournée. Si la liste est destinée à une analyse humaine, elle doit présenter une diversité sans pour autant atteindre une taille démesurée, telle que l'analyste soit dépassé par le foisonnement des données. Si cette liste est soumise à des traitements ultérieurs, destinés notamment à faire émerger des regroupements, il convient de disposer d'un nombre d'éléments suffisamment important. De même, si cette liste doit être confrontée à des listes issues d'autres analyses, il semble judicieux de disposer de listes de taille relativement comparable.

Parmi les règles qualitatives, on mentionnera la règle d'interprétabilité des unités isolées. Il est peu judicieux de récupérer des unités, lexicales ou sémiques, dont l'interprétation n'est pas possible. Ce problème se pose notamment pour des unités à caractère prédicatif : l'approche 'sac de mots' isole certaines unités qui ne sont interprétables que si elles sont combinées à d'autres unités. Ainsi, dans l'étude sur *Outreau* (*cf.* chapitre III.1, 4.1.3.b), la structure 'sac de sèmes' a généré des unités telles que /existence/ ou /découverte/, qui, isolément, posent des problèmes d'interprétabilité (sur quoi portent l'existence et la découverte ?). Il n'existe pas de méthode systématique pour repérer les unités interprétables ou non. Le recours à des listes d'exclusion ou le choix de certaines catégories grammaticales (privilégier les noms par exemple) peuvent contribuer à améliorer la sélection.

3. Organiser les différentes sources d'information

3.1 Fusionner, juxtaposer, composer : quel compromis ?

Il existe de multiples vues sur les textes, et, dans notre cadre, plusieurs axes d'observation de l'évolution de sens, à travers les différents paramètres et observables ainsi que la multiplicité des combinaisons possibles. Les résultats reflètent des perspectives distinctes. La question suivante se pose : jusqu'où aller dans la synthèse des différentes représentations obtenues ? Faut-il les conserver de façon dissociée ou les fusionner en une représentation commune ?

L'analyse des textes n'est pas une science exacte. Une représentation synthétique comporte une part d'arbitraire. Confier la synthèse des résultats à un traitement automatique et se reposer exclusivement sur cette représentation revient à minimiser le contrôle linguistique, pourtant indispensable. Ajoutons que certains résultats ne sont pas réductibles l'un à l'autre, comme le souligne (Pincemin, 2008) : elle présente les propriétés utilisées pour décrire les textes comme des *dimensions descriptives* et souligne que certaines ne peuvent être regroupées, mais devraient être maintenues distinctes[57].

[57] « … certaines dimensions sont orthogonales, au sens où il n'y a pas d'interrelation permettant de passer de

Cependant, la vocation des traitements est de fournir un autre regard sur le texte, qui guide l'interprétation et qui est suggestif. Si les représentations sont trop nombreuses, trop variées et simplement juxtaposées, elles désorienteront le lecteur au lieu d'être des guides interprétatifs. Il semble peu judicieux de traiter une masse de données pour générer de nouvelles données tout aussi foisonnantes.

On doit donc garder à l'esprit qu'une représentation synthétique est une proposition, pas un absolu. Il convient de conserver des traces des résultats intermédiaires. Il semble judicieux de privilégier une ***hiérarchie de représentations*** plutôt qu'une seule vision synthétique ou une constellation indifférenciée de représentations : on propose une représentation synthétique principale (éventuellement 2 ou 3), avec possibilité d'accès aux autres représentations, par exemple en faisant varier des paramètres ou en accédant aux représentations initiales.

3.2 Hiérarchiser les résultats : quelques critères et techniques

Les traitements que nous ciblons en priorité pour l'allocation de signifié font intervenir :

- différentes sources : les ressources lexicographique et textuelles ;
- différents types d'unités observées : domaines, unités lexicales et sèmes dans le corpus ; domaines et sèmes dans le dictionnaire ;
- différents jeux de contrastes au sein du corpus.

Nous proposons quelques principes pour hiérarchiser les représentations issues des divers traitements.

Hiérarchie entre corpus et dictionnaire : priorité au dictionnaire. Le corpus et le dictionnaire donnent des images du sens de la cible lexicale. Le dictionnaire fournit une image initiale, le sens codé. Par exemple, pour *toxique,* le point de départ sera un ensemble de définitions associées aux domaines de la biologie ou de la médecine. Le corpus donne une image du sens relative aux emplois discursifs. On peut ainsi observer des emplois dans les domaines de la biologie, de l'environnement, de la médecine, mais aussi dans le domaine de la finance. Ces emplois se répartiront de façon variable dans le temps et selon les domaines, ils permettront de faire émerger des unités lexicales saillantes (*actifs, titres, créances* par exemple, dans le domaine de la finance) ou des sèmes saillants (/*banque/* par exemple). Les représentations du sens en corpus ne sont pas une fin en soi : pour allouer un signifié, il faut les confronter au sens codé, de façon à amender celui-ci et à proposer un nouveau signifié (dans le cas de *toxique*, ajout d'une nouvelle définition dépendant du domaine de la finance, mais pas de nouvelle définition en environnement, car les emplois dans ce domaine peuvent s'expliquer par le sens codé). En sortie, l'objectif est donc de se ramener au sens codé. Autrement dit, à l'issue des traitements, on cherche à se positionner par rapport à la ressource lexicographique. La représentation des résultats relative au dictionnaire est donc prioritaire sur celle du corpus, qui est seulement une phase intermédiaire des traitements, non la phase finale.

La représentation associée au dictionnaire se construit à partir des informations issues du corpus. Elle résulte de traitements appliqués à des informations elles-mêmes obtenues à partir d'autres traitements. Autrement dit, elle est obtenue par composition : elle est fonction de la représentation des résultats associée au corpus, elle-même fonction des traitements mathématiques appliqués au corpus.

---

l'une à l'autre. Il n'est peut-être pas pertinent d'entredéfinir toutes les dimensions (…). C'est en ce sens que le texte n'est plus *à la croisée* de descriptions finalement liées en faisceau, mais se déploie *entre* les dimensions descriptives » (Pincemin, 2008:959).

Hiérarchie en fonction de la granularité : les domaines, puis les sèmes. Les traitements peuvent s'appliquer à différents types d'observables : domaines, unités lexicales ou sèmes dans le corpus ; domaines ou sèmes dans le dictionnaire.

Selon le principe de détermination du global par le local, on donnera priorité à de l'information correspondant à une granularité plus grossière. Les représentations associées à une granularité plus fine leur seront hiérarchiquement subordonnées.

Synthèse des jeux de contraste pour chaque type d'observable. Dans le corpus, chaque type d'observable peut donner lieu à plusieurs jeux de contrastes : entre différents domaines, au sein d'un domaine, en fonction de la période et en fonction de l'ordre de cooccurrence. Les résultats peuvent s'articuler autour des jeux de contrastes ou autour de type d'unités.

Si le type de contrastes est privilégié comme axe d'organisation, il semble nécessaire de maintenir une distinction entre les différents types d'unités, notamment de conserver distinctement domaines et sèmes, qui constituent des observables non réductibles les uns aux autres.

Pour réduire la masse de résultats, il semble plus pertinent de donner priorité à des représentations relatives à un type d'observable. Ceci revient à synthétiser en un résultat global les résultats issus des différents jeux de contrastes. Les résultats plus détaillés peuvent servir à titre complémentaire : ils permettront d'apprécier le type de contribution de chaque jeu de contraste et ainsi de nuancer l'interprétation.

Quelques techniques pour synthétiser. La façon de regrouper des résultats issus des différents jeux de contrastes dépend de la façon dont se présentent les résultats. Les résultats sont des valeurs affectées soit à des unités seules (résultats sous forme de listes), soit à des couples d'unités de même nature ou de nature différente (résultats sous forme de tableau, ou matrice). La façon de mettre en relation différents jeux de résultats dépend des formats disponibles en entrée ainsi que des formats qu'on souhaite en sortie.

Si les formats des jeux de données sont identiques (par exemple, deux tableaux croisant les unités lexicales et les domaines, avec même nombre de lignes et de colonnes), il sera possible de calculer un indicateur synthétique des résultats en appliquant différentes fonctions aux couples de valeurs, par exemple des fonctions linéaires (moyennes, barycentres), des fonctions de type min ou max, etc. On pourra également regrouper les informations dans une même structure (affectation de couples de pondérations, au lieu de disposer de deux tableaux de pondérations). Si les formats varient, d'autres techniques pourront être utilisées, avec des opérations ensemblistes sur les listes (intersection, réunion ou constitution de sous-ensembles en fonction des éléments communs et propres à chaque liste). Dans une perspective de synthèse, l'objectif est de réduire plusieurs valeurs caractéristiques d'une unité ou d'un couple d'unités à une valeur.

Dans la section à venir, où l'idée est de structurer les données, l'objectif n'est pas de réduire la multiplicité des valeurs mais de s'en servir pour articuler des unités entre elles.

4. Structurer les unités

Rappelons qu'on cherche à caractériser le nouveau sens à partir de fonds et de formes sémantiques. Ceux-ci ne se dégagent pas tant d'une unité ou d'un ensemble d'unités indifférenciées que de la façon dont des unités se combinent, qu'elles se regroupent ou qu'elles s'opposent. De même, au (chapitre I.3, 3.1.2), nous avons souligné l'importance de la structure interne du sémème. Les techniques à même de dégager des structures en fonction des valeurs dont on dispose (les pondérations précédemment décrites) jouent donc un rôle fondamental.

4.1 Critères de structuration

À partir du moment où l'on dispose d'au moins deux coefficients associés à chaque unité considérée (domaines, unités lexicales ou sèmes), on peut envisager de dégager des structures plus complexes que les listes hiérarchiques, qui constituent une structure unidimensionnelle où les éléments sont considérés un par un.

Les coefficients peuvent être affectés à :

- des paires d'unités de même nature (domaines, unités lexicales, sèmes définitoires)
- des paires d'unités de nature différente (sèmes définitoires – sèmes domaniaux ; sèmes - unités lexicales ; unités lexicales – type de contraste ; etc.)

Ces données peuvent se représenter sous forme de tableaux, ou matrices. Les matrices ne sont pas nécessairement carrées (le nombre de lignes peut différer du nombre de colonnes), les valeurs peuvent appartenir à des ensembles de définition variés (elles peuvent être des réels ou des entiers, en fonction de la méthode de calcul utilisée pour les pondérations puis le couplage des résultats).

Les structures ont pour vocation d'organiser l'information en terme d'enrichissement et de reconfiguration ; de permettre à l'interprétant de jongler entre une vue d'ensemble ou un regard nuancé des résultats ; de dégager des ensembles cohérents. On cherche :

- des ensembles cohérents qui soient sémantiquement homogènes, pour refléter des isotopies ;
- des ensembles hétérogènes, pour mettre en évidence le faisceau d'isotopies, c'est-à-dire une combinaison sémantique originale qui distingue le nouveau sens de l'unité lexicale du reste du lexique.

On veut donc trouver des dénominateurs communs et associer des unités de nature différente. On cherche également à compacter l'information en dégageant un petit nombre de grandes tendances, avec des possibilités de raffinement et d'accès à des vues plus détaillées.

4.2 Types de structures : regroupements, réseaux, hiérarchies et dispositions d'ensemble

Les structures possibles sont nombreuses. Nous en retiendrons quatre : les regroupements, les réseaux, les hiérarchies et les dispositions d'ensemble. Un certain nombre de techniques permettent de faire émerger telle ou telle structure, certaines étant spécifiques à un type de structure, d'autres pouvant relever de plusieurs types de structures.

Les ***regroupements*** permettent de répartir les unités en ensembles et de créer une structure essentiellement cloisonnée. Il existe deux sortes de regroupements : ceux qui relèvent du *hard clustering* (structure de partitions) et ceux qui relèvent du *soft clustering* ou *fuzzy clustering* (Manning et Schütze, 2003:499). Lorsque la constitution de groupes ou de classes est sous forme de partition (*hard clustering*), tous les éléments sont répartis dans des classes et ces classes n'admettent pas de chevauchements, autrement dit, tous les éléments appartiennent à une unique classe. Lorsqu'on relâche les contraintes de partitionnement (*soft clustering*), certains éléments peuvent appartenir à plusieurs classes ou n'appartenir à aucune classe. Le fait qu'un élément appartienne à plusieurs classes peut s'exprimer en termes de présence / absence (présence dans les classes A et B, absence de la classe C) ou à travers des pondérations, souvent issues de probabilités (appartenance à 50% à la classe A, à 30% à la classe B et à 20% à la classe C). Les regroupements peuvent également être distingués selon

qu'ils sont obtenus par reconfigurations successives des groupes jusqu'à ce que soit atteint un optimum (regroupements non hiérarchiques) ou par agglomérations ou divisions successives, générant des groupes emboîtés (regroupements hiérarchiques).

Au niveau des méthodes de regroupement :

- La méthode-type de *hard clustering* non hiérarchique est la méthode des k-moyennes (Mac Queen, 1967). Les classes sont constituées selon un objectif de minimisation de l'inertie intraclasse. Partant de k centres initiaux, les éléments les plus proches sont affectés à chaque centre et forment une classe ; un nouveau centre est calculé pour chaque classe ; les classes sont amendées, par réaffectation d'éléments aux nouveaux centres ; et ainsi de suite jusqu'à stabilisation des classes. Les cartes de Kohonen (Lebart *et al.,* 2003:199-200 ; Kohonen, 1989) relèvent aussi de ce type de méthodes : des groupes sont constitués relativement à un ensemble de centres dont la configuration n'est pas libre mais contrainte, contrairement aux k-moyennes (les centres forment un "filet", des liens de contiguïté sont imposés) ; elles ne se limitent pas aux regroupements et interviennent pour définir d'autres types de structures, car elles positionnent les groupes les uns par rapport aux autres.
- Des mélanges de distribution sont utilisés pour les méthodes de *soft clustering* non hiérarchiques : chaque classe correspond à une distribution statistique (par exemple gaussienne), leur importance relative varie. Chaque élément a une probabilité d'apparition selon les différentes distributions, ou de façon complémentaire, chaque classe a une certaine probabilité d'être la classe de rattachement d'un élément donné. L'algorithme d'espérance d'espérance-maximisation (*expectation-maximization*) est une technique de référence pour les mélanges de distribution (Manning et Schütze, 2003:514-527), dont le principe général est grossièrement le suivant : les éléments sont affectés à chaque distribution ; cette affectation permet d'estimer les paramètres du modèle (moyenne et variance par exemple) ; une nouvelle affectation pondérée des éléments à chaque distribution est effectuée, sous critère de maximisation de la vraisemblance, et l'importance relative de chaque distribution est réévaluée ; cette nouvelle affectation permet de recalculer les paramètres du modèle, et ainsi de suite.
- Les méthodes de classification hiérarchique sont ascendantes lorsqu'elles procèdent par agglomérations successives (CAH) ou descendantes par subdivisions successives (CDH). Les CAH peuvent regrouper des classes selon différents critères : en fonction des éléments les plus éloignés de chaque classe, des éléments les plus proches, en fonction des distances entre de barycentres (établis en prenant en compte ou non l'existence, et donc la taille de sous-classes internes à une classe), en fonction de la variance au sein d'une classe, voulue minimale (Caraux et Pinloche, 2005). Les classifications descendantes procèdent par subdivisions successives. Une méthode de classification descendante est intégrée dans le logiciel Alceste et détaillée dans (Reinert, 1983).

De même que, dans le cadre de l'indexation de documents, (Pincemin, 1999:683-688) prend position contre le principe de partitionnement et pour des regroupements autorisant des chevauchements de classes et des éléments non classés, les alternatives au *hard clustering* semblent pertinentes dans le cadre de l'allocation de signifié : certaines unités de sens peuvent être transversales et être associées à plusieurs sous-ensembles constitutifs du sémème. Cette transversalité est au demeurant ce qui permet le lien entre nouveau sens et ancien sens : certaines unités de sens sont héritées de l'ancien sémème, comme l'idée de /propagation/ ou de /caractère néfaste/ dans le cas de *toxique* en contexte de crise financière.

Les ***réseaux*** privilégient les dépendances, les liens entre unités. Le passage d'un élément à l'autre s'effectue par récursivité. Les réseaux répondent plus à une perspective de navigation ou encore de rayonnement à partir d'un élément, plutôt qu'à une perspective de composition - décomposition.

Des réseaux, ou graphes, peuvent être obtenus assez immédiatement à partir des données sous forme de matrice : tout coefficient non nul associé à un couple d'unités définit un lien entre ces unités. Par-delà ces liens obtenus immédiatement, des techniques telles que les marches aléatoires, fondées sur des chaînes de Markov, permettent de reconfigurer le réseau initial de façon non triviale : des unités initialement sans lien seront reliées, des nœuds de forte convergence émergeront à l'issue des navigations en plusieurs étapes, etc. On peut se reporter à (Gaume, 2004) pour une application sur le lexique. Les réseaux peuvent servir à mettre en évidence des zones fortement connectées telles que les cliques (Ploux et Victorri, 1998), qui se rapprochent des regroupements mais restent en lien avec le reste du réseau. Des structures linéaires, en chaînes, peuvent également être construites, comme l'effectue (Martinez, 2003) pour générer des enchaînements de cooccurrences, ou polycooccurrences.

Les ***hiérarchies*** font ressortir l'importance des éléments hiérarchisés et orientent l'ordre de parcours des résultats. Les hiérarchies ont déjà été évoquées pour les unités pondérées séparément, indépendamment de toute structuration préalable. Elles peuvent également s'appliquer aux regroupements et moduler soit la structure externe (mise en évidence de l'importance relative des différents groupes), soit la structure interne (différenciation des éléments constitutifs d'un groupe). Les ***sériations*** peuvent se voir comme un complément aux hiérarchies : elles sont une forme d'organisation unidimensionnelle, qui ne procède pas nécessairement par importance croissante ou décroissante, mais le plus régulièrement possible, de façon à ce qu'il y ait le minimum de variations d'un élément à l'autre.

Les classifications hiérarchiques permettent d'obtenir indirectement des hiérarchies de groupes d'éléments par rapport à un élément donné. Partant de cet élément, les nœuds définissent successivement des groupes complémentaires. Ces groupes peuvent ensuite être hiérarchisés en fonction de la profondeur du nœud qui les articule à la branche où apparaît l'élément ciblé. Des sériations peuvent être obtenues par permutation de groupes d'éléments ou vecteurs ou de selon de différents critères, ayant notamment pour objectif de minimiser l'écart de distances entre deux groupes ou vecteurs de rangs successifs (Caraux et Pinloche, 2005). Pour les séries ordonnées telles que les séries chronologiques, on peut s'appuyer sur des coefficients d'autocorrélation afin de voir si l'organisation des données est couplée à l'ordre de la série. D'autres méthodes d'analyse des séries chronologiques sont présentées dans (Salem, 1988).

Les ***dispositions d'ensemble*** permettent d'aborder les résultats en termes de proximité et d'éloignement entre unités ou groupes d'unités. Elles n'imposent pas de frontière ni de lien, mais se déploient selon un continuum. Elles peuvent guider vers des regroupements ou des dépendances, mais elles maintiennent un certain flou. Elles sont étroitement liées à la visualisation et à une forme d'organisation spatiale, telle que les nuages de points. Ce mode d'organisation est particulièrement intéressant pour mettre en relation des unités de nature différente, par exemple en considérant le positionnement d'unités supra-lexicales de type domaine par rapport à la disposition d'ensemble des unités infra-lexicales. Les dispositions d'ensemble sont étroitement liées à la visualisation.

Au niveau des techniques, on évoquera les méthodes factorielles, dont les deux méthodes principales sont l'analyse des correspondances et l'analyse en composantes principales (Lebart, 2008). Ces méthodes permettent d'appréhender la façon dont s'organise un nuage de points dans un espace multidimensionnel, à partir d'une projection en deux dimensions. Les cartes de Kohonen permettent également d'appréhender des dispositions d'ensemble entre groupes d'éléments. Elles se présentent sous forme d'un espace quadrillé. Chaque groupe correspond à une case. Deux groupes contigus, proches dans la disposition d'ensemble, seront également proches en termes de distance.

5. Visualiser

5.1 Importance de la visualisation

Les traitements mathématiques et informatiques servent à faire face à la masse de données, mais leur intérêt ne se limite pas à la réduction de la masse de données : ils permettent également de porter un autre regard sur le contenu informationnel, qui n'est pas directement accessible à travers la lecture exhaustive des supports. On ne se situe donc pas dans une perspective comparable à celle des résumés automatiques, où l'on voudrait fournir des descriptions clefs en main du nouveau sens (contextes les plus représentatifs des nouveaux sens ou définition rédigée).

Les représentations visuelles (nuages de points, histogrammes, graphes, etc. ; *cf.* section suivante) constituent un moyen d'accès au sens particulièrement efficace. De façon plus illustrative que démonstrative, la toile est un lieu de diffusion de l'information où l'accès au contenu et au sens est primordial, et l'accès au contenu passe de plus en plus par des modes de représentation visuelle tels que les nuages de mots-clés ; on constate également un succès grandissant des cartes heuristiques[58] comme mode de diffusion des connaissances. Avec des moyens techniques plus performants et moins de contraintes éditoriales, les représentations visuelles servent de plus en plus de vecteurs de transmission du contenu informationnel. Sans pour autant être une preuve scientifique, ce succès populaire témoigne en faveur d'une efficacité de ces représentations pour accéder à l'information.

Dans son argumentation en faveur des représentations visuelles, (Polguère, 2002) souligne que la vue est un moyen d'accès à la connaissance particulièrement efficace, car il permet d'accéder rapidement à une grande quantité d'informations. De plus, le recours aux représentations visuelles permet de jouer sur le niveau de granularité de la description du contenu : on peut généralement accéder soit à des informations relevant de la macrostructure, soit à de l'information propre à une microstructure, tout en maintenant un lien entre les différents niveaux d'information.

Ces représentations visuelles sont une aide pour accéder au sens, elles constituent un moyen de guider l'interprétant. Il convient de ne pas les confondre avec le sens lui-même et de les utiliser comme auxiliaires, non comme des absolus. (Polguère, 2002) les qualifie de *métaphores visuelles* et met en garde contre les interprétations abusives :

> « Il arrive fréquemment, lorsque la métaphore semble bien remplir son office, que ceux qui en usent se mettent graduellement à la confondre avec le phénomène même qu'elle est censée modéliser. »

[58] Les cartes heuristique ou *mindmapping* est une forme de représentation des connaissances, des pensées, des concepts. Il consiste en des représentations imagées, schématiques, qui disposent dans l'espace des mots-clés. Voir par exemple (Buzan et Buzan, 2000), (Keller et Tergan, 2005).

L'intérêt des représentations visuelles est d'ouvrir à d'autres formes descriptions, qui orientent vers d'autres modes de représentation du sens. On présente quelques techniques de visualisation, qui ont servi à de nombreux travaux en textométrie.

5.2 Techniques de visualisation

La visualisation peut servir à mettre en relief les structures qui se dégagent des données numériques. Cette mise en valeur peut servir d'illustration à projeter dans un espace de représentation des structures identifiées au préalable par analyse quantitative, elles peuvent compléter de telles analyses quantitatives, ou elles peuvent être la base de l'extraction de structures.

De nombreuses techniques de visualisation permettent de mettre en valeur les saillances et les structures. Elles peuvent focaliser l'attention sur différents aspects : les quantités ou proportions (5.2.1) ; les ensembles et sous-ensembles (5.2.2) ; les connexions ou dépendances (5.2.3) ; la disposition générale (5.2.4) ; la matérialisation dans le texte (5.2.5). Face à cette panoplie, on discutera de la question des représentations multiples et de leur articulation (5.2.6). On ne fera pas un panorama exhaustif des techniques de visualisation, dont l'étude complète constituerait un champ d'investigation en soi. Seules quelques techniques seront présentées, qui n'ont pas toutes été mises en œuvre dans notre cadre expérimental, mais qui présentent des aspects complémentaires à intégrer dans une approche complète.

5.2.1 Visualiser les quantités : les histogrammes, classiques et efficaces

Les histogrammes sont des outils de visualisation basiques. Classiques en statistique, ils sont un mode de représentation qui appartient à une culture commune et ils sont des guides efficaces de l'interprétation.

Les histogrammes mettent en relief les pondérations Le quantitatif ressort particulièrement, l'accès au qualitatif ne se fait que dans un second temps (lecture des axes et des étiquettes de légende).

Ce type de représentation est particulièrement intéressant pour visualiser des séries ordonnées, notamment celles associées à une évolution chronologique. L'axe du temps est généralement l'axe des abscisses. Il met en évidence le déroulement temporel et une éventuelle diffusion, qu'il s'agisse de la cible, d'un sème de son sémème ou d'un ou plusieurs domaines.

Les histogrammes présentent des limites à partir du moment où le nombre d'observables devient trop important : la représentation graphique connaît alors une forme de saturation.

5.2.2 Visualiser les ensembles : tables et structures arborescentes

La visualisation d'ensembles permet d'appréhender des groupes d'unités thématiquement homogènes, issus des unités saillantes en corpus ou internes au sémème. Dans le cadre de nos expériences, des tables constituées manuellement ont permis de mettre en évidence des groupes d'unités correspondant à différentes dimensions sémantiques associées à une cible lexicale.

Les tables de listes et les structures arborescentes permettent de faire ressortir les groupes, les sous-groupes, les imbrications. Ce type de représentation est fortement structurant, parfois même à l'excès : les divisions sont nettes, les frontières clairement définies.

Les modes de visualisation évoqués correspondent à des partitionnements. Il est également possible de faire émerger des ensembles qui partagent des éléments communs et se

chevauchent. Les chevauchements peuvent apparaître indirectement dans les tables de listes, si deux listes partagent des éléments en commun.

Un autre mode de représentation qui, graphiquement, se rapproche des tables, mais qui, de fait, intègre un agencement spatial beaucoup plus poussé est celui des cartes de Kohonen. Celles-ci répartissent les unités selon un quadrillage obtenu à partir de la structure multidimensionnelle des données. Deux cases proches dans la représentation seront également proches numériquement. Même si les cases font apparaître des frontières, la juxtaposition ou au contraire l'éloignement des cases permettent de nuancer et d'assouplir ces frontières : la frontière entre deux cases proches correspondra à une séparation moins marquée qu'entre deux cases distantes.

5.2.3 Visualiser les dépendances : graphes, molécules et chaînes

Une représentation qui met en relief les dépendances intéresse notre démarche à plusieurs égards :

- L'allocation de signifié est sous-tendue par l'idée que le sens se construit à travers l'émergence en discours d'une forme sémantique ou molécule sémique. La notion de molécule sémique implique des composantes entretenant des liens plus ou moins importants avec la cible lexicale et également reliées entre elles.
- L'analyse des emplois discursifs à travers plusieurs grilles de lecture, suite aux jeux de contrastes entre différents espaces textuels ou différents niveaux d'observation (sèmes, unités lexicales, domaines) est susceptible de faire émerger des dépendances, en réseau ou en chaîne. Par exemple, l'émergence d'un sème pourra être relative à celle d'une ou plusieurs unités lexicales, elles-mêmes dépendantes d'un domaine et/ou d'une période.
- Le nouveau sens n'est pas totalement détaché des sens déjà présents dans la définition : certaines unités sont communes au nouveau sens et à d'autres sens, autrement dit, certains sèmes permettent de tisser des liens entre les différentes définitions, ou au contraire, ils n'entretiennent des liens qu'au sein d'une définition particulière. Ajoutons que les dépendances se manifestent entre différents types de sèmes : les sèmes microgénériques ou spécifiques dépendent des sèmes mésogénériques, c'est-à-dire des domaines. Visualiser les dépendances peut donc donner à voir une organisation relative aux domaines et des liens transversaux entre les différentes définitions.

Les dépendances peuvent être mises en évidence à travers des graphes. Ceux-ci peuvent se présenter sous forme de chaînes ou de chemins simples (sans circuits), autrement dit, sous forme de structures linéaires. Ce type de structure peut être adapté pour mettre en relation des unités de nature différente (période de temps, domaine, unité lexicale, sème). Sinon, les dépendances peuvent se présenter sous forme de structures non linéaires, avec des circuits et notamment des composantes connexes (sous-graphe dont tous les sommets sont reliés deux à deux). Ces structures sont intéressantes pour aborder l'organisation d'unités de même nature (articulation de l'ensemble des sèmes entre eux, par exemple).

Les polycooccurrences du logiciel Coocs (Martinez, 2003) permettent de visualiser des chemins simples, c'est-à-dire des structures linéaires orientées partant d'un sommet initial. Ces structures rattachent successivement des unités lexicales à une cible lexicale donnée.

L'application Prox (Gaume, 2004) permet de voir des organisations d'unités sous forme de graphes non orientés, qui donnent à voir des réseaux de synonymes. Les unités lexicales sont reliées selon qu'il existe ou non une relation de synonymie. La longueur des arêtes est définie par une mesure de proximité dans le graphe de synonymes.

Les graphes ne privilégient pas les frontières mais les dépendances, qu'elles résultent de liens directs, de chaînes ou de structures connexes. Ce type de structure favorise la navigation dans les résultats.

Les graphes ne permettent pas de voir directement des groupes cohérents, qui font perdre des finesses d'articulation entre unités en imposant des délimitations, mais qui sont très structurants. Cependant, différentes techniques permettent de constituer des groupes à partir de graphes (*cf.* par exemple (Schaeffer, 2007) pour des détails techniques généraux ou (Pons et Latapy, 2006) pour une technique particulière).

5.2.4 Visualiser la disposition générale : nuages de points et cartographies

Certaines représentations visuelles permettent d'observer la disposition des éléments sans imposer au lecteur de délimitations ou de liens de dépendances et en faisant ressortir la configuration d'ensemble. Ce mode de visualisation maintient un flou au niveau des microstructures, mais il permet d'appréhender les résultats dans leur globalité et de faire émerger des tendances générales en terme d'éloignements, de proximités ou de positionnement d'une unité par rapport à l'ensemble des autres unités.

Ce type de représentation est particulièrement approprié pour faire ressortir des éléments dont le comportement est très atypique par rapport aux autres éléments, ou encore pour mettre en évidence des oppositions marquées. Ainsi, en corpus, ces outils visuels sont susceptibles de faire ressortir des unités caractérisant les nouveaux emplois, atypiques par rapport aux anciens emplois.

Les représentations sur la disposition générale sont particulièrement utiles lorsque les données à analyser se situent dans un espace multidimensionnel. Cette situation est courante : elle se présente dès que les données peuvent être représentées sous forme de matrices à plusieurs lignes et colonnes (par exemple, des matrices unités lexicales – périodes, unités lexicales – domaines, sèmes – domaines, etc.). Les dimensions de l'espace peuvent être alors définies par les colonnes de la matrice, ou par ses lignes. Dans ce cadre, la représentation des unités étudiées se présente souvent sous forme de carte en deux ou trois dimensions, par projection de données multidimensionnelles dans un espace de dimension 2 ou 3. Elle peut notamment être obtenue par des techniques d'analyse factorielle telles que l'analyse en composantes principales ou l'analyse des correspondances, qui projettent les données selon les axes principaux (c'est-à-dire les axes d'élongation maximale du nuage de points).

De telles représentations peuvent notamment être générées à l'aide du logiciel Dtm-Vic (Lebart et Piron, 2011). Ces représentations ont été utilisées notamment dans le cadre des sondages, pour des réponses à des questions ouvertes : les visualisations donnent des informations sur la structure générale du contenu sémantique, dans son ensemble ou relativement à des variables illustratives telles que l'âge des répondants, leur sexe, etc. (Lebart *et al.,* 2003). Dans le même ordre d'idée, des cartes de synonymes ont été réalisées à partir du logiciel VisuSyn (Ploux et Victorri, 1998), par constitution de cliques de synonymes dans un espace multidimensionnel (c'est-à-dire des groupes de synonymes obtenus à partir d'un graphe de synonymes), puis par projection de ces cliques en deux dimensions.

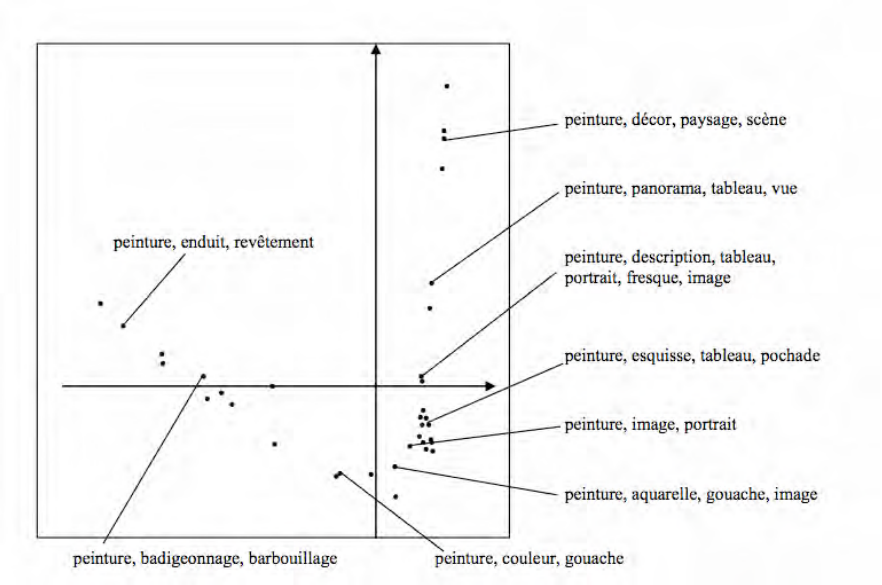


Figure II.2.12 : Exemple de carte de synonymes générée par VisuSyn, reprise de (Victorri, 2002)

Les représentations issues des analyses factorielles sont pertinentes pour faire émerger des oppositions plus que des regroupements. L'interprétation des résultats doit respecter certaines règles. Une proximité apparente au voisinage de l'origine ne correspond pas nécessairement à une proximité dans l'espace multidimensionnel. En particulier, des points proches et situés à proximité de l'origine en deux dimensions pourront être en réalité éloignés. Des techniques d'analyse permettent d'évaluer l'importance de ce biais et ce qui relève de fausses proximités[59]. Par contre, si deux ensembles de points sont éloignés du centre et situés aux deux extrémités d'un axe, ils pourront toujours être interprétés comme opposés. Moyennant les précautions évoquées à propos d'un éventuel biais, la proximité de deux points s'interprétera en termes de corrélation (Lebart *et al.*, 2003:18). Si plusieurs sous-nuages apparaissent dans le plan factoriel, ils pourront être analysés comme des groupes distincts. Si le nuage a une forme parabolique en deux dimensions, cette configuration pourra notamment refléter l'existence d'une série ordonnée, telle qu'une série chronologique (Lebart, 1995:89-94).

Les représentations de nuages de points peuvent être enrichies à l'aide d'autres techniques, susceptibles de faire mettre en valeur des regroupements, d'introduire des distinctions, de relier des données ou d'ajouter de nouvelles informations. Ainsi, il est possible d'utiliser des CAH en amont, d'en extraire des groupes puis de projeter les groupes issus des CAH dans le plan factoriel. (Lebart, 1995:185-206) insiste sur l'intérêt d'articuler classification et représentations factorielles : ces deux techniques sont complémentaires, aussi bien sur le plan théorique que sur le plan interprétatif. On trouvera une illustration par exemple chez

[59] Cela dépend de la forme initiale du nuage de points, selon que l'élongation du nuage est à peu près la même selon tous les axes (comme une sphère ou un ballon rond en 3D) ou que l'élongation est importante selon certains axes par rapport aux autres axes (comme un ballon de rugby en 3D). La projection du nuage en deux dimensions tend à "écraser" des points distants sur des axes autres que les axes principaux ; en deux dimensions, on pourra voir apparaître des points proches autour de l'origine, qui étaient initialement éloignés dans l'espace multidimensionnel. Le biais est important dans le cas d'une sphère, faible lorsque le nuage de point est très allongé selon un ou deux axes principaux. L'analyse des pourcentages d'inertie de chaque axe par rapport aux autres permet d'avoir une idée de la forme du nuage, donc de l'importance du biais (Lebart, 1995:89). L'observation du nuage dans des plans factoriels définis par les axes suivants (axes 2 et 3, par exemple) permet de vérifier les proximités entre points.

(Boussidan *et al.* 2010). Des informations d'une autre nature sont ajoutées dans Proxidocs (Roy, 2007) : les cartes présentent une organisation spatiale des documents textuels dans un espace en deux dimensions en fonction de leur proximité thématique ; des informations complémentaires sont ajoutées au graphique : un jeu de couleurs permet d'identifier le thème dominant de chaque document et un jeu sur la taille des points met en relief la taille des documents apportent des informations qui s'ajoutent à celles mises en relief à travers la disposition.

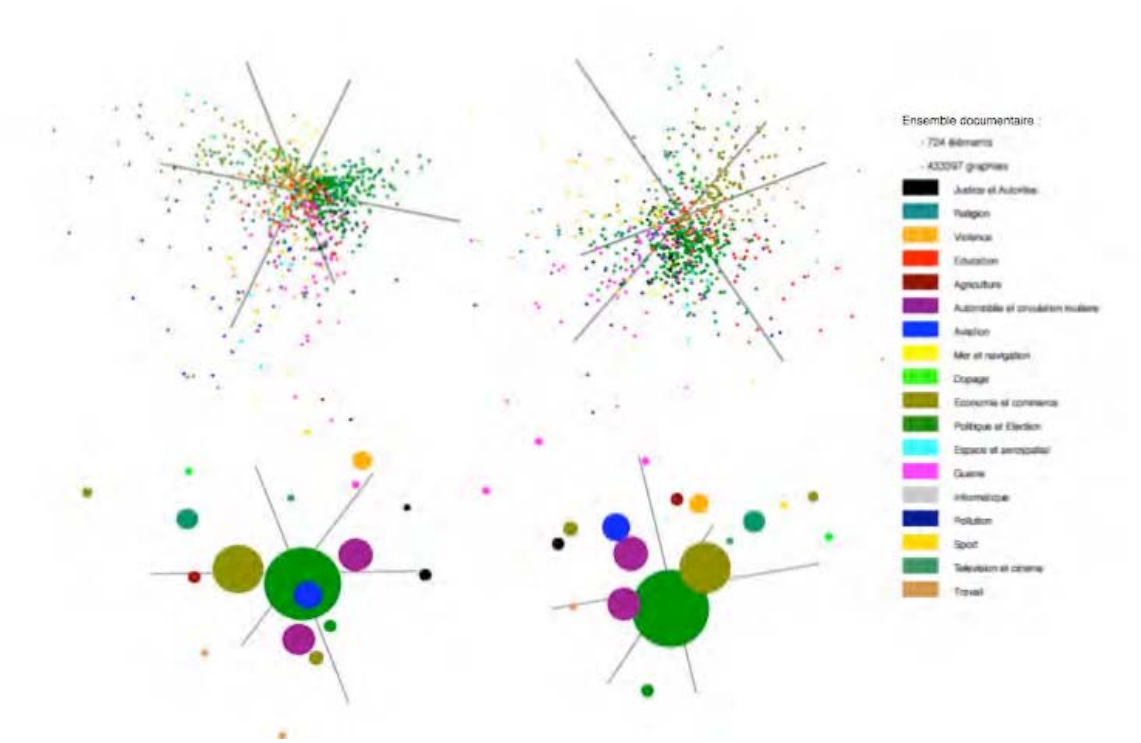
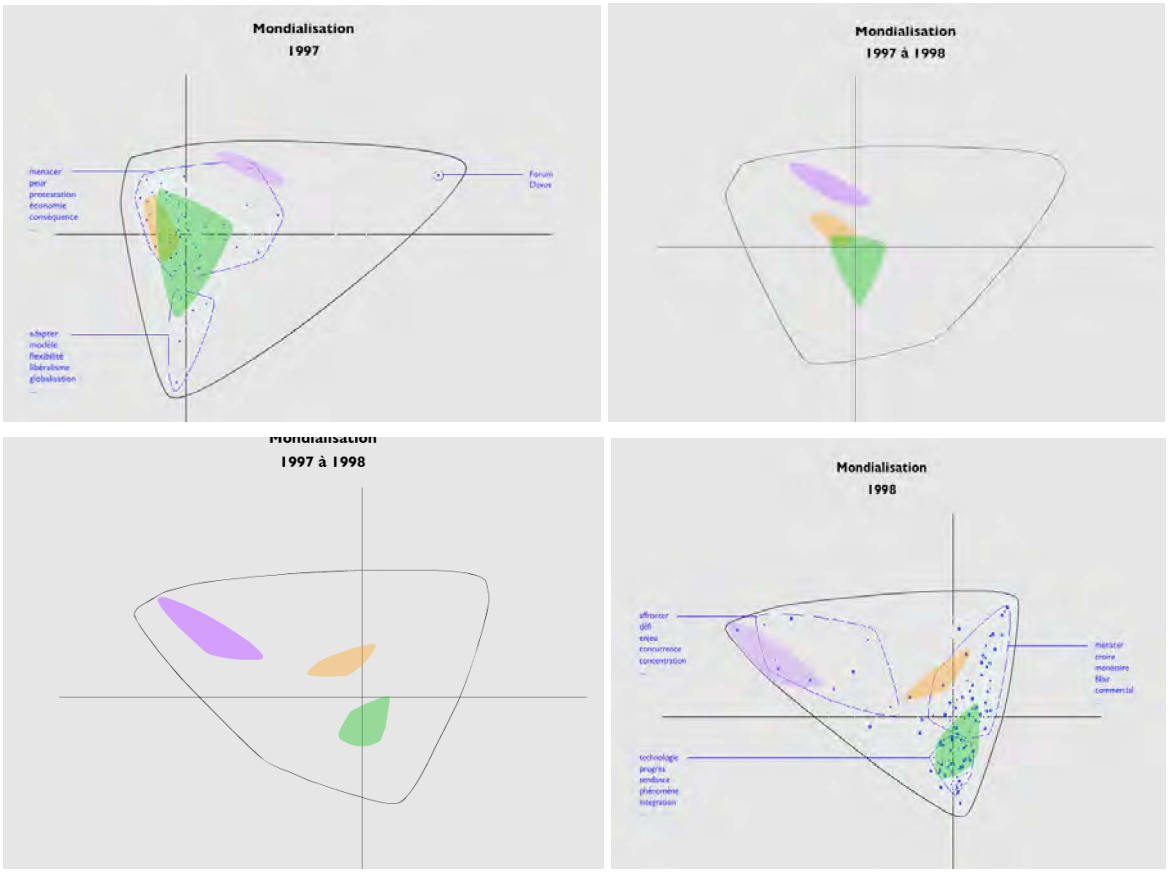


Figure II.2.13 : Cartographie enrichie générée par Proxidocs (Roy, 2007:105)

5.2.5 Visualisations statiques ou dynamiques

Si un grand nombre de représentations sont figées, un certain nombre d'outils génèrent également des représentations dynamiques. Dans notre cadre, une représentation dynamique présente un intérêt particulier, car on cherche à observer la diffusion du nouveau sens : une visualisation dynamique permettrait de voir comment l'évolution de sens progresse dans le temps ou dans les anciens et nouveaux domaines d'emplois.

Dans cet esprit, (Boussidan *et al.*, 2010) proposent des cartes dynamiques pour représenter l'évolution en diachronie de l'environnement cooccurrentiel de *mondialisation*. La représentation dynamique est obtenue par interpolation d'images en 2D associées à plusieurs périodes de temps. Ces images sont obtenues en couplant AFC et classification ascendante hiérarchique : l'AFC permet une représentation bidimensionnelle par projection d'un nuage de points multidimensionnel, la classification hiérarchique permet de délimiter des groupes de cooccurrents. L'interpolation permet de voir en continu le passage d'une image à l'autre, elle est accessible à l'adresse http://dico.isc.cnrs.fr/en/diachro.html.



Figures 5.14 a, b, c, d : Captures d'images successives de la représentation dynamique de l'évolution des cooccurrents de mondialisation entre 1997 et 1998 (Ploux et al., 2011)

Signalons que les représentations dynamiques ne correspondent généralement pas à l'évolution réelle, mais qu'elles résultent souvent d'un lissage des irrégularités et d'une inférence sur les zones lacunaires – ce qui est au demeurant un principe de base de bon nombre de modèles continus appliqués à des données discrètes. Cette démarche met en évidence une régularité d'évolution que des variations mineures d'images successives risqueraient de brouiller. Pour une représentation dynamique réalisée à partir de plusieurs images, il convient d'estimer l'importance des variations des images initiales par rapport au modèle continu (tout comme la variance apporte un éclairage complémentaire à la moyenne dans les études statistiques) ; d'autre part, il faut conserver à l'esprit qu'une représentation dynamique est aveugle sur ce qui s'est passé dans l'intervalle séparant deux images consécutives. Par exemple, il se peut qu'un pic événementiel se situe entre deux périodes de temps à l'origine d'images successives, pour peu qu'elles soient espacées ; une représentation dynamique inférera une évolution sur les périodes intermédiaires, mais elle ne pourra refléter le pic événementiel.

Soulignons enfin que plusieurs représentations figées successives constituent une alternative à une représentation dynamique. Celles-ci fournissent une suite d'arrêts sur image et, en termes de dynamique, elles sont à mi-chemin entre une unique représentation statique et une représentation dynamique, dont l'évolution est continue.

5.2.6 Visualisation ancrée dans le texte : concordanciers nouvelle génération

Les résultats quantitatifs et les représentations visuelles présentées jusque-là offrent un autre regard sur le texte que les données initiales. Pour que cet autre regard soit bien un complément de la réalité textuelle, ces approches gagnent à être complétées par un retour au texte. Ce retour au texte permet de contrôler les résultats et de mettre en regard l'interprétation

en contexte à l'interprétation des résultats de l'analyse mathématique. Il peut prendre deux formes :

- un retour à la source relativement brut, avec une sélection et un jeu sur l'agencement des contextes d'emploi initiaux ;
- un retour aux données textuelles enrichies par les résultats d'analyse. Les textes constituent alors une base de projection des structures qui ont émergé quantitativement.

Le retour à la source n'implique pas une lecture exhaustive de l'ensemble des documents du corpus. Des outils peuvent servir à sélectionner des données textuelles pertinentes ou à les organiser de façon à guider vers l'information pertinente.

Parmi ces outils, les concordanciers jouent un rôle de premier plan. Ils permettent de superposer des cotextes où apparaît la cible lexicale et, ainsi, de les appréhender rapidement. Les trois caractéristiques fondamentales des concordanciers, à savoir le pivot, la taille du cotexte et le critère de tri (Pincemin *et al.*, 2006), ont chacune un rôle propre dans le cadre de l'allocation de nouveau sens. L'organisation des cotextes selon un pivot est adaptée à notre approche focalisée sur une cible néosémique. La taille du cotexte, généralement réduite, offre un regard complémentaire aux traitements effectués : l'analyse statistique s'appuie sur des informations globales, issues de paliers relativement larges ; les cotextes réduits des concordanciers permettent de se concentrer sur le local et de réintégrer des informations éliminées dans le cadre des traitements statistiques (par exemple l'influence de la syntaxe). Des cotextes plus resserrés autour de la cible ne doivent normalement pas donner une nouvelle image du sens, mais ils doivent proposer un autre équilibre dans ce qui ressort : les informations liées au global apparaissent de façon plus latente, les informations dues à des interactions locales ressortent plus en finesse. Enfin, le tri peut être guidé par les résultats d'analyse. Au-delà des tris habituels que proposent les concordanciers, par exemple en fonction de l'ordre alphabétique des énièmes mots du contexte droit ou gauche, on peut envisager des tris guidés par les résultats d'analyse : tris en fonction des domaines ; tris en fonction de certains sèmes et des unités qui les lexicalisent ; tri en fonction des marqueurs de sentiment néologique tels que les guillemets ; tri en fonction de la saillance statistique, etc. Une telle exploitation relève de fonctionnalités des concordanciers définies par (Pincemin *et al.*, 2006), où interviennent les notions de zones et de classes d'objets : les zones sont délimitées non pas par des tailles de contexte mais par la présence de certains types d'unités ; les tris se font en fonction de classes d'objets, où les classes seraient définies par exemple par des sèmes.

La sélection de contenu pertinent peut aussi s'appuyer sur des principes utilisés notamment pour les résumés automatiques. Il s'agit alors de sélectionner un ou des fragments textuels représentatifs des phénomènes observés. Ce type de méthode apparaît également dans (Lebart et Salem, 1988:117-127), à propos de la sélection de réponses modales, c'est-à-dire les réponses les plus représentatives[60], dans le cadre de sondages avec réponses à des questions ouvertes. De même, (Kilgariff *et al.*, 2008) proposent un système, GDEX, pour sélectionner des exemples pertinents afin d'illustrer le sens de collocations dans un dictionnaire d'anglais langue étrangère[61]. Dans notre cadre, il s'agirait d'extraire les paragraphes ou documents jugés les plus représentatifs des nouveaux emplois.

[60] La sélection de la réponse modale est obtenue par calcul d'un indice qui dépend du nombre de formes spécifiques et leur degré de spécificité dans chaque réponse.

[61] Les exemples sont sélectionnés à partir de pondérations affectées aux fragments textuels. Les poids dépendent de plusieurs paramètres : la présence de délimiteurs de phrase, la longueur du fragment, la présence d'expressions anaphoriques, la fréquence des mots ainsi que leur saillance. Les paramètres sont réglés à partir d'une étude manuelle réalisée à petite échelle.

Utiliser texte comme base de projection est pertinent pour faire émerger des phénomènes latents, non immédiatement accessibles à la lecture. En particulier, les isotopies et les formes sémantiques n'apparaissent pas directement dans le texte. L'allocation d'un nouveau signifié nécessite d'avoir identifié des unités de sens saillantes, qui se distinguent par une présence marquée et récurrente dans les cotextes d'emploi. En s'appuyant aux unités lexicales auxquelles sont affectées ces unités de sens, il est possible de faire ressortir ces dernières à partir du matériau textuel, par exemple à travers des jeux de couleurs, des mises en relief (en style gras par exemple) ou par annotation sémique ciblée (ajout des unités saillantes). Ce retour au texte enrichi en informations additionnelles apparaît notamment chez (Perlerin, 2004:141-147), qui présente une application pour visualiser les thèmes à partir d'un surlignage en couleur des unités lexicales associées à ce thème.

Figure II.2.15 : Visualisation de la métaphore conceptuelle de la météorologie boursière, extraite de (Perlerin, 2004:204)

5.2.6 Visualisations multiples : que choisir ?

Les différentes représentations visuelles ne mettent pas toutes en évidence la même perspective, elles offrent des vues complémentaires. Lorsqu'on dispose de plusieurs représentations, la diversité permet des recoupements, susceptibles de faire ressortir les biais interprétatifs induits par telle représentation ou telle autre. Par ailleurs, certaines représentations peuvent mettre en évidence des particularités qui n'émergent pas à travers d'autres représentations. Les multiples vues sur les résultats constituent une forme de mise à l'épreuve d'une représentation donnée, donc une forme de contrôle.

6. Validation

La validation des résultats est incontournable. Elle peut s'appuyer sur des méthodes linguistiques principalement manuelles ou recourir à des techniques mathématiques. Dans le cadre des expériences réalisées, la validation des résultats a été effectuée manuellement.

La validation s'est appuyée sur des retours au texte. Les résultats d'analyse ont été confrontés à l'interprétation des emplois contextualisés, soit à partir de la lecture de documents où apparaissait la cible lexicale, soit à partir de concordances centrées sur la cible lexicale.

Nous avons également eu recours aux connaissances d'un expert. Dans le cadre de l'expérience sur le corpus *Outreau*, le recours à un regard de spécialiste, celui de M. Lecolle, a contribué à valider les résultats. La validation a pris deux formes : 1) une confrontation des résultats quantitatifs à une analyse manuelle réalisée en amont et détaillée dans (Lecolle, 2007b) ; 2) une organisation manuelle de données en parallèle des traitements quantitatifs, selon un format similaire aux sorties des traitements : M. Lecolle a évalué le caractère saillant ou non des unités observées aux différentes périodes de temps, et ce sans connaissance a priori des traitements mathématiques appliqués ni des résultats quantitatifs.

Par ailleurs, nous avons procédé à des recoupements de résultats. Ces recoupements ont été réalisés à partir de sources extérieures, notamment pour vérifier le caractère néologique d'une cible. Ils se sont aussi fondés sur la comparaison de différents axes d'approche. Les apports de l'annotation en traits sémantiques ont par exemple pu être validés par confrontation aux résultats obtenus sans annotation.

La question de la validation est intervenue dans notre cadre pour des expériences relevant d'un stade exploratoire. Après consolidation de la phase exploratoire, des traitements de plus grande ampleur peuvent être envisagés. Même à un stade plus avancé, il convient de garder des moyens de contrôle et de validation du signifié alloué. Les techniques de visualisation articulées à la représentation textuelle constituent une forme de retour au texte favorable à la validation des résultats. Ajoutons que, quelle que soit la chaîne de traitement mise en œuvre, il est nécessaire de conserver des traces des traitements intermédiaires, pour pouvoir éventuellement identifier quel niveau de traitement est à l'origine d'une divergence entre des parcours interprétatifs et les résultats quantitatifs.

D'autres formes de validation existent, notamment mathématiques, mais elles n'ont pas été utilisées dans le cadre de nos expériences. Les expériences réalisées ont plus vocation à confirmer le potentiel de certaines pistes et à illustrer les étapes d'un traitement qu'à valider une procédure complète. Nous reviendrons sur un autre type de validation au (chapitre III.2, 5.3.2).

Le chapitre qui s'ouvre propose une procédure qui s'appuie sur une partie des techniques évoquées dans ce chapitre. L'objectif n'est pas de tester une panoplie d'outils et techniques, mais de définir un enchaînement cohérent en accord avec l'analyse d'outils réalisée dans ce chapitre et étayée par des expériences illustratives basées sur les ressources présentées au chapitre II.1.

Partie III.
Vers un modèle applicatif : mise en œuvre et perspectives

Chapitre III.1
Proposition d'une procédure d'allocation de signifié par granularité sémantique décroissante

Le protocole d'allocation de signifié a été élaboré à partir d'une approche inductive. Diverses expériences ont été réalisées pour valider des lignes de force théoriques qui n'étaient pas nécessairement articulées entre elles à l'origine. Ces expériences forment un ensemble morcelé, constitué d'une variété de cibles lexicales et de corpus d'observation, mais elles ont contribué à faire émerger un système cohérent.

La procédure présentée dans ce chapitre propose une chaîne de traitements qui respecte les principes fondamentaux du modèle théorique (approche du sens à travers une représentation en sèmes, contrastes entre fonds et formes sémantiques, jeu sur la granularité des descripteurs, distinction entre activation, inhibition et enrichissement) et qui intègre les ressources utilisées pour représenter les emplois textuels et les sens codés. Les expériences qui ont servi à l'élaboration du modèle ont été articulées aux différentes étapes du processus. De ce fait, nous ne proposons pas une expérience de validation filée de bout en bout, mais un enchaînement d'étapes formant un tout cohérent, accompagné d'expériences illustratives propres à chaque étape.

Les étapes principales du processus sont la présélection de cibles lexicales (section 1), l'analyse des changements au niveau supra-lexical (domaines) (section 2), l'analyse des changements au niveau lexical (section 3), puis l'analyse des changements au niveau infra-lexical (traits sémantiques définitoires, autres que les domaines) (section 4).

1. Présélection de cibles lexicales

Plusieurs cibles lexicales ont servi de support aux expériences. Ces cibles ont été observées dans différents corpus et elles ont servi à valider différents aspects de la procédure. Nous décrivons les critères qui ont guidé la sélection des cibles (1.1), puis nous détaillons en quoi elles se rattachent à de la néologie sémantique (1.2). Enfin, nous donnons un aperçu général de la façon dont les expériences rattachées à chaque cible se distribuent au cours du processus (1.3).

1.1 Critères utilisés pour la présélection

La présélection de cibles lexicales a été guidée par deux critères :

- par des indices conscients, tels que les guillemets ou les marqueurs de nouveauté, comme dans les exemples ci-dessous :

> « Le deuxième tour a marqué un net coup d'arrêt à ce que la plupart des instituts de sondage annonçaient comme un **« tsunami »** comparable à la déferlante rose qui avait submergé les conseils régionaux en 2004. » *Ouest France,* 18/06/2007
>
> « Vendredi après-midi, George Bush et son secrétaire au Trésor Henry Paulson ont confirmé qu'ils étaient prêts à dépenser "des centaines de milliards de dollars du contribuable" afin de mettre en place un mécanisme permettant aux investisseurs (banques, assurances, fonds...) de se débarrasser de leurs actifs "**toxiques". C'est-à-dire** de tous les produits financiers structurés à base ou non de subprimes devenus invendables depuis le début de la crise. » *Libération*, 20/09/2008
>
> «…aujourd'hui l'économie dite purement financière représente cinquante fois (en termes de transactions) l'**économie dite réelle**. » *Le Figaro*, 23/09/2008

- par des sentiments de locuteurs, couplés ou non à des analyses manuelles de discours. L'étude du sens d'*Outreau* a été impulsée par un travail d'expert sur la question (Lecolle, 2007b). D'autres exemples déjà cités, tels que *brouteur*, ont été présélectionnés à partir de suggestions étayées uniquement par un sentiment néologique de locuteur, et non des indices de détection.

De façon plus générale, la présélection peut reposer sur une détection automatique ou semi-automatique, guidée par le repérage d'indices conscients. Elle peut être complétée par des candidats proposés par des locuteurs, ce qui répond à un principe proposé par (Beust, 2007) qui reprend et adapte au TAL l'idée d'*énaction* (Varela, 1996)[62] : la présélection répond à un couplage entre système et personne, autrement dit, on prend en compte une composante humaine (le ressenti d'utilisateurs) dans la procédure de présélection des candidats. Techniquement, ce couplage peut se faire par le biais d'un module spécifique où les utilisateurs sont amenés à proposer des candidats qu'ils jugent néologiques. Un principe similaire, de type communautaire, régit l'introduction de nouvelles entrées dans Wikipédia (les utilisateurs proposent des entrées candidates, qui sont ensuite soumises à validation). Plutôt que d'intégrer des suggestions d'utilisateurs orientées vers l'enrichissement d'une base néologique, il est aussi possible d'aller rechercher des informations dans des pépinières à candidats. Ceci revient à faire de la veille sur des sources réputées réactives, comme des blogs sensibles à l'évolution de la langue et dont les billets peuvent parfois receler des candidats pertinents (blog *Langue Sauce piquante* du Monde (http://correcteurs.blog.lemonde.fr/), le *Dicomoche* (http://www.dicomoche.net/)). L'interaction avec des utilisateurs peut être coûteuse et demande un certain discernement, mais elle peut constituer un bon complément à une détection semi-automatique.

1.2 Présentation des cibles étudiées

Trois types de cibles ont servi de support aux expériences :

- des cibles néologiques, qu'on peut subdiviser en deux groupes :
  - des cibles relevant de la néologie sémantique : *toxique*, *moléculaire, tablette* et *tsunami* ;

[62] « Varela propose le concept d'énaction (ou cognition incarnée) pour permettre d'appréhender l'action adaptative de tout organisme vivant à son environnement. (…) L'énaction est au départ une théorie du vivant mais cette théorie peut être réinvestie dans d'autres champs de recherche. (…)
L'idée au centre de ce que nous appelons ici une Sémantique Computationnelle Enactive (SCE) est une certaine façon de considérer ce qu'est le sens dans des interactions homme-machine (que ce soit le sens d'un énoncé, d'une phrase, d'un texte, d'une collection de documents …). » (Beust, 2007)

  - des cibles relevant d'autres types de néologie mais ramenées au cadre de la néologie sémantique : *Outreau*, néologie de forme (car non codé dans le dictionnaire) mais converti artificiellement en néologie sémantique par construction manuelle d'un ancien sens ; *économie réelle*, phrasème néologique qui ne sera étudié que pour la question de l'enrichissement en traits sémantiques (donc sans considérer les difficultés liées à l'interaction des sens des composantes simples) et dans le but de valider les apports de l'annotation en traits sémantiques par rapport aux informations présentes sur le plan lexical ;
- des cibles à statut intermédiaire, qui, sans être à proprement parler de la néologie sémantique, s'y rattachent par certains aspects : *tempête*, qui connaît des emplois métaphoriques en contexte de crise (la métaphore étant un procédé fondamental de la néologie sémantique) ; *numérique*, qui connaît une diffusion massive dans un grand nombre de domaines (l'accroissement des emplois et la propagation dans de nombreux domaines sont des mécanismes qu'on cherche à observer pour la néologie sémantique) ; *pourri* en contexte de crise (il y a ambiguïté sur le statut : s'agit-il d'un emploi métaphorique qui n'a pas de caractère de nouveauté ou de nouveaux emplois propres au domaine financier, participant du foisonnement néologique de *toxique* ?) ;
- des cibles témoins, non néologiques, utilisées comme points de comparaison : *actifs, dangereux, délétère, raz-de-marée*.

Le nouveau sens des cibles néologiques ainsi que leur corpus d'étude est présenté globalement dans le tableau ci-dessous, puis détaillé dans les paragraphes suivants pour les cibles *toxique, Outreau* et *économie réelle*. Ces cibles font l'objet d'analyses relativement fines par la suite et il est nécessaire d'apporter des compléments d'information sur leur profil.

Cible lexicale	Corpus	Nouveau sens
toxique (adj)	'Crise financière' (nouveau sens) ; 'Monde Diplomatique' (anciens sens)	Qualifie les produits et instruments financiers à l'origine de la crise financière de 2008.
moléculaire	'Factiva'	Nouveaux emplois dans le domaine de la gastronomie (*cuisine moléculaire*) qui renvoient à un jeu sur les textures, les couleurs et les formes en cuisine créant un effet artistique et qui mettent en avant une pratique régie par des principes scientifiques.
bouclier	'Crise financière' (nouveau sens) ; *Le Monde* (ancien sens)	Dispositif fiscal qui impose un plafonnement des sommes dues légalement (*bouclier fiscal*, *bouclier social)*.
tablette	'Factiva'	Emplois liés à une innovation technologique, la *tablette numérique*, qui connaît une diffusion de grande ampleur dans la société et, de façon corrélée, une diffusion massive dans les discours.
tsunami	'Crise financière'	Emploi métaphorique, pour qualifier des événements violents et destructeurs (la crise financière dans notre corpus d'étude).
Outreau (n.p.)	'Outreau' (Lecolle, 2007b)	Évolution du sens géographique (nom de ville) vers le sens d'*erreur judiciaire par excellence.*
économie réelle	'Crise financière'	Partie de l'économie rattachée à la sphère industrielle, ou encore aux produits et biens de consommation, par opposition à la partie spéculative de l'économie (économie financière ou capitalisme financier).
pourri	'Crise financière' (nouveau sens) ; *Le Monde* (ancien sens)	Qualifie des produits et instruments financiers, de la même façon que *toxique.*

Tableau III.1.1 : Liste des cibles lexicales à caractère néologique

1.2.1 Toxique : de nouveaux emplois en contexte financier

a- Nouveau sens dans le corpus 'Crise financière'

Dans le corpus 'Crise financière', l'adjectif *toxique* qualifie presque exclusivement des produits et instruments financiers. Ces produits désignent les titres générés par les *subprimes,* crédits immobiliers hypothécaires. Le non remboursement des créances a rendu ces titres illiquides[63]. Ces titres devenus invendables sont à la source de l'effondrement du système financier. Deux extraits sont donnés à titre d'illustration sur les emplois de *toxique* :

> « Depuis plus d'un an, le caractère contagieux de la crise du subprime est avéré. La dissémination des **produits appelés " toxiques** " a dispersé les pertes à travers le monde. » (*Le Figaro*, 16 septembre 2008)

> « Au départ dédié au rachat des **actifs toxiques** des banques, afin de soulager le bilan de celles-ci, le plan a finalement été utilisé à recapitaliser les établissements. » (*Le Figaro,* 13 janvier 2009)

b- Ancien sens dans le TLFi et dans le Monde Diplomatique

Les définitions de *toxique* du TLFi (*cf.* figure III.1.2) sont associées à cinq domaines : BIOLOGIE, CHIMIE, MEDECINE, PHARMACOLOGIE ET PHYSIOLOGIE. Au sens propre, la prégnance du vivant et l'ancrage dans le monde physique, réel, sont fortement marqués dans les définitions. Le caractère nocif et dangereux s'applique à des organismes ou des êtres animés. Au sens figuré (*"Qui distille plus ou moins ouvertement des propos médisants ou dépréciateurs")*, si *toxique* peut se rapporter non à une substance mais à un message, la cible reste toujours un individu ou une communauté d'individus.

Seule la définition par métonymie, "*Qui est à l'origine ou est la cause d'une pollution, d'une intoxication*", pourrait correspondre au sens de *toxique* en contexte de crise. Mais, d'une part, cela nécessiterait d'interpréter *intoxication* selon un sens métaphorique non répertorié dans le TLFi et *pollution* à partir de sa définition par extension, d'autre part, les exemples d'illustration[64] de cette définition, sur l'*industrie toxique* et l'*alimentation toxique*, n'orientent pas vers une telle interprétation.

➔ **TOXIQUE**, subst. masc. et adj.
I. *Subst. masc., CHIM., MÉD., BIOL.* Produit d'origine animale, végétale ou minérale qui provoque l'intoxication, la destruction d'un organisme vivant.
Au fig. Ce qui agit négativement sur l'individu, son psychisme ou son physique.
II. *Adjectif*
A. *CHIM., MÉD., BIOL.* Qui est ou qui contient un poison.
En partic. [En parlant d'un produit alimentaire] Qui contient des toxines.
-- *Au fig.* Qui distille plus ou moins ouvertement des propos médisants ou dépréciateurs.
B. *PHYSIOL., PHARMACOL.*
1. Qui provoque un empoisonnement, une intoxication.
2. *Dose, équivalent toxique*. Quantité de produit à partir de laquelle celui-ci devient un poison pour un organisme.
3. Qui est dû à la présence d'un poison dans l'organisme, à l'action d'un poison sur un organe, un tissu.
C. *P. méton.* Qui est à l'origine ou est la cause d'une pollution, d'une intoxication.

Figure III.1.2 : Définitions associées à l'entrée toxique dans le TLFi

[63] Un titre *illiquide* est un titre financier qu'il est difficile de revendre, de négocier sur les marchés (site spécialisé *abcbourse*, http://www.abcbourse.com/apprendre/lexique.aspx?s=i). L'adjectif *illiquide* est à caractère néologique, de fait, il est actuellement absent d'un certain nombre de ressources lexicographiques. Pour le caractère néologique d'*illiquide*, on peut par exemple se reporter au blog *Langue Sauce Piquante* (http://correcteurs.blog.lemonde.fr/2009/03/09/illiquide/).

[64] Exemples de la définition par métonymie de *toxique* : "Les *principales industries toxiques que nous ayons à connaître sont celles qui exposent l'ouvrier à l'intoxication par le plomb, le mercure, l'arsenic, le cuivre* (MACAIGNE, *Précis hyg.*, 1911, p. 312). *Il existe un cornage par paralysie récurrentielle, accompagné de phénomènes paraplégiques; il disparaît avec la cessation de l'alimentation toxique et ne peut être regardé comme rédhibitoire* (BRION, *Jurispr. vétér.*, 1943, p. 246).".

Dans le corpus du *Monde Diplomatique*, les emplois de *toxique* se rapportent aux domaines biologique ou chimique. Ils renvoient à des questions d'impact industriel, de pollution de l'environnement ou de risques pour la santé. Les principaux substantifs qualifiés par *toxique* sont *déchets, gaz*, *substances, émissions, produits, décharges*. Les contextes d'apparition mettent en évidence des sens en adéquation avec les définitions du dictionnaire. *Toxique* ne qualifie jamais un substantif se rapportant au domaine financier.

1.2.2 Nom propre *Outreau* : un nouveau sens par stratification

Le nom propre *Outreau* correspond, en 2002, à un nom de ville du Pas-de-Calais. Il acquiert un sens stabilisé de « parangon des scandales judiciaires » suite aux événements qui s'y sont déroulés.

Le nouveau sens d'*Outreau* se construit progressivement, il évolue par strates. (Lecolle, 2007b) propose de distinguer cinq périodes pour observer l'évolution de sens et elle dégage plusieurs dimensions sémantiques qui apparaissent au cours des périodes d'observation, ou sont au contraire éliminées, atténuées ou modifiées. Cinq catégories principales recouvrent les évolutions de sens.

- La *dimension locative* caractérise le sens d'*Outreau* en période 1. Elle est apparente à travers diverses facettes : emplacement géographique, structure urbaine, ou collectif propre au lieu (habitants). Même si le sens locatif d'*Outreau* ne disparaît jamais complètement, d'autres sens le supplantent progressivement.
- La *dimension policière et judiciaire* est présente aux cinq périodes, mais sous différentes formes. La période 1, moins marquée, se positionne en amont de la procédure pénale, elle recouvre des aspects policiers (arrestations) et l'enclenchement de la procédure (mise en examen). Les notions de réseau pédophile et d'inceste y sont très présentes, ainsi qu'en période 2. Le traitement judiciaire s'étale des périodes 2 à 4, avec le déroulement du procès de Saint-Omer en période 2, l'attente du verdict en période 3 et le verdict en période 4. La période 5, qui correspond au procès en appel et à la commission d'enquête parlementaire consécutive, porte à nouveau sur la procédure judiciaire mais aussi sur l'institution judiciaire comme objet d'étude en raison de ses dysfonctionnements, ce qui ajoute une dimension politique.
- *L'émotion populaire* est aussi sous-jacente dans le sens affecté à *Outreau*. Son influence dans le déroulement de l'affaire est dénoncée en période 5, on peut donc supposer qu'elle influe implicitement sur le sens d'*Outreau* aux périodes précédentes, à travers une condamnation.
- Les sens de *fiasco judiciaire* et d'*« erreur judiciaire par excellence »* sont caractéristiques de la période 5, même si l'erreur judiciaire émerge déjà aux périodes 2, 3 et 4.
- La *dimension politique*, absente initialement, est surtout présente en période 5 avec l'ouverture d'une enquête parlementaire, mais apparaît avec la prise de recul sur les dysfonctionnements du système judiciaire.

Outreau constitue un cas limite par rapport aux objets d'étude que nous avons définis. En effet, formellement, il s'apparente à une néologie de forme : il n'a pas de sens codé dans le TLFi car il s'agit d'un nom propre. Le TLFi étant un dictionnaire de langue, il n'existe pas d'entrée pour les noms propres. Un dictionnaire encyclopédique, en revanche, pourrait intégrer des noms propres, dont des noms de ville, et il permettrait de ramener la néologie associée à *Outreau* à de la néologie sémantique. Un traitement automatique construit sur les entrées du TLFi ne peut donc affecter d'ancien sens à *Outreau*, il retourne un signifié vide. Cependant, l'étude d'une unité lexicale dotée d'un signifié vide reste valide sur plusieurs plans : il est toujours possible d'étudier les variations lexicales propres aux textes et de

procéder à l'enrichissement du sémème. Seule la reconfiguration du sémème sort du champ d'étude : elle n'a aucune pertinence en l'absence de traits sémantiques à reconfigurer. Pour résoudre ce problème, nous avons utilisé un artifice pour nous ramener à la néologie sémantique et à notre champ d'étude : une entrée a été élaborée à partir des propositions de M. Lecolle. Cette entrée comporte deux définitions, une correspondant à l'ancien sens, l'autre au nouveau sens :

➔ **OUTREAU**
I. Ville française du Pas-de-Calais
II. Erreur judiciaire liée à la découverte et croyance en l'existence d'un réseau pédophile, puis à la réfutation publique de cette croyance.

Figure III.1.3 : Définitions proposées pour les ancien et nouveau sens d'Outreau

L'ancien sens de ville française est analogue au sens codé. Le nouveau sens est introduit pour d'autres raisons, liées au type d'analyse qui sera effectué sur *Outreau*. La première étape d'un traitement est de contraster deux périodes seulement. Cette étape ne sera pas réalisée, on effectuera directement l'analyse de l'évolution progressive dans le temps, à partir du découpage en cinq périodes. Le contraste de deux périodes a pour vocation de fournir des traits sémantiques candidats pour appartenir au nouveau sens. L'évolution progressive sur plusieurs périodes est destinée à affiner les résultats obtenus à l'issue du contraste, d'une part pour valider les candidats, d'autre part, pour nuancer leur importance relative et pour préciser la façon dont ils se combinent au sens codé. L'étude d'une telle définition permet donc d'étudier la reconfiguration du sémème (en terme d'activation et inhibition), ainsi que de valider les nouveaux traits sémantiques et de les articuler au sens codé.

1.2.3 Économie réelle

La notion d'*économie réelle* renvoie à une partie de l'économie associée aux échanges de biens, de services, de capitaux ou de travail, ancrée dans une dimension matérielle et industrielle, par opposition à une dimension spéculative. La lexie compte 176 occurrences dans le corpus 'Crise financière' (pour une présentation du corpus, *cf.* chapitre II.1, 2.5.1). À la lecture des paragraphes contenant *économie réelle*, la crise économique apparaît comme une pathologie contagieuse ou comme une catastrophe naturelle se propageant de la sphère financière, considérée comme virtuelle, à la sphère industrielle, correspondant à l'économie dite réelle.

1.3 Aperçu général de la distribution des cibles dans les expériences illustratives

Les cibles présélectionnées seront utilisées alternativement au cours des différentes étapes du processus. Le tableau ci-dessous donne une vue d'ensemble de la répartition des expériences selon le stade du traitement. Il est organisé en sept colonnes. La colonne I précise le niveau de granularité de l'analyse (supra-lexical, lexical, infra-lexical) . Les colonnes II à IV portent sur l'objet et la finalité de l'analyse :

- en colonne II est précisée la partie du sémème de la cible sur laquelle on se focalise. Comme on procède par granularité décroissante, les sèmes observés dans un premier temps sont les domaines du sens codé (sèmes mésogénériques), ceux observés dans un second temps sont les traits sémantiques autres que les domaines (sèmes microgénériques ou spécifiques) ;
- en colonne III sont décrits les observables. Pour le premier niveau de granularité, les observables sont des domaines, observés de façon générale puis de façon relative à la cible (distinction entre les domaines associés au sens codé ou non). Pour les niveaux de granularité plus fins, on observe d'abord les unités lexicales, puis les traits sémantiques ;

- en colonne IV, on précise les différents axes d'analyse, dont la finalité varie. Certains axes portent sur la reconfiguration du sémème, d'autres sur son enrichissement, d'autres encore sur la constitution de classes, etc.

Les colonnes V à VII décrivent les supports des expériences illustratrices, c'est-à-dire :

- en colonne V le sous-corpus où émerge le nouveau sens ;
- en colonne VI, le corpus par rapport auquel le sous-corpus est contrasté ;
- en colonne VII, les cibles observées dans chaque corpus et sous-corpus.

La dernière colonne indique la section où l'expérience illustrative est présentée.

<table>
<tr><th>I</th><th>II</th><th>III</th><th colspan="3">IV</th><th>V</th><th>VI</th><th>VII</th><th>VIII</th></tr>
<tr><th>Niveau</th><th>Sémème de la cible</th><th>Observables (granularité)</th><th colspan="3">Analyse</th><th>Sous-corpus</th><th>Corpus</th><th>Cibles</th><th>Section</th></tr>
<tr><td rowspan="2">Supra-lexical</td><td rowspan="2">Domaines (sèmes mésogénériques)</td><td>Domaines textuels (domaines donnés)</td><td colspan="3">• ∃ évolution?
• Domaines émergents
• Régularité d'évolution
• Structuration</td><td>Factiva (vois., pér)</td><td>Factiva</td><td>tsunami toxique tablette moléculaire (numérique)
Témoins: dangereux délétère tempête raz-de-marée</td><td>2.2</td></tr>
<tr><td>Domaines relatifs à la cible (domaines projetés)</td><td colspan="3">Reconfiguration Enrichissement (domaniaux)</td><td>Crise (vois.)</td><td>Crise
MD (vois.)</td><td>toxique</td><td>2.3</td></tr>
<tr><td rowspan="4">Lexical</td><td rowspan="4"></td><td rowspan="4">Unités lexicales</td><td rowspan="3">Sélection et pondération (cooc. d'ordre 1)</td><td>confrontation de domaines</td><td>Crise (vois.)</td><td>Crise
MD (vois.)</td><td>toxique</td><td>3.2.1</td></tr>
<tr><td>au sein d'un domaine</td><td>Crise (vois.)</td><td>Crise</td><td>toxique</td><td>3.2.2</td></tr>
<tr><td>par période</td><td>Outreau (vois., pér)</td><td>Outreau</td><td>Outreau</td><td>3.2.3</td></tr>
<tr><td colspan="2">Sélection et pondération (cooccurrents d'ordre 2)</td><td>Crise (vois.)</td><td>Le Monde (vois.)</td><td>bouclier pourri toxique tempête tsunami
Témoin : actif</td><td>3.3</td></tr>
<tr><td rowspan="5">Infra-lexical</td><td rowspan="5">Traits autres que les domaines (sèmes microgénériques ou spécifiques)</td><td rowspan="5">Traits sémantiques</td><td colspan="2" rowspan="2">Reconfiguration du sémème (cooccurrents d'ordre 1)</td><td>Crise (vois.)</td><td>Crise
MD (vois.)</td><td>toxique</td><td rowspan="2">4.1.3</td></tr>
<tr><td>Outreau (vois, pér)</td><td>Outreau</td><td>Outreau</td></tr>
<tr><td rowspan="3">Enrichissement (sur les cooccurrents d'ordre 1)</td><td>∃ apport des traits?</td><td>Crise (vois.)</td><td>Crise</td><td>économie réelle</td><td rowspan="3">4.1.4</td></tr>
<tr><td>Renforcement sém/lex</td><td>Outreau (vois., pér)</td><td>Outreau</td><td>Outreau</td></tr>
<tr><td>Nouveauté sém/lex</td><td>Crise (vois.)</td><td>Crise
MD (vois.)</td><td>toxique</td></tr>
<tr><td colspan="3">Constitution de classes (sur les cooccurrents d'ordre 2)</td><td>Crise (vois.)</td><td>Le Monde (vois.)</td><td>bouclier pourri toxique tempête tsunami
Témoin : actif</td><td>4.2.2</td></tr>
</table>

Tableau III.1.4 : Expériences illustratives des différentes étapes de la procédure[65]

[65] Légende – MD : corpus du *Monde Diplomatique* ; Crise : corpus 'Crise financière' ; renforcement sém/lex : renforcement apporté par les traits sémantiques aux unités lexicales ; nouveauté sém/lex : enrichissement apporté par les traits sémantiques distinct de ce qui apparaît au niveau des unités lexicales ; (vois.) : restriction aux

2. Niveau supra-lexical : des domaines en corpus aux domaines du sens codé

La structure principale du processus est déterminée par le critère d'évolution du global vers le local : la caractérisation du sens se construit à partir d'unités de granularité décroissante. Dans ce paragraphe, nous étudions des descripteurs de niveau global. Nous parlerons de domaines. Les domaines sont une restriction des descripteurs supra-lexicaux, qui incluent également les thèmes et les isotopies globales.

Nous cherchons à :

- préciser certaines des caractéristiques des descripteurs supra-lexicaux (sous-section 2.1) ;
- proposer des façons d'associer au corpus différentes étiquettes de domaines (sous-section 2.2) ;
- définir comment exploiter les domaines pour caractériser le sens d'une cible lexicale en corpus (sous-section 2.3) ;
- articuler les domaines et leur profil en corpus à la partie du signifié correspondante, c'est-à-dire aux sèmes mésogénériques. Dans notre cadre, les sèmes mésogénériques sont assimilés aux domaines du TLFi (sous-section 2.4).

Deux types de domaines, issus du TLFi ou propres au corpus, peuvent être utilisés. Ces deux types de domaines ne sont pas nécessairement équivalents. Dans les cas où notre terminologie pourrait prêter à confusion, nous préciserons s'il s'agit des domaines issus du TLFi (domaines lexicographiques), ou des domaines qui correspondent aux descripteurs supra-lexicaux du corpus (domaines textuels).

2.1 Caractéristiques des unités supra-lexicales

2.1.1 Présence transverse à l'axe du temps

Pour observer la diffusion d'une unité lexicale au cours du temps relativement aux domaines, deux cas de figure se présentent :

- les domaines constituent le point fixe par rapport auquel on observe la diffusion de la cible lexicale ;
- la cible lexicale constitue le point fixe par rapport auquel on sélectionne un environnement textuel et on étudie l'évolution des différents domaines.

Dans le cas où les domaines sont le point fixe, la présence des domaines à chaque période paraît nécessaire pour rendre comparables les différentes périodes. Il n'est pas indispensable que la répartition des domaines soit équilibrée aux différentes périodes, une normalisation permet de rectifier les déséquilibres. Cependant, un relatif équilibre reste préférable[66].

Dans le cas où la cible est le point fixe, on peut observer une présence évolutive des domaines au cours du temps. Le critère de présence de la cible lexicale préside à la constitution du corpus ; si le nouveau sens se caractérise par un changement de domaine, le nouveau domaine peut être totalement absent à l'instant initial. Le corpus 'Outreau' de (Lecolle, 2007), utilisé dans nos expériences, correspond à ce cas de figure : il a été constitué

voisinages de la cible ; (pér) : restriction à une période.

[66] « Dans les études lexicométriques, il est prudent de faire en sorte que les textes que l'on réunit en un même corpus à des fins de comparaison aient des longueurs comparables » (Salem, 1988:107 cité par (Delanoé, 2010)). L'équilibre entre sous-corpus correspondant aux différentes périodes d'un découpage chronologique est donc souhaitable. Les études menées sur le corpus *Outreau* ont de ce fait posé certains problèmes au niveau de l'interprétation des résultats quantitatifs : le découpage en période est manuellement légitime, car sémantiquement cohérent, mais il présente des déséquilibres importants.

à partir d'articles journalistiques sélectionnés en fonction de la présence du nom propre *Outreau*. Cependant, l'évolution n'a de sens que si la base textuelle dont est extrait le corpus présente le même équilibre de domaines au cours du temps. Autrement dit, la répartition des domaines est supposée comparable aux différentes périodes dans la base textuelle originelle, avant que soit appliqué le filtre de présence de la cible. Dans le corpus 'Outreau', on observe par exemple que le domaine POLITIQUE y est absent initialement et apparaît dans les dernières périodes du corpus. L'émergence du domaine POLITIQUE a été jugée significative lors de l'analyse manuelle. Cette conclusion repose sur l'hypothèse que, dans la base textuelle d'origine, sans le filtre relatif à la cible lexicale, le domaine POLITIQUE avait autant de chances d'être présent à la période initiale qu'à la période finale. Autrement dit, on suppose implicitement que la base de données dont est extrait le corpus présente une certaine homogénéité de répartition des domaines aux différentes périodes.

2.1.2 Nombre réduit

Les unités supra-lexicales sont destinées à fournir des informations de granularité sémantique grossière. On ne recherche pas des nuances qualitatives, mais les grandes lignes qualitatives. Les unités supra-lexicales ont pour finalité de servir de dénominateur commun aux unités lexicales ou aux traits sémantiques, qui peuvent quant à eux comporter une grande diversité. Leur rôle est de contribuer à la résolution de l'homonymie et à de la désambiguïsation grossière. À ce stade, un nombre réduit d'unités supra-lexicales répond à notre objectif, tandis qu'un nombre trop important introduirait prématurément des nuances et une dispersion de l'information.

Empiriquement, on constate que les clés de structuration de l'information sémantique (domaines lexicographiques, rubriques de journaux, etc.) sont de l'ordre d'une à quelques dizaines au niveau le plus global. Ainsi, dans le portail de recherche assistée, le TLFi propose en ligne 21 domaines principaux, qualifiés de *centres d'intérêt* et ensuite subdivisés en domaines plus précis[67]. Les grands domaines du *Petit Larousse* sont au nombre de 33 (Khalid, 2008:66). Au niveau des ressources textuelles, le journal du *Monde* en ligne (www.lemonde.fr) propose comporte 32 sous-rubriques assimilables à des domaines (telles qu'Économie, POLITIQUE, GASTRONOMIE, FOOTBALL), réparties dans 3 rubriques principales[68] (ACTUALITES, SPORT et LOISIRS).

Les constats empiriques sur le nombre de divisions principales invitent à préférer un petit nombre d'étiquettes pour caractériser le global, qui reste de l'ordre de quelques dizaines, de préférence en restant en deçà ou aux alentours de 30.

2.1.3 Généricité

Les étiquettes de domaines doivent respecter un critère de généricité : un domaine doit être une caractérisation sémantique commune à plusieurs unités lexicales. De plus, elles se doivent d'en préciser le sens et de contribuer à leur désambiguïsation. Autrement dit, les étiquettes de domaines sont des aides pour qualifier le sens d'une unité lexicale et pour regrouper sous un même chapeau différentes unités.

[67] Ils sont accessibles au niveau du module de recherche assisté. Celui-ci propose une structure de domaines. Cette strcture est accessible sur le site du TLFi atilf.atilf.fr, dans l'onglet "recherche assistée", à partir du lien cliquable "Cliquez ici pour choisir la discipline" de la troisième rubrique.

[68] Les sous-rubriques ont été associées à des domaines si elles validaient les tests suivants : vérification de la relation *"traite de"* et *"n'est pas de type petite annonce"*. Ainsi, la sous-rubrique ECONOMIE de la rubrique ACTUALITES vérifie les deux relations ; la sous-rubrique CARNET ne vérifie par la première relation ; la sous-rubrique ANNONCES AUTO ne vérifie pas la deuxième relation.

2.1.4 Proximité des domaines textuels avec les domaines lexicographiques

Deux types de domaines sont utilisés : les domaines textuels et les domaines lexicographiques. Les domaines textuels sont caractéristiques d'un corpus et ils peuvent être donnés au départ, résulter d'une annotation en traits sémantiques ou être construits (*cf.* chapitre II.1, 1.2). Les domaines lexicographiques sont ceux associés au sens codé, ils proviennent du dictionnaire. Les domaines textuels sont utilisés pour structurer les significations de la cible en corpus. Les domaines lexicographiques structurent le contenu sémantique de la cible hors contexte et par rapport au lexique de la langue.

Les lignes de structuration principales du corpus, assimilées à des domaines, sont le reflet de caractéristiques propres au corpus. Le découpage proposé résulte d'une vision du monde particulière, associée aux usages dont le corpus est représentatif. Les domaines lexicographiques sont aussi un découpage qui témoigne d'une certaine représentation du monde, propre à l'époque de la rédaction du TLFi et de nature ontologique[69]. Il est presque certain qu'ils ne coïncideront pas parfaitement avec les domaines du corpus. Cependant, pour permettre le lien avec les sens codés, il est indispensable de pouvoir établir des corrélations entre les domaines textuels et lexicographiques.

Sous réserve que les découpages textuels et lexicographiques ne soient pas trop éloignés, la correspondance entre domaines lexicographiques et textuels peut être effectuée a priori, par projection des domaines dans le corpus, ou a posteriori, à partir d'un rapprochement manuel ou de procédures automatisées.

2.2 Établir le profil domanial du corpus

La caractérisation domaniale du corpus peut être obtenue selon plusieurs procédés : à partir d'informations données dans le corpus (2.2.1), à partir d'une projection des domaines lexicographiques (2.2.2) ou à partir d'une construction non triviale exploitant les données et métadonnées du corpus (2.2.3).

2.2.1 Les domaines donnés

Les domaines donnés sont des étiquettes préalablement affectées aux documents constitutifs du corpus, ce sont des clés de structuration de la base textuelle dont est issu le corpus.

Pour un corpus journalistique, les domaines pourront être définis à partir des rubriques (par exemple, pour le journal *L'Humanité,* les rubriques A LA UNE, POLITIQUE, SOCIAL-ECONOMIE, SOCIETE, MONDE, CULTURE, SPORTS, MEDIAS pourront être considérées comme des étiquettes de domaines). Le rôle structurant des domaines a été démontré notamment dans l'étude de (Illouz *et al*., 1999) sur un corpus issu du journal *Le Monde*[70]. Pour une base de données

[69] Lors de la rédaction du TLFi, les emplois spécialisés dépendaient d'une liste de domaines préalablement établie, en fonction de découpages présents dans des ressources de type encyclopédique : « Le travail spécifique pour cette partie du corpus a consisté à nous donner, préalablement à toute sélection, une liste, de type encyclopédique, des domaines dans lesquels pouvait se distribuer la production des textes de cette section de notre documentation ; liste qui serait elle-même le reflet des activités principales de la nation ou de l'ethnie française au cours des deux siècles considérés. Pour des raisons pratiques, il ne pouvait s'agir que d'une liste d'ensembles à très large extension, nous contentant, pour les sous-ensembles, de renvoyer aux documents utilisés pour l'établissement de cette liste : la Classification décimale universelle (Édition abrégée, La Haye, F.I.D., 1958), les plans de l'Encyclopédie française, de l'Encyclopédie de la Pléiade, etc. » (Imbs, 1971:XXIV). Cette liste originelle a connu quelques amendements, à partir de propositions de rédacteurs, mais la structure principale est conditionnée par des principes plus ontologiques que linguistiques.

[70] (Illouz *et al.*, 1999) extraient des parties de corpus (tranches de mots d'une taille donnée), les structurent à partir de leurs profils lexicaux (application de la méthode de Sammon, projection non-linéaire en deux dimensions d'un espace multi-dimensionnel (Sammon, 1969)) et observent si la structure émergente reflète des

journalistiques telle que Factiva, les rubriques propres à chaque journal diffèrent, mais une nomenclature de *sujets* (tels que ARTS ET SPECTACLES, POLITIQUE/RELATIONS INTERNATIONALES, ENVIRONNEMENT, SANTE ou encore SCIENCE ET TECHNOLOGIE) est proposée de façon commune à toutes les sources journalistiques et peut servir de définition des domaines.

Les domaines donnés ont l'avantage d'être assez directement accessibles, d'où une économie au niveau des traitements appliqués au corpus. De plus, s'appuyer sur ces domaines revient à respecter une cohérence qui a conditionné la production du discours considéré.

En revanche, les limites suivantes se présentent, qui justifient l'intérêt des domaines projetés ou construits :

- Le corpus ne dispose pas nécessairement d'étiquettes assimilables à des domaines, ou son panel d'étiquettes n'est pas homogène. C'est le cas du corpus 'Crise financière'. Celui-ci est construit à partir de deux sources journalistiques distinctes : *Le Figaro* et *L'Humanité*, dont les rubriques respectives ne coïncident pas.
- Les domaines donnés tels que les rubriques correspondent bien à une structure propre aux ressources journalistiques, mais ils ne définissent pas nécessairement le découpage optimal (Illouz *et al.*, 1999) : il peut par exemple y avoir une forte hétérogénéité interne à une rubrique, qui devrait être décomposée en sous-ensembles, ou des regroupements possibles entre des rubriques ou parties de rubriques disjointes. Le découpage optimal est fonction de ce que l'on cherche à observer. Dans notre cadre, le découpage supra-lexical doit pouvoir être mis en correspondance avec les domaines du *TLFi*.

2.2.2 Les domaines projetés

La répartition des documents textuels peut être établie à partir des domaines lexicographiques.

La démarche générale consiste d'abord à annoter sémantiquement le corpus en domaines, c'est-à-dire à ajouter aux unités lexicales les étiquettes de domaines qui leur sont associées dans la ressource lexicographique. Ensuite, des traitements numériques permettent de sélectionner des domaines pertinents pour chaque document. On peut ainsi extraire les domaines saillants ou évaluer leur degré de saillance à l'aide d'indices de significativité statistique. Différents profils peuvent alors être obtenus, selon le critère appliqué, par exemple selon qu'on choisit :

- d'affecter le domaine le plus significatif à chaque document. On retrouve le même type de profil que pour les domaines donnés où, généralement, un document est associé à une unique rubrique ;
- d'affecter un nombre variable de domaines à chaque document, sans pondération. Ceci revient à trier les documents en deux groupes (caractérise le document / ne caractérise pas le document). Ce tri peut se faire en fonction d'un seuil. Les valeurs affectées aux domaines servent à trier les domaines en fonction du seuil, mais elles ne sont pas conservées à l'issue du tri, une fois le groupe de domaines sélectionné ;
- d'affecter un ensemble pondéré de domaines à chaque document. Les pondérations permettent de hiérarchiser les domaines selon leur degré de significativité dans le document, la caractérisation est donc plus nuancée. Un seuil peut toujours être utilisé, mais les valeurs affectées aux domaines sont conservées après application du seuil.

proximités entre textes d'une même rubrique.

Des travaux en cours au sein du laboratoire[71] témoignent de la faisabilité d'une telle caractérisation des textes en domaines. Ces travaux ont donné lieu à des expériences de profilage domanial par annotation de corpus. Le profilage repose sur une annotation des textes constitutifs d'un corpus journalistique (issu du *Monde Diplomatique*) à partir de domaines du TLFi. Des poids sont ensuite affectés aux domaines à travers trois jeux de comparaison : comparaison des textes au corpus journalistique, du corpus journalistique à un corpus de référence et des textes au corpus de référence. Les poids sont obtenus par un calcul de spécificités (Lafon, 1984) , puis soumis à un critère couplant un seuil et une fonction booléenne basique afin de décider entre la sélection ou l'exclusion d'un domaine. Cette méthode offre ainsi une possibilité d'affecter aux textes des domaines pondérés, en procédant par exclusion plutôt que par sélection.

L'intérêt d'une projection est qu'un lien direct est établi entre le corpus et la ressource lexicographique : les expériences précédemment citées fournissent une table de correspondance entre les documents textuels et les domaines du TLFi.

Cependant, la projection a priori est aussi une prise de risque : on suppose implicitement que la méthode de projection est efficace et que la caractérisation résultante en domaines a une certaine qualité. Si ce n'est pas le cas, la question suivante se pose : peut-on faire émerger des contrastes pertinents si la structure qui sert de base est de mauvaise qualité ? Autrement dit, il faut que la déformation induite par la projection ne soit pas pénalisante. Or le rôle structurant des domaines projetés n'offre pas les garanties des domaines donnés, dont la validité a déjà été mise en évidence (Illouz *et al*., *cf*. supra). Le problème du couplage entre le corpus et la ressource de référence n'est que repoussé si on renonce à une projection en amont de tout traitement, mais cela permet d'obtenir des différences entre emplois de la période initiale et emplois ultérieurs, à partir d'informations propres au corpus. Ajoutons que la projection de domaines a un coût en termes de traitements informatiques, notamment dans de grands corpus (coût d'une annotation sémantique en domaines puis de l'extraction des domaines statistiquement saillants pour chaque document du corpus).

2.2.3 Les domaines construits

On peut souhaiter s'appuyer sur la structure propre au corpus sans utiliser d'étiquettes préalablement données, par exemple parce que les étiquettes de domaines manquent d'homogénéité, parce qu'elles ne qualifient pas le contenu textuel de façon suffisamment précise ou pertinente, etc., ou encore parce que de telles étiquettes sont absentes. Dans ce cas, il est possible de recourir à diverses techniques pour construire des domaines. Ceux-ci peuvent être obtenus à partir de regroupements thématiques de documents, comme ceux de la plateforme ProxiDocs (Roy et Beust, 2005) : l'interaction entre un utilisateur et le logiciel ThemeEditor (Beust, 2002) génère des classes sémantiques, dont la distribution permet de générer des regroupements thématiques. Il est aussi possible de s'appuyer sur l'extraction de mots-clés, comme l'effectuent (Rossignol et Sébillot, 2002), puis de mettre en relation ces mots-clés avec les domaines lexicographiques. Les grandes lignes de caractérisation peuvent prendre une forme lexicalisée, avec un ou des mots-clés par exemple, mais pas nécessairement. Ainsi, dans le cas d'une classification thématique, l'étiquette correspondant à un groupe de documents n'est pas forcément explicite. Dans cette situation, un domaine correspond à un sous-ensemble sémantiquement cohérent.

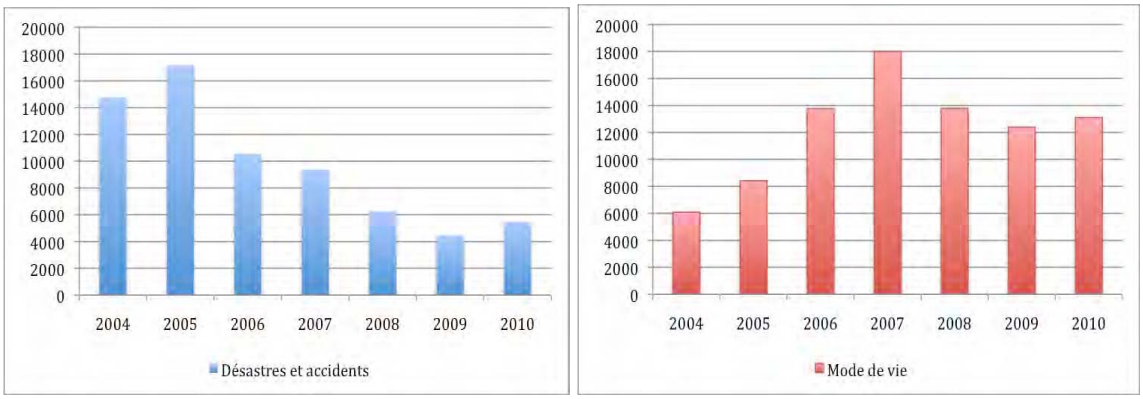
[71] Un protocole de caractérisation thématique par annotation domaniale a été développé par M. Grzesitchak dans le cadre de ses travaux sur le profilage thématique. Ces travaux ne sont pas finalisés à l'heure actuelle.

2.2.4 Expériences réalisées

Dans nos expériences, les domaines ont été obtenus selon les différents procédés précédemment décrits :

- Le corpus 'Crise financière' est un corpus thématique, rattaché aux domaines de l'économie et de la finance. La caractérisation domaniale du corpus est construite : elle est déduite des critères de sélection, à savoir la pertinence des documents par rapport au thème de la crise financière, évaluée manuellement, ainsi qu'un filtrage partiel des documents par mot-clé (en l'occurrence, le mot-clé *crise* a servi à sélectionner les articles d'une des deux sources journalistiques).
- Le corpus 'Outreau' est un corpus construit autour d'un mot-clé. Le profil domanial du corpus a été étudié par projection des étiquettes de domaines du TLFi dans les documents.
- Le corpus 'Factiva' a utilisé des domaines donnés, issus des métadonnées du site Factiva (les SUJETS du moteur de recherche avancée disponible sur Factiva).

Le profil domanial d'un corpus est établi lorsqu'à chaque document ont été associés le ou les domaines correspondants (domaines donnés, projetés ou construits), éventuellement affectés d'une pondération. De plus, chaque document est défini par sa position chronologique, c'est-à-dire qu'il est associé à une période de temps. On en déduit un profil domanial du corpus à chaque période. Le profil domanial peut être représenté de façon détaillée document par document ou de façon synthétique à l'échelle du corpus. La représentation détaillée doit être conservée pour établir le profil domanial d'éventuels sous-corpus, tels que le sous-corpus des voisinages d'une cible : les domaines qui caractérisent un sous-corpus se déduisent des documents qui constituent ce sous-corpus. Le profil synthétique définit la répartition d'ensemble des domaines dans le corpus. Il est utile pour normaliser des résultats, notamment en cas de répartition déséquilibrée des domaines. Ainsi, dans Factiva, les sources journalistiques évoluent au cours du temps, de même que le nombre d'articles affectés à un sujet donné (croissance marquée, faible, décroissance, …). Dans l'exemple ci-dessous[72], le nombre de documents du sujet MODE DE VIE et celui du sujet DESASTRES ET ACCIDENTS connaissent des évolutions marquées et inverses (croissance dans le premier cas, décroissance dans le second cas).



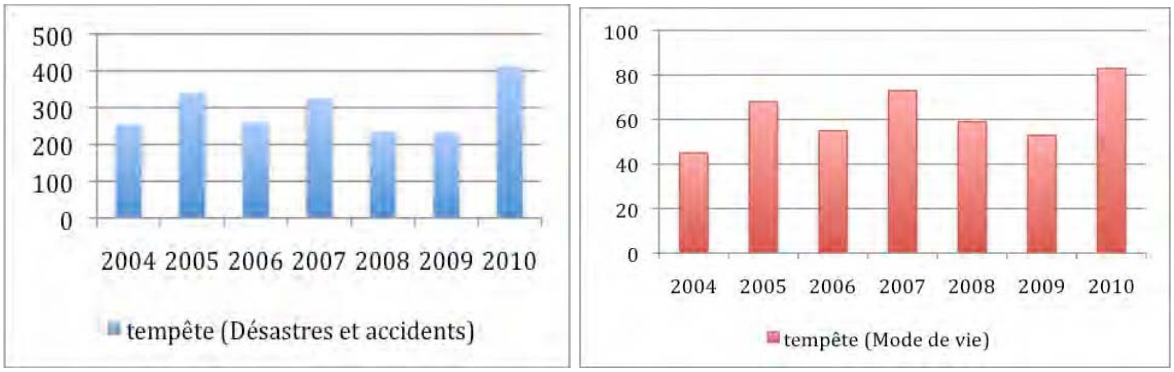
Figures 6.5 a et b : Nombre total de documents par période pour les domaines (a) DESASTRES ET ACCIDENTS et (b) MODE DE VIE dans le corpus Factiva

Sans normalisation, des phénomènes néologiques risquent de ne pas être identifiés, ou d'être repérés à tort. Ainsi, un nombre d'occurrences constant au cours du temps dans DESASTRES ET ACCIDENTS correspond à un accroissement relativement à la taille du corpus, donc une

[72] Les sources journalistiques sont *Le Figaro*, *Libération, Ouest-France, L'Equipe, L'Express, Les Echos, L'Expansion, La Tribune*, sur la période de 2004 à 2010.

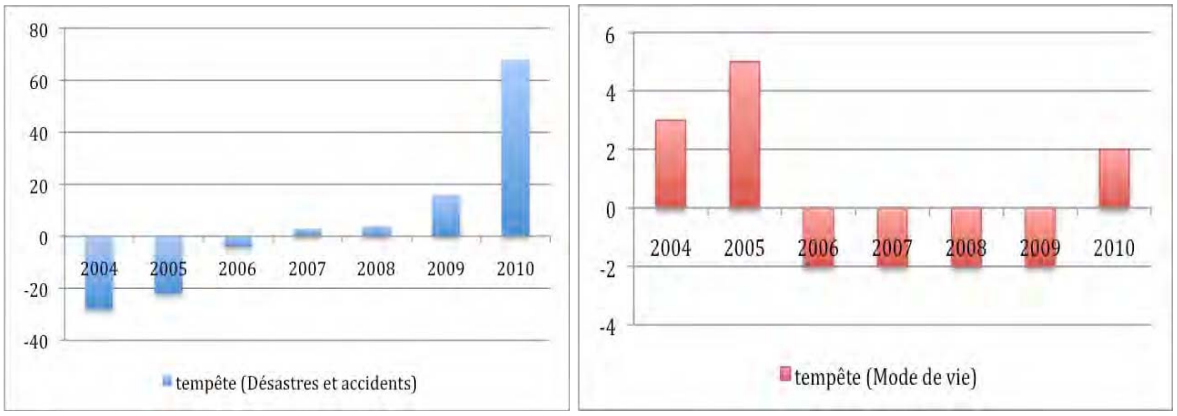
néologie potentielle, mais, sans normalisation, la diffusion ne sera pas identifiée ; à l'inverse, toute unité lexicale présente proportionnellement à la taille des sous-corpus périodiques de la catégorie MODE DE VIE semblera se diffuser au cours du temps – donc être candidate à la néosémie.

À ce titre, l'exemple de *tempête* est éclairant. Dans les domaines MODE DE VIE et DESASTRES ET ACCIDENTS, les fréquences donnent l'impression que l'évolution est à peu près similaire, avec une stagnation ou, dans le cas de MODE DE VIE, un léger accroissement entre 2004 et 2009, puis un accroissement plus marqué en 2010.



Figures 6.6 a et b : Nombre de documents par période contenant tempête pour les domaines (a) DESASTRES ET ACCIDENTS et (b) MODE DE VIE dans le corpus Factiva

Une normalisation est réalisée de façon à ajuster l'évolution des domaines au voisinage de *tempête* en fonction de l'évolution des domaines dans l'ensemble du corpus. Pour cela, on effectue un calcul des spécificités pour chaque domaine, dont les paramètres sont le nombre d'occurrences de la *tempête* par période, le nombre d'occurrences total, le nombre de documents à la période considérée et le nombre de documents total, à domaine fixé. À l'issue de cette normalisation, un accroissement régulier de *tempête* apparaît dans DESASTRES ET ACCIDENTS, ce qui témoigne d'une diffusion ; dans MODE DE VIE, la présence de *tempête* ne varie que faiblement, voire diminue légèrement.



Figures 6.7 a et b : Spécificités calculées à partir du nombre de documents par période contenant tempête pour les domaines (a) DESASTRES ET ACCIDENTS et (b) MODE DE VIE dans le corpus Factiva

Disposer d'un profil domanial du corpus au cours du temps permet donc d'éviter ce type de biais : la diffusion au cours du temps sera observée relativement au nombre total de documents associés à chaque sujet dans l'ensemble du corpus.

En bref, pour tout triplet (domaine, document, période), on disposera d'une valeur associée, correspondant au lien qu'entretient le domaine avec le document (en termes de présence/absence ou de degré d'association) et à l'appartenance du document à la période. Cette distribution sur les triplets (domaine, document, période), intermédiaire, sert à construire la distribution sur les couples (domaine, période) : on disposera d'une valeur associée à tout couple, qui reflète la répartition des domaines période par période et qui rend ces périodes comparables.

2.3 Établir le profil domanial de la cible lexicale

Pour établir le profil d'une unité lexicale en fonction des domaines, on exploite les fréquences de la cible relativement aux profils domaniaux des documents (empreintes de fréquence thématisées, *cf.* chapitre I.3, 2.1.2 et 2.2.3). Nous procédons en quatre étapes : la conservation ou le rejet de cibles potentielles selon l'évolution des domaines (2.3.1); l'identification de domaines émergents pour les cibles retenues (2.3.2) ; l'analyse détaillée de l'évolution dans le temps des domaines (2.3.3) ; la structuration des domaines (2.3.4).

2.3.1 Amender l'ensemble des cibles lexicales en fonction de leur évolution domaniale

La première étape consiste à vérifier que les cibles présélectionnées entrent bien dans notre champ d'investigation – c'est-à-dire qu'elles présentent une évolution domaniale (thématique dans un cadre plus général)[73].

Dans les illustrations qui suivent, les cibles présélectionnées sont *toxique, délétère, dangereux ; tsunami, tempête, raz-de-marée ; moléculaire ; tablette ; numérique*. Elles se répartissent en deux catégories :

- Les cibles effectivement pressenties ou identifiées comme néologiques, à savoir *toxique, tsunami, moléculaire* (gastronomie), *numérique* (extension d'emploi pour qualifier tout ce qui touche aux technologies de l'information et de la communication, comme les hautes technologies ou l'informatique ; cas probable de dédomanialisation, avec évolution d'emplois spécialisés dans les domaines mathématique, électronique et informatique vers un usage courant), *tablette* (pour les *tablettes numériques*).
- Les cibles non néologiques, utilisées comme témoins. Il s'agit de *délétère, dangereux, tempête, raz-de-marée*.[74]

À ce stade des analyses, un corpus découpé en deux périodes suffit. Les périodes du corpus 'Factiva' correspondent à des tranches annuelles, pour les années 2004 et 2010.

L'analyse a pour point de départ un découpage du corpus en sous-corpus par domaine et par période. Deux types d'indices peuvent être associés à chaque document : un indicateur binaire de présence/absence par document (indicateur utilisé dans les expériences illustratives de Factiva), ou la fréquence absolue de l'unité lexicale dans le document. À partir des indices associés aux documents et à partir du profil domanial des documents, on déduit un indice d'occurrence des cibles pour chaque couple (domaine ; période). Dans les expériences illustratives, l'indice affecté à un couple (domaine ; période) est le nombre d'articles appartenant au domaine et contenant la cible.

[73] L'ensemble des cibles présélectionnées peut être élargi à la totalité du corpus : toute unité est candidate. Il n'y a alors pas de véritable présélection. La sélection proprement dite se construira sur la base de l'évolution domaniale, non sur d'autres types d'indices. Une telle procédure est lourde, sa mise en œuvre expérimentale nécessite du temps et des moyens conséquents. Cela dépasse l'ambition de nos expériences, illustrations à petite échelle, mais cette piste mériterait de faire l'objet d'une étude de plus grande ampleur.

[74] Ces unités ont été choisies en rapport avec les cibles néologiques (*toxique* et *tsunami*). Elles forment des groupes d'unités sémantiquement proches, pour permettre une meilleure comparaison.

La distribution domaniale des cibles, c'est-à-dire le nombre de documents contenant la cible par période et par domaine, est ramenée au profil domanial du corpus. Pour illustrer les biais lorsqu'on n'utilise pas le profil général du corpus, considérons la distribution domaniale de *tablette (cible)* en 2004 et 2010, et plus particulièrement les domaines ARTS ET SPECTACLES et MODE DE VIE.

Distribution domaniale de *tablette*

	ARTS ET SPECTACLES	MODE DE VIE	Autres domaines
2004	36	23	/
2010	117	73	/

Tableau III.1.8 : Nombre de documents contenant* tablette *en 2004 et 2010 pour les domaines ARTS ET SPECTACLES et MODE DE VIE

Considérés isolément, c'est-à-dire sans le profil domanial du corpus, les indices associés à la cible pourraient amener à des conclusions hâtives. À période fixée, la présence de *tablette* est apparemment plus marquée dans ARTS ET SPECTACLES que dans MODE DE VIE, en 2004 comme en 2010. À domaine fixé, il y a apparemment un accroissement dans le temps, qui semble grosso modo de même importance dans ARTS ET SPECTACLES et dans MODE DE VIE. Le profil domanial général du corpus montre des déséquilibres, par rapport auxquels il convient de réajuster les résultats.

Profil domanial du corpus,
restreint à ARTS ET SPECTACLES et MODE DE VIE

	ARTS ET SPECTACLES	MODE DE VIE	Autres domaines
2004	49 228	6 079	x
2010	42 304	12 803	x

Tableau III.1.9 : Nombre de documents associés aux domaines ARTS ET SPECTACLES et MODE DE VIE en 2004 et 2010.

Les indices initiaux, c'est-à-dire le nombre de documents par domaine et par période contenant *tablette* du tableau (tableau III.1.8), sont remplacés par des valeurs de spécificités. Celles-ci sont d'abord calculées à période fixée, c'est-à-dire en fonction du nombre de documents sur l'ensemble des domaines à chaque période.

Spécificités calculées à période fixée

	ARTS ET SPECTACLES	MODE DE VIE	Autres domaines
2004	1	12	x
2010	3	13	x

Tableau III.1.10 : Spécificités à période fixée, calculées pour ARTS ET SPECTACLES et MODE DE VIE relativement à l'ensemble des domaines utilisés pour les expériences

Le calcul des spécificités appliqué à période fixée (tableau III.1.10) témoigne d'une présence relative de *tablette* dans MODE DE VIE plus significative que dans ARTS ET SPECTACLES. Dans un second temps, les spécificités sont calculées à domaine fixé, c'est-à-dire en fonction du nombre de documents sur l'ensemble des deux périodes pour chaque domaine.

Spécificités calculées à domaine fixé

	ARTS ET SPECTACLES	MODE DE VIE	Autres domaines
2004	-13	-1	x
2010	13	1	x

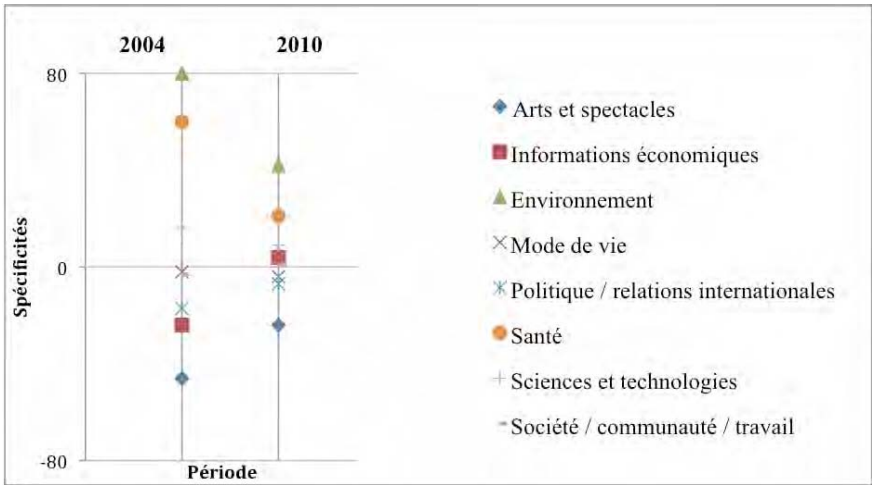
Tableau III.1.11 : Spécificités à domaine fixé, calculées relativement aux deux années 2004 et 2010 exclusivement

À domaine fixé (tableau III.1.11), l'accroissement est ténu pour MODE DE VIE, tandis qu'il est fortement marqué pour ARTS ET SPECTACLES que pour MODE DE VIE. Ainsi, si le profil domanial de *tablette* est ramené au profil domanial du corpus, on est amené à réviser les conclusions qui auraient pu être faites a priori.

Au-delà de cet exemple, une utilisation plus générale des fréquences des cibles en fonction de leur profil domanial et de celui du corpus pourrait se décliner comme suit. Le tri des cibles présélectionnées s'effectue sur critère d'*accroissement minimal* de la présence de la cible en deuxième période, dans au moins un domaine. Ce critère dépend d'un seuil de significativité statistique. Il convient également d'appliquer un critère de *présence minimale* pour filtrer les cibles dont la présence reste trop rare, quels que soient la période et le domaine considérés. Ce critère dépend d'un seuil de fréquence[75].

On propose de distinguer deux cas de figure, selon le nombre de domaines émergents.

Dans le premier cas, l'accroissement permet d'isoler un domaine particulier. C'est le cas avec *toxique*, pour lequel il y a un comportement très atypique en INFORMATIONS ECONOMIQUES, avec un accroissement marqué entre 2004 et 2010. Ce comportement atypique peut être observé de deux façons. Considérons d'abord la ventilation des domaines de *toxique* en 2004 et 2010 (figure III.1.12 ci-dessous).



[75] Les seuils de fréquence n'affectent pas les résultats des expériences à petite échelle présentées dans ce paragraphe, mais, dans des expériences de plus grande ampleur, ils permettraient un filtrage préalable d'unités. Celles-ci seraient considérées comme trop rares pour valider le critère de diffusion.

Figure III.1.12 : Ventilation des domaines associés à toxique en 2004 et en 2010[76]

Un domaine se distingue des autres : la position d'INFORMATIONS ECONOMIQUES évolue nettement par rapport aux autres domaines (passage de spécificités négatives à positives et évolution du rang relatif aux autres domaines de l'avant-dernière position à la quatrième position), tandis que ces derniers conservent à peu près la même configuration.[77]

La particularité d'INFORMATIONS ECONOMIQUES apparaît aussi en comparant les évolutions calculées à domaine fixé (tableau III.1.13) : l'indice de spécificité est supérieur de 23 pour ce domaine, alors que pour tous les autres domaines, il est inférieur à 4.

	ARTS ET SPECTACLES	INFORMATIONS ECONOMIQUES	ENVIRONNEMENT	MODE DE VIE	POLITIQUE – RELATIONS INTERNATIONALES	SANTE	SCIENCES ET TECHNOLOGIES	SOCIETE – COMMUNAUTE - TRAVAIL
Spécificités	3	**23**	-4	1	4	-2	0	2

Tableau III.1.13 : Saillance des domaines associés à toxique en 2010 relativement à 2004 (accroissement dans le temps)

L'émergence d'un domaine particulier peut donc s'observer de deux façons (figure III.1.12 et tableau III.1.13). À domaine fixé, il y a un accroissement significatif des emplois dans le domaine en deuxième période. Ou alors le poids relatif du domaine par rapport aux autres calculé à période fixée s'accroît de façon significative entre les deux périodes. Ce cas de figure est susceptible de renvoyer à une domanialisation.

Dans le second cas, l'accroissement affecte de façon à peu près similaire un ensemble des domaines. Ce cas de figure est celui de *numérique* (six domaines émergents), *tablette* (quatre domaines émergents) et *tsunami* (cas de figure moins net, deux domaines émergents seulement). On reviendra sur le détail des domaines aux paragraphes suivants (2.3.2 et 2.3.4 ; *cf.* tableaux 6.15 et 6.16), où sont présentés les graphiques illustratifs. Un tel accroissement peut être le témoin d'une forte diversification des domaines d'emploi et être l'indice d'une dédomanialisation du sens de l'unité lexicale.

[76] Les valeurs des spécificités les plus élevées en valeur absolue ont été réduites à une valeur seuil, d'une part pour des raisons de lisibilité du graphique (représentation plus compacte des autres valeurs, moins adaptée pour distinguer les contrastes), d'autre part parce que ce changement ne modifie pas l'interprétation : les surreprésentations ou sous-représentations particulièrement marquées par rapport aux autres continuent d'apparaître comme telles.

[77] Interprétation du diagramme : étant donné le mode de calcul des spécificités, il ne faut pas observer un point isolément et observer l'accroissement de sa spécificité d'une période à l'autre, mais considérer l'évolution de l'ensemble de points de chaque période. Par exemple, une spécificité de -40 en 2004 et de -20 en 2010 correspond à une sous-représentation marquée. L'écart de 20 entre les deux valeurs peut être un simple effet de variation de la taille du sous-corpus, il ne doit donc pas donner lieu à une interprétation abusive d'accroissement des emplois. On constate d'ailleurs qu'il y a un tassement global de l'échelle de spécificité, qui est précisément le type de phénomène qu'entraîne une évolution de la taille du sous-corpus. En revanche, un point doit s'interpréter par rapport à l'ensemble des autres domaines. On peut interpréter le passage d'une spécificité négative à une spécificité positive comme suit : un domaine sous-représenté par rapport aux autres domaines devient surreprésenté, donc il s'impose par rapport à d'autres domaines. Un autre axe d'interprétation consiste à observer l'évolution de rang d'un domaine : par exemple, INFORMATIONS ECONOMIQUES passe de la deuxième spécificité la plus faible en 2004 (donc la 7e spécificité la plus forte) à la quatrième spécificité la plus forte, il prend ainsi le pas sur trois autres domaines, à savoir POLITIQUE / RELATIONS INTERNATIONALES, MODE DE VIE et SOCIETE / COMMUNAUTE / TRAVAIL, tandis que ces trois domaines gardent la même position relative.

Après application de seuils de fréquences et de spécificités, la liste des cibles lexicales peut être amendée : il y aura élimination si la diffusion n'est pas suffisamment significative, ni pour un domaine donné, ni en considérant les domaines dans leur ensemble. Les résultats expérimentaux sont présentés ci-dessous.

		Cible retenue (existence d'un accroissement minimal ; seuil de spécificité : 5 pour un domaine au moins)	Nature de la diffusion	
			Accroissement distinguant un domaine particulier	Accroissement comparable sur plusieurs domaines
Cible	***toxique***	✓	✓	✗
	délétère	✗		
	dangereux	✗		
	tsunami	✓	✗	✓
	raz-de-marée	✗		
	tempête	✓	✗	✓
	numérique	✓	✗	✓
	tablette	✓	✗	✓
	moléculaire	✗		
Légende – ✓ : critère en tête de colonne validé ; ✗ **:** critère en tête de colonne non validé				

Tableau III.1.14 : Cibles conservées à l'issue de l'analyse des domaines

Les cibles pressenties comme néologiques, à savoir *toxique, tsunami* et *tablette* sont identifiées comme néologiques d'après les critères d'évolution domaniale. *Numérique*, qui présente des caractéristiques similaires à celles de la néologie en termes de diffusion (emplois démultipliés et diffusion non plus spécialisée mais généralisée, quel que soit le domaine) fait également partie des cibles retenues. En revanche, malgré les nouveaux emplois en gastronomie, *moléculaire* n'est pas retenu. Son emploi en gastronomie reste rare : il relève d'un champ trop spécialisé, pas suffisamment représenté par rapport aux autres domaines. De plus, la diffusion n'est pas suffisamment importante pour que les seuils de significativité soient franchis, même si une analyse manuelle témoigne d'une évolution du nombre d'emplois. La majorité des cibles témoins non néologiques, à savoir *dangereux*, *délétère* et *raz-de-marée*, sont éliminées par la procédure, conformément aux attentes. En revanche, *tempête* est conservé car ses emplois connaissent une diffusion notable. La diffusion n'implique pas nécessairement un changement de sens. Si on considère les fréquences de *tempête*, le taux de présence apparaît toujours élevé (ce qui est au demeurant le même cas de figure que *numérique*), donc les domaines d'emploi ne sont pas nouveaux. Une étude plus poussée des fréquences avec notamment l'intégration de seuils de fréquence maximaux amènerait à éliminer *numérique* et *tablette*.

2.3.2 Associer un ou des domaines à chaque cible

Les données qui ont permis d'amender l'ensemble des cibles lexicales peuvent être réutilisées pour associer à chaque cible un ou des domaines où se diffuse vraisemblablement le nouveau sens. À l'étape précédente, l'attention était focalisée sur des informations quantitatives, telles que des franchissements de seuils ou le nombre de domaines émergents. Dans l'étape présente, les informations recherchées sont qualitatives. Les sorties du traitement sont des étiquettes de domaines susceptibles de qualifier le nouveau sens.

Les domaines émergents peuvent être repérés de deux façons : à travers un accroissement significatif de la présence de la cible au cours du temps et au sein des documents relevant du

domaine considéré (étude à domaine fixé) ; à travers l'émergence d'un nouveau domaine dans la tête de liste des domaines d'emplois (comparaison des ventilations de domaines). Dans les illustrations (tableau III.1.15, *cf.* infra, issu de l'analyse des résultats expérimentaux), les deux approches ont tendance à converger, mais elles peuvent fournir des éclairages complémentaires, notamment dans le cas de diffusion dans plusieurs domaines. Les exemples de *tsunami* et *tablette* sont particulièrement frappants : *tablette* se diffuse dans certains domaines en particulier, mais dans l'ensemble, la hiérarchie des domaines d'emplois reste la même et les contrastes demeurent ; *tsunami* se diffuse plus dans certains domaines, mais en parallèle, l'importance de ses emplois entre les différents domaines tend à s'homogénéiser, les contrastes s'estompent.

	Domaines présentant un accroissement significatif en 2010 (évalué à domaine fixé)	**Évolution de la ventilation des domaines**
toxique	**INFORMATIONS ECONOMIQUES**	**INFORMATIONS ECONOMIQUES** s'impose
tsunami	**ARTS ET SPECTACLES** **SOCIETE – COMMUNAUTE – TRAVAIL**	Évolution d'une ventilation très contrastée vers une ventilation faiblement contrastée ➔ évolution vers une présence à peu près équilibrée entre tous les domaines ➔ renforcement relatif d'**INFORMATIONS ECONOMIQUES**, **ARTS ET SPECTACLES, SOCIETE - COMMUNAUTE – TRAVAIL**
tempête	**INFORMATIONS ECONOMIQUES** **POLITIQUE – RELATIONS INTERNATIONALES** **SOCIETE – COMMUNAUTE - TRAVAIL**	Évolution d'une ventilation contrastée vers une ventilation qui reste contrastée, mais de façon moindre ➔ renforcement relatif d'**INFORMATIONS ECONOMIQUES** et, dans une moindre mesure, de **POLITIQUE – RELATIONS INTERNATIONALES**
numérique	**POLITIQUE – RELATIONS INTERNATIONALES** **MODE DE VIE** **ARTS ET SPECTACLES** **SOCIETE – COMMUNAUTE – TRAVAIL** **SANTE** **SCIENCES ET TECHNOLOGIES**	Ventilation contrastée en 2004 et 2010, qui reste approximativement la même ➔ Aucun domaine ne prend notablement le pas sur les autres
tablette	**ARTS ET SPECTACLES** **INFORMATIONS ECONOMIQUES** **POLITIQUE – RELATIONS INTERNATIONALES** **SCIENCES ET TECHNOLOGIES**	Ventilation moyennement contrastée en 2004 et 2010, qui reste approximativement la même. **INFORMATIONS ECONOMIQUES** gagne un peu de terrain dans la ventilation

Tableau III.1.15 : Domaines associés à chaque cible selon l'axe d'analyse

L'analyse des accroissements (évolution des spécificités) est à compléter par une analyse de la présence de la cible, rare ou non (évolution des fréquences)[78]. Par exemple, *tablette* et *tempête* présentent tous deux des accroissements propres à certains domaines et des évolutions de ventilations similaires. Cependant, comme le montre l'évolution à domaine fixé (colonne de gauche du tableau III.1.15), *tablette* passe de rare à non rare entre 2004 et 2010 dans un certain nombre de domaines (INFORMATIONS ECONOMIQUES, POLITIQUE – RELATIONS

[78] La rareté de la présence est déterminée par un seuil de fréquence. Celui-ci est calculé en fonction de la taille du sous-corpus considéré (sous-corpus défini par un couple domaine-période). La fonction de calcul est $\sqrt{t}/6$, où t est la taille du sous-corpus. La fonction racine nous a paru un intermédiaire satisfaisant entre une constante, qui ne prend pas en compte le déséquilibre des sous-corpus, et une fonction linéaire, dont les écarts de seuils paraissaient excessifs par rapport aux écarts de taille de sous-corpus. La constante multiplicative 1/6 a été choisie empiriquement, pour que les seuils restent de l'ordre de la dizaine. Le calcul comporte une part d'arbitraire, comme tout choix de seuil.

INTERNATIONALES, SCIENCES ET TECHNOLOGIES), tandis que *tempête* n'était rare dans aucun domaine ni en 2004, ni en 2010. On peut supposer que, pour *tempête*, il ne sera pas question de nouveau sens, mais simplement d'un usage plus répandu. Pour *tablette*, l'existence de nouveaux domaines d'emplois renforce l'hypothèse d'un nouveau sens, que la confrontation au sens codé permettra de valider ou d'infirmer.

2.3.3 Tester la régularité de la diffusion domaniale

Aux étapes précédentes, on a comparé deux périodes de temps seulement. Il est ainsi possible d'identifier un contraste, mais cela ne renseigne pas sur la progression de la diffusion. Celle-ci peut avoir des profils variés : diffusion linéairement croissante, pic de diffusion puis recul progressif (phénomène de mode ou événement marquant par exemple), croissance irrégulière, etc. Avec deux périodes seulement, on accroît le risque de sélectionner à tort certains candidats (par exemple, si la deuxième période correspond à un pic ponctuel) ou d'écarter des candidats valables (par exemple, si la deuxième période correspond à un creux dans une évolution irrégulière).

Évaluer la régularité de la diffusion n'a pas tant pour vocation d'effectuer un nouveau tri des candidats que de nuancer l'analyse. Par exemple, pour les néologismes de mode, un certain nombre de questions sont délicates à trancher : faut-il les exclure ? Les garder, et si oui, moyennant quel accroissement avant et après un pic de saillance ? Si on se maintient à des seuils supérieurs à ceux avant le pic, mais qu'il y a tout de même décroissance d'année en année, peut-on considérer que le nouveau sens tend à s'implanter dans la langue (donc qu'une allocation de signifié mérite d'être réalisée) ? Dans notre cadre, le contraste de deux périodes sert de critère initial, même s'il comporte une part d'arbitraire. L'allocation de signifié sera réalisée selon ce critère, mais la régularité de la diffusion pourra servir à distinguer les candidats selon leur profil d'évolution, à établir une priorité dans les signifiés à allouer (on privilégiera les néologies dont la diffusion s'inscrit dans la durée) et à nuancer la caractérisation domaniale (on peut considérer que le nouveau sens entretient une relation plus forte avec un domaine qui se maintient dans le temps comme domaine d'emploi qu'avec un domaine présent ponctuellement). Cela peut également servir de signal d'alerte (indicateur sur la fiabilité du choix d'un candidat par exemple).

Afin d'illustrer ce point, considérons l'évolution des domaines de 2004 à 2010 pour les cibles *toxique, tsunami, tablette, numérique* et *tempête*. Pour chaque cible, deux graphiques d'évolution sont présentés :

- les courbes d'évolution des spécificités à domaine fixé ;
- les évolutions des ventilations, obtenues par calcul de spécificité sur l'ensemble des domaines à période fixée.

Les principales tendances observables sont ensuite récapitulées. Il apparaîtra que les tendances varient d'une cible lexicale à l'autre : selon la cible, on observe un pic ou une croissance régulière, avec un comportement atypique d'un domaine ou à l'inverse une évolution similaire des différents domaines.

Avant de présenter les résultats, il nous semble nécessaire d'apporter quelques précisions sur l'interprétation des graphiques, dont la lecture demande certaines précautions. L'encadré qui suit précise comment les interpréter et quels écueils il faut éviter.

Les spécificités sont calculées selon un paramètre (domaine ou période), l'autre paramètre étant fixé, puis les résultats sont mis en miroir en fonction de l'autre paramètre. Pour les courbes à domaine fixé et pour les ventilations de domaines, l'ordre dans lequel interviennent

les paramètres est inversé. Cette inversion affecte les valeurs des spécificités et l'interprétation des résultats.

Courbes à domaine fixé. La comparaison de valeurs ponctuelles (spécificités) se fait à domaine fixé, c'est-à-dire pour des points d'une même courbe. Pour comparer les domaines entre eux, il faut confronter les courbes dans leur ensemble, et non point par point. Comparer les valeurs ponctuelles dans différents domaines à une même période amènerait à des interprétations abusives. Il serait aberrant de dire que tel domaine est surreprésenté par rapport à tel autre en 2006. De même, il convient d'être méfiant par rapport à l'amplitude des courbes : deux courbes homothétiques délivreront la même information. L'important est la tendance évolutive : on pourra comparer la régularité de l'accroissement, ou encore l'existence de pics au même moment.

Ventilations de domaines. La comparaison des valeurs ponctuelles se fait à période fixée, c'est-à-dire pour un ensemble de points alignés verticalement. Pour comparer les périodes entre elles, il faut confronter les "brochettes" de points dans leur ensemble. Il faut par exemple éviter de suivre un point d'une période à l'autre, à l'inverse de ce qu'on effectue pour les courbes à domaine fixé. Par exemple, la spécificité d'un domaine A peut passer de 5 en 2004 à 10 en 2005, mais si l'ensemble des spécificités a tendance à se disperser comme un faisceau de 2004 à 2005, et si le domaine A se fait supplanter par d'autres domaines (sa spécificité devient inférieure à d'autres de 2004 à 2005), alors on ne pourra conclure à un accroissement, mais plutôt à un recul par rapport à d'autres domaines. Il faut se méfier de l'amplitude de la ventilation : si les domaines sont ventilés de la même façon, mais que seule l'amplitude de la ventilation varie, il ne faudra pas conclure à une diminution des domaines les plus spécifiques positivement et à un accroissement des domaines ayant les plus fortes spécificités négatives. L'important est la façon dont évolue la configuration des domaines : on pourra observer si un domaine passe d'une sous-représentation à une surreprésentation par rapport aux autres domaines (passage d'une spécificité négative à une spécificité positive), ou s'il gagne au cours du temps des rangs dans le classement des domaines par spécificités décroissantes.

Les résultats sont les suivants :

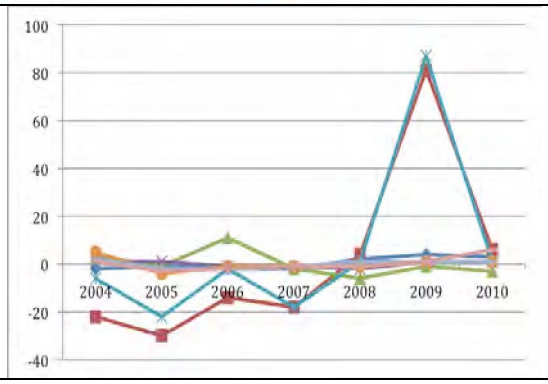
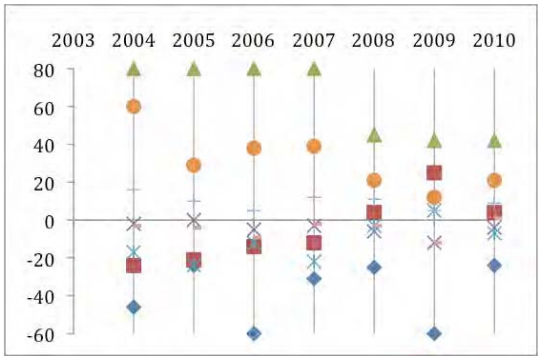
	Évolution des spécificités calculées à domaine fixé de 2004 à 2010	Évolution des ventilations de 2004 à 2010
toxique	Arts et spectacles ; Informations économiques ; Environnement ; Mode de vie ; Politique / relations internationales ; Santé ; Sciences et technologies ; Société / communauté / travail	
		
	En **INFORMATIONS ECONOMIQUES** et **POLITIQUE**, contraste entre les périodes de 2004 à 2007 (sous-emploi) et les périodes 2008 à 2010 (suremploi). Pic en 2009. Accroissement d' **INFORMATIONS ECONOMIQUES** entre 2004 et 2009.	**INFORMATIONS ECONOMIQUES** prend de l'importance dans la ventilation : ce domaine passe de saillances négatives de 2004 à 2007 à des saillances positives de 2008 à 2010. Il prend progressivement le pas sur plusieurs autres domaines au cours du temps.
	L'émergence d'un domaine particulier, **INFORMATIONS ECONOMIQUES**, est nette. Ce domaine connaît un accroissement progressif. L'évolution progressive va dans le sens d'une domanialisation de *toxique* dans le domaine économique.	

Tableau III.1.16.a : Évolution des domaines de 2004 à 2010 dans le voisinage de toxique

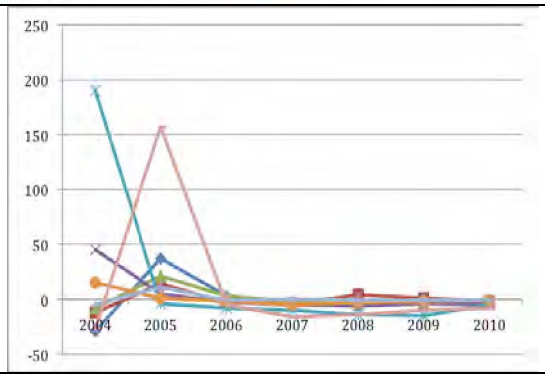
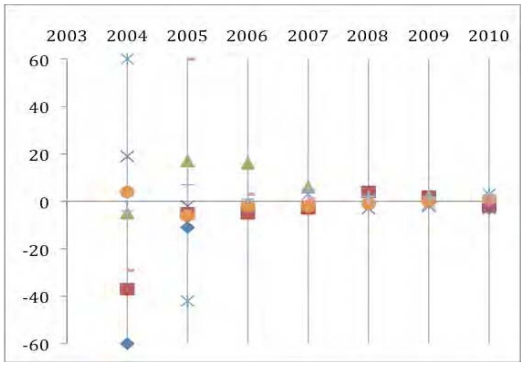
	Évolution des spécificités calculées à domaine fixé de 2004 à 2010	Évolution des ventilations de 2004 à 2010
tsunami	Arts et spectacles ; Informations économiques ; Environnement ; Mode de vie ; Politique / relations internationales ; Santé ; Sciences et technologies ; Société / communauté / travail	
		
	Pic très important en 2005 (dans certains cas, en 2004) dans presque tous les domaines.	Évolution vers un tassement de plus en plus marqué. En particulier, **ENVIRONNEMENT**, le domaine le plus proche de l'ancien sens, domine presque tous les autres domaines en 2005 et 2006, puis il s'éclipse par rapport à d'autres domaines.
	L'évolution période par période confirme l'émergence d'un nouveau sens. Celle-ci semble déclenchée par un pic événementiel. L'évolution de la ventilation des domaines peut s'interpréter comme un emploi métaphorique qui se généralise, car les déséquilibres entre domaines sont de moins en moins nets : tous les domaines deviennent domaines d'emploi.	

Tableau III.1.16.b : Évolution des domaines de 2004 à 2010 dans le voisinage de tsunami

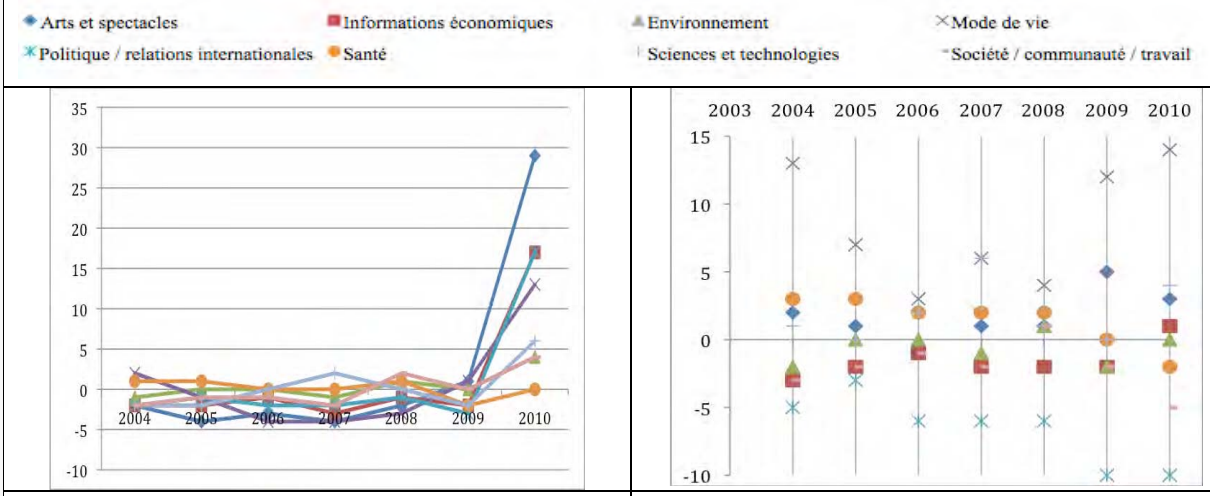
	Évolution des spécificités calculées à domaine fixé de 2004 à 2010	Évolution des ventilations de 2004 à 2010
tablette		
	Pic en 2010 quel que soit le domaine, excepté **SANTE**	Ventilation similaire d'année en année, mais à partir de 2009, une reconfiguration s'amorce, puis se renforce en 2010. **SANTE** s'efface, **INFORMATIONS ECONOMIQUES** s'impose (gain de trois rangs, passage d'une sous-représentation (spécificité négative) à une surreprésentation en 2010), **SCIENCES ET TECHNOLOGIES** se renforce (gain de deux rangs).
	L'évolution des domaines témoigne d'une amorce de changement, qui se manifeste de façon assez brutale. Le changement pourrait amener une reconfiguration des domaines d'emploi, avec l'apparition de plusieurs nouveaux d'emploi, **INFORMATIONS ECONOMIQUES** et **SCIENCES ET TECHNOLOGIES**. Comme le changement se produit en 2010, le recul n'est pas suffisant pour savoir si l'évolution de sens perdure ou non.	

Tableau III.1.16.c : Évolution des domaines de 2004 à 2010 dans le voisinage de tablette

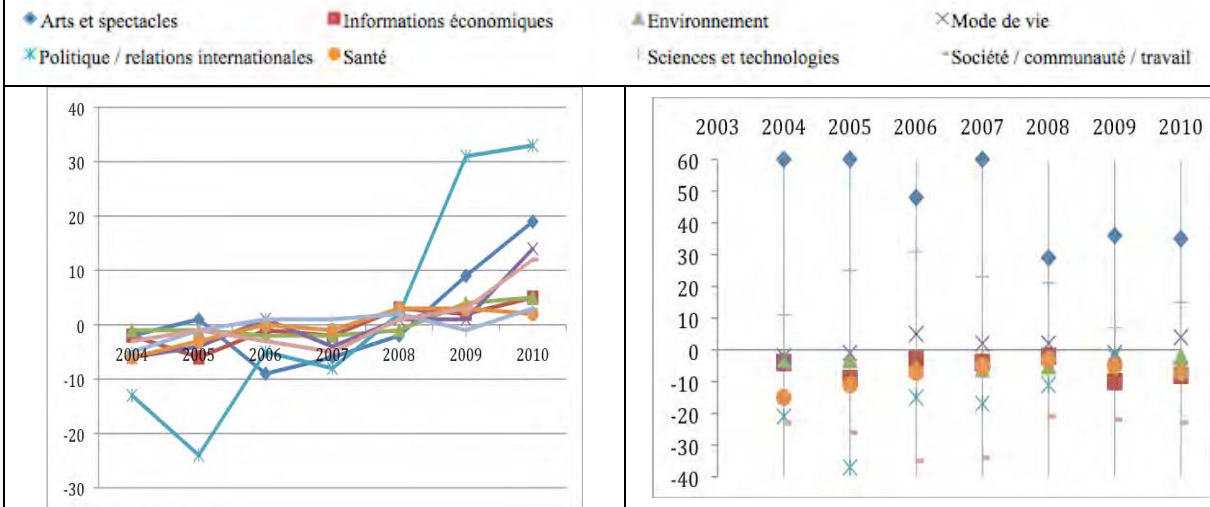
	Évolution des spécificités calculées à domaine fixé de 2004 à 2010	Évolution des ventilations de 2004 à 2010
numérique		
	Croissance progressive de 2006 à 2010 dans la plupart des domaines (sauf pour **SCIENCES ET TECHNOLOGIES ;** présence également peu marquée en **SANTE** et **ENVIRONNEMENT**).	**POLITIQUE**, fortement minoritaire par rapport aux autres domaines de 2004 à 2007, gagne du terrain à partir de 2008. Par ailleurs, la configuration évolue peu.
	La diffusion dans les domaines est générale et progressive. Elle est plus marquée en **POLITIQUE** que dans les autres domaines, ce qui pourrait éventuellement amener soit à intégrer **POLITIQUE** au nouveau sens, soit à revoir la hiérarchie des domaines d'emploi pour donner plus de poids à **POLITIQUE**.	

Tableau III.1.16.d : Évolution des domaines de 2004 à 2010 dans le voisinage de numérique

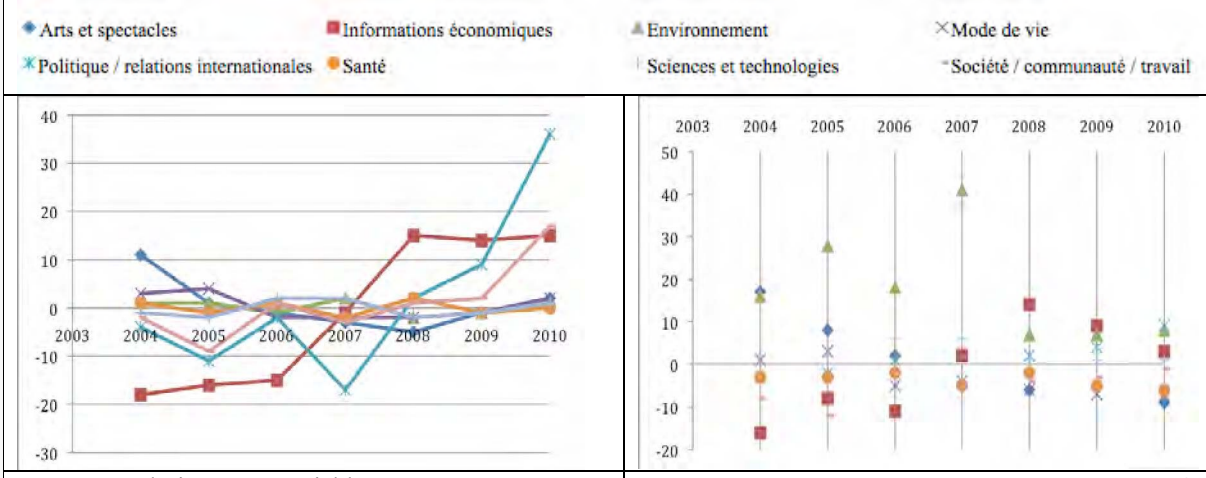
	Évolution des spécificités calculées à domaine fixé de 2004 à 2010	Évolution des ventilations de 2004 à 2010
tempête		
	Les évolutions sont variables. **INFORMATIONS ECONOMIQUES, POLITIQUE / RELATIONS INTERNATIONALES** et **SOCIETE / COMMUNAUTE / TRAVAIL** présentent globalement un accroissement.	**INFORMATIONS ECONOMIQUES et POLITIQUE / RELATIONS INTERNATIONALES** s'imposent dans la ventilation, tandis qu'**ARTS ET SPECTACLES** s'efface.
	Il n'y a pas de loi générale qui se dégage nettement de l'évolution des domaines. Une explication éventuelle serait que le sens métaphorique est déjà implanté et s'adapte au fil des événements.	

Tableau III.1.16.e : Évolution des domaines de 2004 à 2010 dans le voisinage de tempête

2.3.4 Organiser les domaines de la cible

Si plusieurs domaines ont été associés à une cible lexicale, il convient de les organiser de deux façons : en les hiérarchisant et en les regroupant.

Une hiérarchie des domaines peut être établie à partir de critères variés :

- lorsqu'on dispose de deux périodes seulement, on peut utiliser un tri par significativité décroissante des significativités à domaine fixé de la nouvelle période par rapport à l'ancienne, ou un tri par valeur décroissante de la différence de significativité entre la nouvelle période et l'ancienne période dans la ventilation de domaines ;
- lorsqu'on dispose de plusieurs périodes, on peut s'appuyer sur les coefficients des pentes de régressions linéaires[79], calculées à domaines fixés ou relativement aux ventilations de domaines.

D'après notre perspective, le critère de diffusion d'un emploi au sein d'un domaine (émergence d'un nouveau domaine) nous semble prioritaire sur un accroissement dans la ventilation (reconfiguration des domaines d'emploi, évolution de leur poids relatif). Les valeurs obtenues et, par conséquent, l'organisation des domaines en fonction de ces valeurs peuvent varier selon le découpage en période choisi (restriction à deux périodes ou multiplicité de périodes) et selon le type d'observation retenu (à domaine fixé ou dans la ventilation des domaines).

Le tableau ci-dessous donne l'exemple de *toxique* et *tsunami*, pour lesquels les domaines ont été classés selon les axes d'analyse précédemment cités : pour les deux années 2004 et 2010 ou pour l'ensemble des années entre 2004 et 2010 ; à domaine fixé ou selon la ventilation des domaines. Chaque couple d'axes d'analyse (couple (2004 et 2010, à domaine fixé), couple

[79] Les pentes de régression linéaire sont les droites qui passent au plus près des points représentatifs des spécificités par période. Les droites de régression linéaire sont obtenues par la méthode des moindres carrés. La distance utilisée dans la méthode des moindres carrés est la distance euclidienne.

(2004 à 2010, ventilation des domaines), etc.) permet d'affecter une valeur à la cible lexicale dans chaque domaine (la valeur est la spécificité lorsqu'il n'y a que deux périodes et la pente de régression linéaire lorsqu'il y a toutes les périodes). Autrement dit, pour un couple d'axes d'analyse et à cible fixée, on dispose d'une valeur associée à chaque domaine. Les domaines sont triés par valeur décroissante et un rang leur est affecté.

Cible		*toxique*				*tsunami*			
Année		2004 et 2010		2004 à 2010		2004 et 2010		2004 à 2010	
Accroissement des spécificités dans le temps		à domaine fixé	dans la ventilation	à domaine fixé	dans la ventilation	à domaine fixé	dans la ventilation	à domaine fixé	dans la ventilation
Rang	**1**	ECO	ECO	ECO	ECO	ART	ART	SOC	SOC
	2	POL	ART	POL	SOC	SOC	ECO	POL	POL
	3	ART	POL	SOC	SCI	ECO	SOC	ART	ECO
	4	SOC	SOC	SCI	MOD	ENV	ENV	ENV	ART
	5	MOD	MOD	ART	ART	SCI	SCI	ECO	ENV
	6	SCI	SCI	SAN	SAN	SAN	SAN	SAN	SAN
	7	SAN	SAN	MOD	ENV	MOD	MOD	SCI	MOD
	8	ENV	ENV	ENV	POL	POL	POL	MOD	SCI

Légende. ART = Arts et spectacles ; ECO = Informations économiques ; ENV = Environnement ; MOD = Mode de vie ; POL = Politique / relations internationales ; SAN = Santé ; SCI = Sciences et technologies ; SOC = Société / communauté / travail

Tableau III.1.17 : Hiérarchie des domaines en fonction de l'accroissement des spécificités dans le temps

Les classements se rejoignent dans l'ensemble. Par exemple, INFORMATIONS ECONOMIQUES est en tête pour *toxique* ; SOCIETE, INFORMATIONS ECONOMIQUES et ARTS ET SPECTACLES sont dominants pour *tsunami*. Cependant, quelques écarts nets peuvent apparaître dans les classements. Par exemple, pour *toxique*, sur l'ensemble des périodes (2004 à 2010), POLITIQUE apparaît en tête de liste à domaine fixé mais en dernière position dans la ventilation. Ceci signifie que, même si *toxique* est nettement plus employé en POLITIQUE qu'il ne l'était auparavant dans ce même domaine, le domaine d'emploi POLITIQUE reste mineur par rapport aux autres domaines d'emploi. De même, si l'on considère *tsunami* à domaine fixé, POLITIQUE est le domaine qui connaît le plus faible accroissement en ne considérant que les périodes 2004 et 2010, alors que, sur l'ensemble des années 2004 à 2010, il a le deuxième plus fort accroissement. Ce phénomène peut s'expliquer par un accroissement des emplois pendant un certain nombre d'années, puis une chute des emplois en 2010 qui se démarque de l'évolution des années précédentes. Le fait de se restreindre à deux années dissociées ne permet pas de savoir ce qui se passe dans l'intervalle, et un pic ou une croissance sur quelques années peuvent ne pas apparaître.

Le regroupement de domaines est possible en confrontant les profils d'évolution. Pour cela, un découpage en deux périodes n'est pas suffisant, il faut disposer d'un découpage en plusieurs périodes. Une corrélation entre profils d'évolution permettra d'établir des groupes d'affinités, ou au contraire des oppositions. Ainsi, les domaines présentant un pic de saillance aux mêmes périodes seront associés (cas des néologies à la mode), de même que ceux qui se caractériseront par une croissance progressive régulière ou par une croissance progressive exponentielle. La constitution de tels regroupements repose sur l'hypothèse que des profils d'évolution similaires correspondent à des liens sémantiques plus marqués.

Dans les expériences sur le corpus 'Factiva', les corrélations ont été calculées pour tout couple de domaines à partir des matrices de spécificités utilisées précédemment pour observer les évolutions de domaines, à savoir :

- les spécificités calculées à domaines fixés. Les coefficients de corrélation permettent de voir si les courbes présentent les mêmes tendances d'évolution ;
- les spécificités calculées pour les ventilations de domaines. Les coefficients de corrélation permettent de voir comment évoluent les positions relatives des deux domaines dans la ventilation.

D'un point de vue technique, les corrélations sont calculées comme suit :

- Pour les spécificités à domaine fixé : pour un domaine donné, une valeur de spécificité est affectée à chaque période. On dispose donc d'un vecteur de spécificités pour le domaine considéré, avec un coefficient associé à chaque période. La corrélation de deux domaines est établie entre les vecteurs de spécificités de chacun des domaines.
- Pour les ventilations : pour une période donnée, une valeur de spécificité est affectée à chaque domaine. On dispose d'un vecteur de spécificités pour la période considérée, avec un coefficient par domaine. Les vecteurs obtenus pour les différentes périodes peuvent être réunis pour générer une matrice domaine-période, avec une valeur affectée à chaque couple (domaine ; période). La corrélation de deux domaines est obtenue à partir de la suite de spécificités établie à chaque période pour chacun des domaines.

Dans les deux cas, on obtient un coefficient de corrélation pour tout couple de domaines, ce qui permet de générer une matrice carrée domaine-domaine, dont les entrées ont pour valeur les coefficients de corrélation.

Pour structurer les domaines en fonction de leurs profils d'évolution, on applique deux techniques sur les matrices de corrélation : une classification hiérarchique, qui permet de dégager des regroupements successifs de domaines, et une méthode de sériation, qui permet d'ordonner les domaines de façon à ce qu'il y ait le moins de variations possible entre deux domaines successifs. D'un point de vue technique, la sériation consiste à permuter les lignes et les colonnes de la matrice, de façon à ce que des coefficients ayant des valeurs proches se trouvent à proximité l'un de l'autre dans la matrice. Ainsi, on peut faire émerger des blocs de valeurs proches. Les méthodes de sériation et de classification hiérarchique utilisées pour faire émerger des regroupements et des blocs ont été appliquées à l'aide du logiciel PermutMatrix (Caraux et Pinloche, 2005 ; *cf.* chapitre II.2, 4.2, pour les méthodes de structuration de données). Ce logiciel permet de visualiser les résultats à partir de matrices dont les coefficients ont été transformés en dégradés de couleurs : des dégradés de rouge pour les coefficients de corrélation positifs (rouge vif pour les valeurs maximales), des dégradés de vert pour les coefficients négatifs (vert vif lorsqu'on tend vers -1) ; le noir correspond à des valeurs proches de 0.

Les tableaux ci-dessous présentent les résultats pour les différentes cibles lexicales. Pour ***visualiser*** les regroupements, on présente le résultat de PermutMatrix, c'est-à-dire la matrice colorée obtenue après sériation et CAH. Ensuite sont listés des groupes d'***affinités***, tels que tout couple de domaines au sein du groupe a un coefficient de corrélation supérieur à 0,8. Ces groupes émergent assez explicitement sur la matrice des corrélations, au niveau des CAH (feuilles regroupées par des nœuds situés assez bas dans l'arbre ; nos arbres sont orientés horizontalement, le bas des arbres est à droite). Il s'agit donc d'affinités particulièrement marquées. Les ***oppositions*** se lisent simplement sur la matrice comme les regroupements dont les nœuds sont situés le plus haut dans l'arbre. La visualisation, les affinités et les oppositions

sont complétées par des commentaires qui s'appuient sur ces résultats, ainsi que sur l'analyse des coefficients de corrélation.

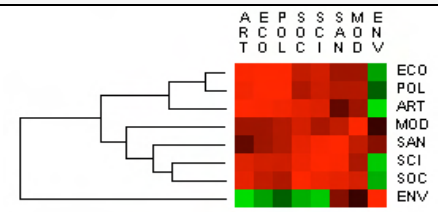
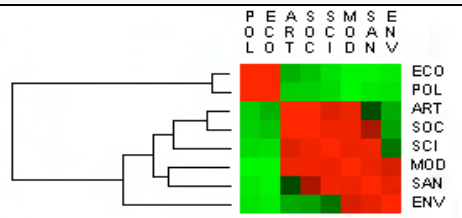
toxique		
	À domaine fixé	Ventilations
Visualisation		
Affinités particulièrement marquées	• {ECO, ART} • {ECO, POL} • {SCI, SAN}	• {ECO, POL} • {SOC, ART}
Oppositions	• ENV vs { ECO, POL, ART, MOD, SOC, SCI, SAN}	• {ECO, POL} vs {MOD, SAN, SCI, SOC, ART, ENV}
À domaine fixé, les coefficients de corrélation sont très variables : certains domaines ont des évolutions étroitement associées à d'autres (coefficients proches de 1), d'autres ont des évolutions sans liens (coefficients proches de 0), d'autres encore évoluent plutôt de façon contraire (coefficient négatif, qui se rapproche de -1). INFORMATIONS ECONOMIQUES (ECO) est associé fortement à ARTS ET SPECTACLES (ART) et POLITIQUE / RELATIONS INTERNATIONALES (POL), comme le montre le regroupement des trois premières lignes de la CAH : ces trois domaines tendent à évoluer de la même façon. Par contre, l'évolution d' INFORMATIONS ECONOMIQUES est sans lien avec celle d'ENVIRONNEMENT (ENV) et de SANTE (SAN), domaines associés aux anciens sens : le nœud qui relie ces domaines intervient tardivement dans la CAH, et les coefficients de corrélation sont proches de 0, ce qui témoigne d'une absence de corrélation. Les corrélations issues des ventilations de domaines affichent des tendances similaires, mais ARTS ET SPECTACLES se dissocie du couple {INFORMATIONS ECONOMIQUES, POLITIQUE / RELATIONS INTERNATIONALES} pour s'associer à SOCIETE / COMMUNAUTE / TRAVAIL (SOC).		

***Tableau III.1.18.a : Regroupements des domaines en fonction de leurs corrélations pour la cible* toxique**

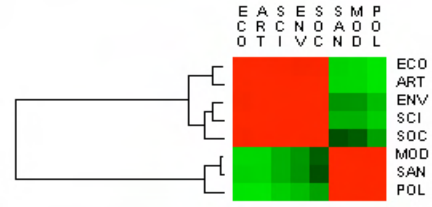
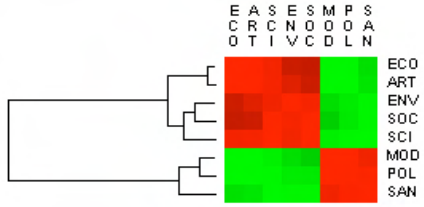
tsunami		
	À domaine fixé	Ventilations
Visualisation		
Affinités particulièrement marquées	• {ECO, ART, ENV, SCI, SOC} • {MOD, POL, SAN}	• {ECO, ART} • {SCI, SOC, ENV} • {MOD, POL, SAN}
Oppositions	{ECO, ART, ENV, SCI, SOC} vs {MOD, POL, SAN}	{ECO, ART, ENV, SCI, SOC} vs {POL, MOD, SAN}
Existence de sous-ensembles de domaines très fortement corrélés positivement 2 à 2. Pour des domaines de sous-ensembles différents, les corrélations sont faibles ou négatives, d'où des évolutions à tendance inverse. Ce phénomène peut correspondre à l'existence d'événements liés successifs, à savoir le pic événementiel, auquel succède la diffusion d'emplois métaphoriques.		

Tableau III.1.18.b : Regroupements des domaines en fonction de leurs corrélations pour la cible tsunami

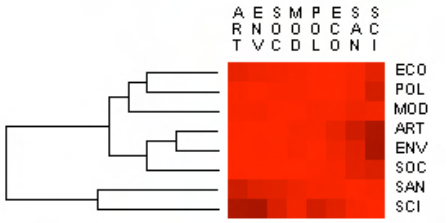
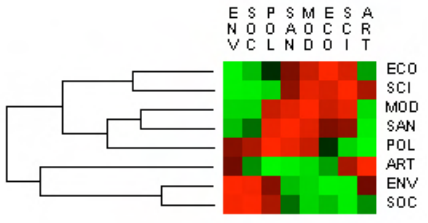
numérique		
	À domaine fixé	Ventilations
Visualisation		
Affinités particulièrement marquées	• {ECO, POL} • {SOC, ART, ENV} • {SOC, POL, ENV} • {SOC, MOD}	/
Oppositions	{SAN, SCI} vs {ECO, POL, MOD, ART, ENV, SOC}	{ECO, SCI, MOD, SAN, POL} vs {ART, ENV, SOC}
À domaine fixé, les corrélations sont toutes positives. Tous les domaines ont donc une même tendance évolutive. Ces évolutions ne sont pas totalement identiques : les coefficients de corrélation s'échelonnent entre 0 et 1, donc, pour certains couples de domaines, les évolutions ne sont que moyennement corrélées. La ventilation ne fait pas émerger d'évolutions fortement couplées. Les oppositions qui sont présentées ne sont pas particulièrement nettes. Les domaines ont donc tendance à évoluer tous de la même façon, sans que des comportements propres à des sous-ensembles de domaines se dégagent de façon claire.		

Tableau III.1.18.c : Regroupements des domaines en fonction de leurs corrélations pour la cible numérique

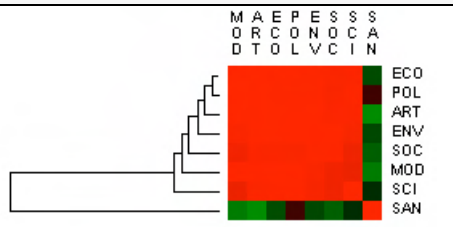
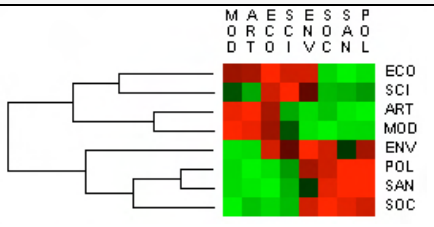
tablette		
	À domaine fixé	Ventilation
Visualisation		
Affinités particulièrement marquées	• {ECO, POL, ART, ENV, SOC} • {ECO, POL, ART, ENV, MOD} • {ECO, POL, ART, SCI}	• {SAN, POL}
Oppositions	SAN vs {ECO, POL, ART, ENV, SOC, MOD, SCI}	{ECO, SCI, ART, MOD} vs {ENV, POL, SAN, SOC}
À domaine fixé, l'évolution de SANTE (SAN) est décorrélée de celle des autres domaines. À cette exception près, toutes les évolutions sont assez, voire très fortement corrélées (coefficient de corrélation compris entre 0,6 et 1). La tendance d'évolution est la même pour tous les autres domaines que SANTE. Les évolutions dans la ventilation font émerger une configuration différente, avec des alliances et des oppositions qui n'apparaissaient pas à domaine fixé. Les différences de regroupements et oppositions entre domaines fixés et ventilations sont l'indice d'une diffusion générale, marquée par une reconfiguration de l'importance relative des domaines d'emploi.		

Tableau III.1.18.d : Regroupements des domaines en fonction de leurs corrélations pour la cible tablette

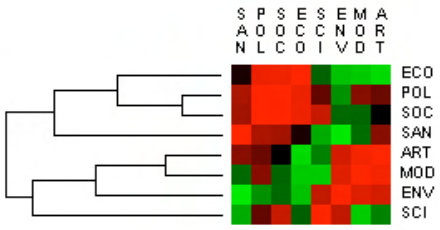
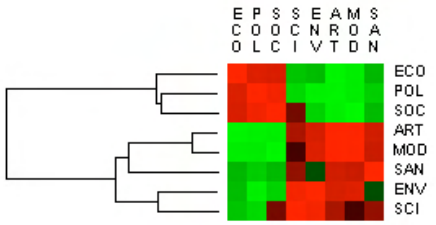
tempête		
	À domaine fixé	Ventilation
Visualisation		
Affinités particulièrement marquées	• {POL, SOC}	• {ART, MOD}
Oppositions	{ECO, POL, SOC, SAN} vs {ART, MOD, ENV, SCI}	{ECO, POL, SOC} vs {ART, MOD, SAN, ENV, SCI}
À domaine fixé, les évolutions ne font pas émerger de grands groupes de domaines ayant des évolutions similaires (c'est-à-dire fortement corrélées). Les coefficients de corrélation ont des valeurs très variables. Une seule corrélation forte apparaît pour le couple {POLITIQUE / RELATIONS INTERNATIONALES (POL), SOCIETE / COMMUNAUTE / TRAVAIL (SOC)}. Dans les ventilations, le positionnement par rapport à l'ensemble des domaines évolue plutôt de la même façon pour le triplet {ÉCONOMIE, POLITIQUE / RELATIONS INTERNATIONALES et SOCIETE / COMMUNAUTE / TRAVAIL}, ainsi que pour les cinq autres domaines. Les évolutions des différents domaines semblent donc peu liées entre elles.		

Tableau III.1.18.e : Regroupements des domaines en fonction de leurs corrélations pour la cible tempête

Les résultats des deux types de corrélation se rejoignent et permettent de préciser les regroupements ou oppositions.

2.3.5 Bilan

Les étapes précédemment décrites correspondent à différentes façons d'étudier les empreintes de fréquence thématisées et d'en extraire de l'information. À l'issue du processus, on dispose :

- ***d'une liste révisée de cibles lexicales***. Les cibles présélectionnées ont été divisées en deux sous-groupes, selon qu'elles sont associées ou non à des variations domaniales remarquables. Les cibles sans variation domaniale ne répondent pas à notre critère prioritaire, elles seront donc écartées par la suite ;
- ***d'une première qualification du nouveau sens à partir d'étiquettes de domaines***. À chaque cible lexicale est associé un ou plusieurs domaines dans lesquels a lieu une diffusion significative de la cible. Dans le cas où plusieurs domaines sont présents, des coefficients leur sont affectés pour évaluer l'importance de la diffusion. Ces valeurs permettent de hiérarchiser les domaines, et ainsi de déterminer une priorité dans leur ordre d'analyse par la suite. Dans le cas d'un découpage en plusieurs périodes de temps, on peut obtenir des regroupements de domaines selon la similarité des profils d'évolution. Enfin, il est important de disposer d'un mode de visualisation de l'évolution (histogramme, cartographie thématique), pour fournir des outils de contrôle et guider l'interprétant à travers une vue d'ensemble.

À ce stade, la qualification de l'évolution est textuelle : elle se fait à partir d'informations propres au corpus. L'étape suivante consiste à rattacher les informations issues du corpus à celles propres à la ressource lexicographique, pour pouvoir comparer les deux ressources et transposer les variations observées en corpus à la représentation du sens codé.

2.4 Du profilage domanial aux domaines du sens codé

On souhaite rattacher les variations domaniales en corpus au dictionnaire. Sur le plan théorique, cela revient à faire le parallèle entre les thèmes ou isotopies globales, restreints ici aux domaines textuels, et les domaines associés au sens codé (c'est-à-dire des sèmes mésogénériques associés au sens codé, qui sont des sèmes de degré de généralité relativement important). Pour y parvenir, on établit dans un premier temps des correspondances générales entre les domaines textuels, qui sont des domaines donnés dans l'expérience Factiva, et les domaines lexicographiques. Ensuite, à partir de ces correspondances, on peut trier les domaines lexicographiques et textuels associés à la cible lexicale : les informations se répartiront selon leur contribution à un enrichissement ou à une reconfiguration du sémème.

2.4.1 Correspondances entre domaines lexicographiques et textuels

Le lien entre domaines lexicographiques et domaines textuels peut être établi manuellement. L'intérêt d'un appariement manuel a posteriori est double. D'une part, il n'est pas nécessaire d'apparier tous les domaines du corpus, mais seulement les domaines des documents où apparaît la cible (par exemple, pour *toxique*, on n'observe pas d'évolution pour MODE DE VIE ou encore ARTS ET SPECTACLES, les occurrences propres à ces domaines ne seront pas analysées plus avant ; les observations se focaliseront sur INFORMATIONS ECONOMIQUES, seul domaine véritablement saillant) ; le nombre de relations à établir est donc réduit. D'autre part, un tel appariement reste soumis au regard humain, il est donc contrôlé.

L'articulation manuelle des domaines n'est pas forcément immédiate. Il peut y avoir notamment des distorsions entre les nomenclatures, qui peuvent être dues à :

- des lacunes. Par exemple, le sujet PRIVACITE / PROTECTION DES DONNEES présent dans Factiva n'a pas de correspondance immédiate dans le TLFi ;
- des chevauchements. Factiva comporte un sujet SPORTS ET LOISIRS, alors que dans le TLFi, SPORTS et LOISIRS sont distincts ;
- des décalages de niveaux. RELIGION est un domaine de premier niveau dans le TLFi et DEMOGRAPHIE un domaine de niveau 3 (sous-domaine de SOCIOLOGIE, lui-même sous-domaine d'ETHNOLOGIE), tandis que RELIGION et DEMOGRAPHIE sont tous deux au même niveau dans Factiva ;
- des incertitudes sur le champ recouvert par une étiquette de domaine. Dans Factiva, ÉDUCATION DES ENFANTS est un affilié à MODE DE VIE, de même niveau qu'ÉDUCATION : doit-on y voir une structure qui ne partitionne pas les articles, mais qui comporte des sous-ensembles susceptibles de se recouvrir ?

Une alternative est d'effectuer un rapprochement à partir de procédures informatisées. Différentes solutions sont possibles :

- ***Projection des domaines lexicographiques dans le corpus***. Cette opération donne en sortie une table de correspondance entre domaines lexicographiques et textuels. La table de correspondance est utilisée après une analyse du comportement de la cible relative à la structure proprement textuelle. L'ordre des traitements n'est pas neutre. Les transformations pour faire émerger l'évolution en corpus peuvent se voir comme l'application de fonctions mathématiques ; associer des domaines du corpus aux domaines lexicographiques à partir des projections est une autre fonction. L'application successive de fonctions correspond à de la composition, qui est une opération non commutative. Autrement dit, l'ordre dans lequel on travaille ne donnera pas nécessairement les mêmes résultats ;

- ***Projections des domaines du corpus dans la ressource lexicographique.*** Les correspondances ne sont pas établies selon la distribution des domaines lexicographiques dans les textes, c'est-à-dire relativement à la structure des domaines textuels, mais en positionnant les domaines textuels par rapport à la structure de la ressource lexicographique. Par exemple, supposons que les domaines sont de type domaines construits et représentés par des classes de mots-clés, dans l'esprit de (Sébillot et Rossignol, 2002), ou qu'ils ont une étiquette qui prend la forme d'une unité lexicale. Les mots-clés ou étiquettes vont créer le lien entre domaines construits et domaines lexicographiques. On peut établir quels domaines lexicographiques interviennent dans les définitions de ces mots-clés ou de quels domaines dépendent les mots-clés lorsqu'ils sont présents dans une définition. À titre d'illustration, considérons l'étiquette de domaine SANTE, un des sujets utilisés pour la caractérisation du corpus 'Factiva'. On fait l'hypothèse que les principaux domaines associés aux définitions contenant l'unité *santé* seront les domaines lexicographiques les plus proches du sujet SANTE. Une requête complexe sur le TLFi permet d'analyser les domaines qui pourraient être associés à SANTE. L'exemple présenté ci-dessous a été analysé à l'aide du moteur de recherche complexe intégré dans le TLFi, qui effectue des recherches sur différentes unités en fonction du type d'objet où elles se situent (définition, exemple et autres parties constitutives d'une entrée) et des liens d'inclusion ou de dépendance qu'elles entretiennent avec d'autres objets (dépendance à un domaine par exemple). La requête effectuée ici, présentée en figure 6.19.a, s'interprète de la façon suivante : le lemme *santé* a été recherché dans des définitions dépendantes d'un domaine technique. Un extrait des résultats est présenté immédiatement après (figures 6.19.b et c).

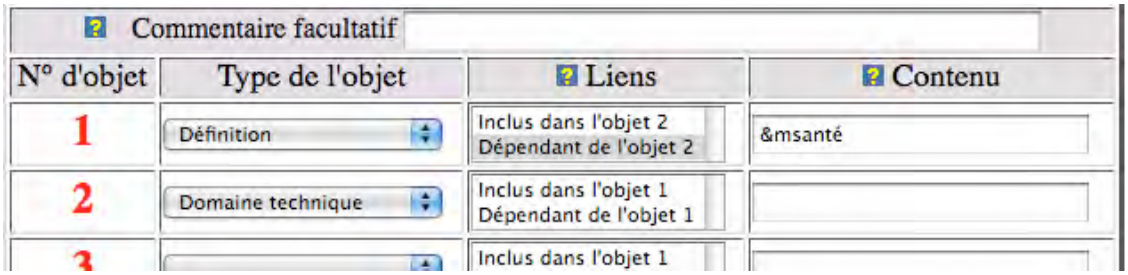
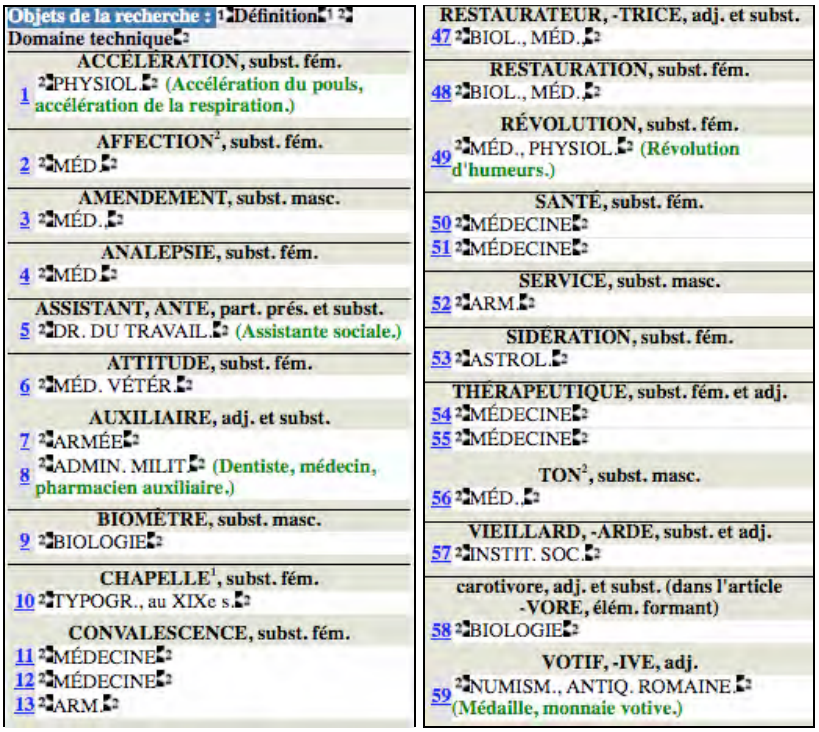


Figure 6.19.a : Requête complexe sur les définitions contenant santé dépendant d'un domaine



Objets de la recherche :	Définition	Domaine technique
ACCÉLÉRATION, subst. fém.		
1		PHYSIOL. (Accélération du pouls, accélération de la respiration.)
AFFECTION[2], subst. fém.		
2		MÉD.
AMENDEMENT, subst. masc.		
3		MÉD.
ANALEPSIE, subst. fém.		
4		MÉD.
ASSISTANT, ANTE, part. prés. et subst.		
5		DR. DU TRAVAIL. (Assistante sociale.)
ATTITUDE, subst. fém.		
6		MÉD. VÉTÉR.
AUXILIAIRE, adj. et subst.		
7		ARMÉE
8		ADMIN. MILIT. (Dentiste, médecin, pharmacien auxiliaire.)
BIOMÈTRE, subst. masc.		
9		BIOLOGIE
CHAPELLE[1], subst. fém.		
10		TYPOGR., au XIXe s.
CONVALESCENCE, subst. fém.		
11		MÉDECINE
12		MÉDECINE
13		ARM.

RESTAURATEUR, -TRICE, adj. et subst.	
47	BIOL., MÉD.
RESTAURATION, subst. fém.	
48	BIOL., MÉD.
RÉVOLUTION, subst. fém.	
49	MÉD., PHYSIOL. (Révolution d'humeurs.)
SANTÉ, subst. fém.	
50	MÉDECINE
51	MÉDECINE
SERVICE, subst. masc.	
52	ARM.
SIDÉRATION, subst. fém.	
53	ASTROL.
THÉRAPEUTIQUE, subst. fém. et adj.	
54	MÉDECINE
55	MÉDECINE
TON[2], subst. masc.	
56	MÉD.
VIEILLARD, -ARDE, subst. et adj.	
57	INSTIT. SOC.
carotivore, adj. et subst. (dans l'article -VORE, élém. formant)	
58	BIOLOGIE
VOTIF, -IVE, adj.	
59	NUMISM., ANTIQ. ROMAINE (Médaille, monnaie votive.)

Figures 6.19.a et b : Début et fin de liste des résultats de la recherche complexe

L'analyse des résultats permet de faire émerger MEDECINE, et dans une moindre mesure BIOLOGIE comme domaines lexicographiques associés à SANTE. 59 définitions dépendant d'un domaine technique contiennent l'unité lexicale *santé*. Parmi les domaines techniques dont dépendent ces définitions, MEDECINE apparaît 28 fois, BIOLOGIE 7 fois et les autres moins de 5 fois. SANTE peut alors être associé à un domaine unique, MEDECINE, qui est le domaine dominant. On peut également établir des associations pondérées, en affectant par exemple un poids de 1 à MEDECINE, de 1/4 à BIOLOGIE, etc. La mise en correspondance de domaines textuels ou, plus largement, de tout autre type d'unité supra-lexicale observé en corpus, et de domaines lexicographiques est une question à part entière, elle constitue un champ d'étude en soi.

2.4.2 Tri des domaines pour préciser l'enrichissement ou la reconfiguration du sémème

Deux types de domaines sont soumis à un tri : les domaines textuels identifiés comme émergents et les domaines lexicographiques associés au sens codé.

Les domaines émergents issus du corpus peuvent se répartir en deux sous-ensembles, selon qu'ils sont associés ou non aux domaines affectés au sens codé :

- S'ils sont associés aux domaines lexicographiques du sens codé de la cible lexicale, ils contribuent à une reconfiguration du sémème : une acception aura tendance à s'imposer. L'analyse de la reconfiguration doit être complétée par l'observation du comportement en corpus des autres domaines de la définition (*cf.* infra).
- S'ils ne sont pas associés aux domaines lexicographiques de la cible, ils témoignent d'un enrichissement du sémème. Si un seul domaine apparaît comme nouveau, on est probablement face à une domanialisation. Si plusieurs domaines sont nouveaux, soit plusieurs sens spécifiques à différents domaines émergent (domanialisations multiples), soit un sens transdomanial se dessine : ce sens n'est pas spécifique à un domaine, mais il est général ou qualifié par plusieurs domaines.

La répartition des domaines émergents selon le critère d'enrichissement ou de reconfiguration est à compléter par une analyse des domaines associés au sens codé. On observera si, en corpus, les domaines textuels auxquels ils correspondent s'effacent, stagnent ou s'imposent. S'ils s'effacent, on pourra y voir l'inhibition d'une acception, donc une perte de domaine. La stagnation ne permet pas de conclusion. L'émergence témoigne d'une acception domaniale qui s'impose, parce qu'elle passe de rare à courante, ou parce que l'acception prendra le dessus sur les autres et deviendra dominante. Une troisième possibilité est qu'un nouveau sens se développe au sein d'un domaine déjà caractéristique de la cible lexicale.

Pour qualifier la reconfiguration ou l'enrichissement domaniaux, il faut donc s'appuyer sur une table de correspondance entre domaines textuels et domaines lexicographiques, déjà évoquée au point précédent (dans les cas où les domaines textuels sont donnés ou construits ; pour les domaines projetés avant tout traitement, la correspondance est immédiate à l'issue du profilage supra-lexical). Ensuite des techniques sont mises en œuvre pour projeter les domaines du sens codé dans cette table et observer leur évolution. Comme précédemment, ces techniques s'appuieront sur des critères de significativité. La significativité est estimée pour les occurrences de la cible relativement au reste du corpus si on ne dispose que d'une période ; pour les occurrences à une période récente relativement aux occurrences initiales ; ensuite, pour préciser le profil d'évolution, elle peut être observée sur plusieurs périodes successives.

À l'issue de cette étape, on dispose donc d'une caractérisation mésosémantique de l'évolution de sens avec d'une part des nouveaux domaines susceptibles de qualifier le nouveau sens, une idée de l'importance de leur émergence (indice de significativité) et leur association à d'autres domaines au comportement similaire ; d'autre part, une représentation de l'évolution des anciens domaines associés, donc de leur reconfiguration.

Le profil obtenu peut alors être précisé à travers des unités lexicales, puis des traits microsémantiques. L'étude d'un niveau de granularité plus fin s'effectue conditionnellement aux domaines.

2.4.3 Illustration : emplois financiers de *toxique*

La cible lexicale de ce cas d'étude est *toxique.* L'étude des domaines associés à *toxique* repose sur la confrontation de deux corpus : le corpus 'Crise financière', témoin des nouveaux emplois, et le corpus 'Monde Diplomatique' pour les anciens emplois (*cf.* chapitre II.1, 1.5.1 pour la présentation des corpus).

Au niveau de la méthodologie, les corpus ont été annotés en domaines lexicographiques (annotation en traits sémantiques par Semy et restriction des annotations aux traits mésogénériques, c'est-à-dire aux domaines), autrement dit, les domaines en corpus sont des domaines projetés. Ensuite, les spécificités des domaines ont été calculées à partir de deux types de comparaisons : des comparaisons des paragraphes du corpus 'Crise financière' contenant *toxique* au reste du corpus 'Crise financière' ; des comparaisons des paragraphes du corpus 'Crise financière' contenant *toxique* aux paragraphes 'Monde Diplomatique' contenant *toxique*. Les résultats ont été analysés d'abord sous l'angle de la reconfiguration, puis sous l'angle de l'enrichissement : on considère d'abord les domaines affectés à *toxique* dans sa définition lexicographique, puis les domaines les plus saillants.

a- Reconfiguration : inhibition des domaines du sens codé

Le sémème affecté à *toxique* à partir du dictionnaire comporte cinq domaines : BIOLOGIE, CHIMIE, MEDECINE, PHARMACOLOGIE et PHYSIOLOGIE. Les valeurs affectées à ces domaines à

l'issue de la projection des domaines lexicographiques sont présentées dans le tableau III.1.20[80].

Corpus		Corpus 'Crise financière' (découpage: voisinages de *toxique* et son complémentaire)		Corpus des voisinages de *toxique* du 'Monde Diplomatique' et de 'Crise financière'	
		Fréquence dans le corpus	Spécificités	Fréquence dans le corpus	Spécificités
Domaines	BIOLOGIE	402	-	220	-5
	CHIMIE	462	-	256	-4
	MEDECINE	934	-	547	-
	PHARMACOLOGIE	45	-	25	-
	PHYSIOLOGIE	393	-3	206	-12

Tableau III.1.20 : Spécificités des domaines issus des définitions de toxique relativement au corpus 'Crise' et au corpus des voisinages

Au sein du corpus 'Crise financière', les domaines associés au sens codé ont tous des spécificités faibles, voire légèrement négatives (spécificité de -3 pour PHYSIOLOGIE). Les domaines du sens codé ne sont que faiblement repris par le cotexte de *toxique*. Autrement dit, il n'y a pas d'isotopie locale, donc pas d'activation des domaines considérés.

La confrontation des voisinages de *toxique* dans le corpus 'Crise financière' et dans le *Monde Diplomatique* vient confirmer ce qui ressort des résultats précédents et y apporte un regard complémentaire. Trois des cinq domaines du sémème, PHYSIOLOGIE, BIOLOGIE ET CHIMIE, sont nettement sous-représentés dans le corpus 'Crise financière' et les deux derniers domaines ne ressortent pas. Autrement dit, dans les anciens emplois incarnés par les données du *Monde Diplomatique*, la récurrence des domaines du sens codé est beaucoup plus marquée que dans les nouveaux emplois associés à la crise. Ceci pourrait s'expliquer par des isotopies domaniales dans le *Monde Diplomatique* qui ne se reproduisent pas dans le corpus 'Crise financière'. Les isotopies domaniales des emplois antérieurs disparaîtraient donc dans les nouveaux emplois, sans pour autant que les domaines du sens codé qui ne constituaient pas d'isotopie s'imposent et reconfigurent les patrons d'activation domaniale.

Le croisement des deux éclairages met donc en évidence une inhibition de l'ensemble des domaines communément associés à *toxique* et amène à conclure à une dédomanialisation de *toxique* dans les nouveaux emplois associés à la crise financière.

b- Enrichissement : émergence d'un nouveau domaine

Après la perte des domaines, l'enrichissement en nouveaux domaines est étudié à travers les domaines saillants, c'est-à-dire les plus spécifiques (voir tableau III.1.21).

[80] L'absence de valeurs correspond à une spécificité inférieure en valeur absolue au seuil fixé, qui est ici de 2.

Corpus	Corpus 'Crise financière' (découpage: voisinages de *toxique* et son complémentaire)			Corpus des voisinages de *toxique* du 'Monde Diplomatique' et de 'Crise financière'		
		Fréquence dans le sous-corpus	Spécif.		Fréquence dans le sous-corpus	Spécif.
Domaines les plus spécifiques	CARACTEROLOGIE	43	21	**FINANCES**	**192**	**27**
	PEDAGOGIE	49	15	CARACTEROLOGIE	43	18
	PSYCHANALYSE	54	14	DRAMATURGIE	32	12
	FINANCES	**192**	**8**	JEUX	148	11
	PHILOSOPHIE	183	5	PEDAGOGIE	49	10
				PSYCHANALYSE	54	8
				BOURSE	40	6

Tableau III.1.21 : Domaines les plus fortement surreprésentés au voisinage de toxique relativement aux deux corpus

Quatre domaines sont surreprésentés à la fois par rapport au corpus 'Crise' et par rapport aux voisinages des emplois du *Monde Diplomatique :* CARACTEROLOGIE, PEDAGOGIE, PSYCHANALYSE et FINANCES. Ils constituent des cibles privilégiées pour un enrichissement car ils distinguent *toxique* du contexte large dans lequel il apparaît, c'est-à-dire de l'ensemble du corpus 'Crise', et de ses anciens cotextes d'emploi, tels qu'ils apparaissent dans le *Monde Diplomatique*.

L'émergence des trois premiers domaines, CARACTEROLOGIE, PEDAGOGIE et PSYCHANALYSE, non conforme à l'analyse manuelle, résulte d'un biais dû à une annotation en traits sémantiques sans désambiguïsation préalable d'*actifs*, cooccurrent privilégié de *toxique*.[81] FINANCES est le seul des quatre domaines à ne pas être affecté par ces biais. Il est saillant à la fois par rapport aux anciens voisinages de *toxique*, et, cas plus remarquable, au sein même du corpus 'Crise financière', où la finance est pourtant omniprésente. Son rang dans le classement des domaines par spécificité décroissante et la valeur des spécificités élevées témoignent de son importance. Ces indices soulignent la présence d'une isotopie locale du domaine FINANCES et invitent à voir une redomanialisation de *toxique* dans le domaine financier.

D'autres domaines n'émergent que dans un seul jeu de contraste :

- Les domaines JEUX et DRAMATURGIE ne sont surreprésentés que par rapport aux voisinages dans les anciens emplois. Dans le corpus 'Crise financière' en revanche, leur comportement n'est pas remarquable, autrement dit, JEUX et DRAMATURGIE s'expriment à peu près autant, voire plus dans le reste du corpus que dans les voisinages de *toxique*. On peut donc en déduire qu'il s'agit de domaines plus caractéristiques du corpus que du sens de *toxique* et qu'ils pourraient éventuellement correspondre à une isotopie textuelle.
- Le domaine PHILOSOPHIE ressort au sein du corpus 'Crise' avec une spécificité de 5 mais n'est que faiblement surreprésenté dans le corpus des voisinages (spécificités de 2).

[81] Les trois domaines proviennent presque exclusivement d'une même forme, *actif*, qui compte 41 occurrences dans le sous-corpus considéré. Or *actif* est un cooccurrent lexical privilégié de *toxique*, mais son sens désigne alors un produit financier, non un type de caractère ni un individu qui s'implique émotionnellement ou de manière participative. L'absence de désambiguïsation des définitions et l'association quasi-exclusive de domaines rares à certains lemmes fait donc émerger des domaines non pertinents. Plusieurs solutions permettent de faire face à ce problème : un filtrage en fonction du nombre de lemmes sources ; un filtrage en fonction de la présence dans le TLFi ; la mise en place d'une procédure de désambiguïsation avant annotation, qui exige des investigations poussées et qui mériterait de faire l'objet de travaux ultérieurs. On reviendra rapidement sur ce point au (chapitre III.2, 4.1).

L'interprétation reste délicate, mais on peut supposer que *toxique* a tendance à attirer dans ses voisinages un certain nombre de lexies porteuses du domaine PHILOSOPHIE.

3. Niveau lexical : préciser les domaines et préparer l'approche en traits sémantiques

Une fois la caractérisation domaniale établie, l'analyse du nouveau sens est réalisée à partir de descripteurs de granularité plus fine, relatifs aux domaines, à savoir les unités lexicales. Dans un second temps, le nouveau sens sera qualifié à l'aide de descripteurs de granularité encore plus fine, les traits sémantiques, étudiés dans la section 4.

Les domaines donnent une idée assez générale du nouveau sens de la cible lexicale. La sélection d'unités lexicales relativement à un domaine permet de préciser le sens au sein de ce domaine. Les unités lexicales ont un rôle intermédiaire, elles sont destinées à établir des liens avec le niveau de granularité sémantique inférieure, c'est-à-dire avec les traits sémantiques (sèmes microgénériques ou spécifiques).

Tout comme pour les domaines, le rôle et les propriétés des unités lexicales seront précisés. Ensuite, on détaillera l'extraction d'unités lexicales saillantes selon leur statut (les cooccurrents, puis les concurrents).

3.1 Caractéristiques des unités lexicales

Les caractéristiques des unités lexicales seront abordées sous trois angles, correspondant aux trois autres types d'observables auxquels elles s'articulent : la cible lexicale qu'elles qualifient, les domaines dont elles dépendent et les traits sémantiques dont elles préparent l'étude.

3.1.1 Deux types de relations à la cible : cooccurrents d'ordre 1 et d'ordre 2

Pour qualifier le sens de la cible, on considèrera comme informatives les unités lexicales qui présentent des affinités avec la cible par combinaison ou par substitution. Les unités en affinité par combinaison seront recherchées parmi les cooccurrents simples, ou cooccurrents d'ordre 1. Les unités en affinité par substitution seront recherchées parmi les cooccurrents d'ordre 2. Elles correspondent à un ensemble d'unités lexicales qui partagent un même ensemble de cooccurrents. On parlera dans ce dernier cas de paradigme de substitution ou de concurrents.

L'unité textuelle où seront recherchés les cooccurrents d'ordre 1 est le paragraphe. Celui qui définit la cooccurrence d'ordre 2 est le syntagme. Nous renvoyons le lecteur au (chapitre II.2, 1.1.1 et 1.1.2) pour plus de précisions sur l'interprétation de chaque type de cooccurrents et sur le choix de l'unité textuelle.

3.1.2 Dépendance aux domaines

Notre analyse repose sur des observations par granularité décroissante. Les observables plus fines sont destinées à préciser les informations apportées par les observables de granularité supérieure. Autrement dit, les unités lexicales précisent les domaines et elles s'étudient de façon relative aux domaines. A ce titre, elles caractérisent le sens de différentes façons :

- (a) comme unités communes à différents domaines d'emploi ;
- (b) comme unités propres à un domaine et relevant de la thématique générale de ce domaine ;

- (c) comme unités propres à un domaine d'emploi, sans pour autant relever de la thématique générale.

a- Unités communes à plusieurs domaines d'emploi

Elles distinguent les paliers de voisinage locaux du reste du domaine et sont aussi caractéristiques de l'environnement lexical dans d'autres domaines.

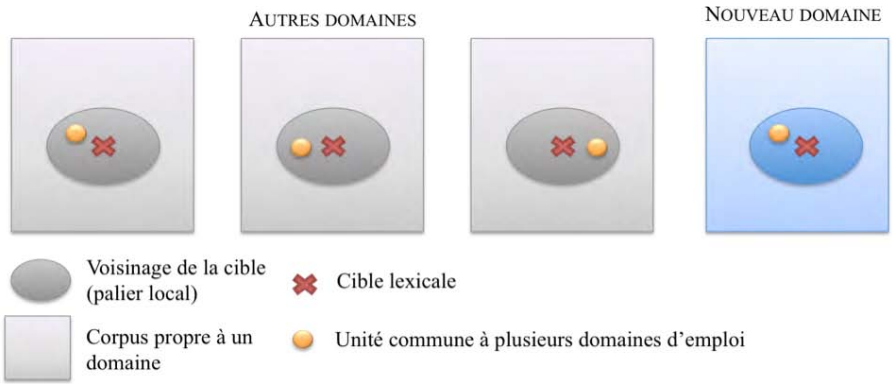


Figure III.1.22.a : Articulation d'une unité lexicale aux domaines – présence transverse à différents domaines

Ces unités témoignent d'une activation de traits sémantiques, elles lexicalisent des traits de l'ancien sémème et permettent de créer une passerelle entre acceptions. C'est le cas d'unités renvoyant à l'idée de violence de *tsunami* dans le contexte de crise financière (cette idée est aussi présente dans d'autres domaines), ou à l'idée de diffusion et d'impact négatif de *toxique.* Ces unités sont typiques du mécanisme qui constitue un des ressorts de la métaphorisation.

b- Unités propres au nouveau domaine d'emploi et relevant de la thématique générale

Elles distinguent l'environnement lexical du domaine de l'environnement lexical des autres domaines d'emploi.

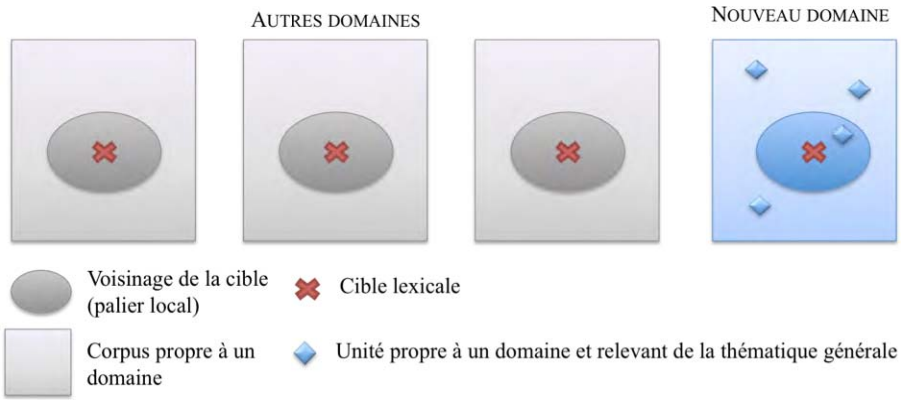


Figure III.1.22.b : Articulation d'une unité lexicale aux domaines – présence générale au sein du nouveau domaine

C'est notamment le cas des unités lexicales *crise*, *banque* ou *financier* en affinité avec *toxique*. Elles contribuent à un enrichissement qui intègre la thématique du domaine, donc elles sont plutôt de type générique.

c- Unités propres au nouveau domaine d'emploi et différentes de la thématique générale

Elles se distinguent de la thématique générale du domaine et apportent des précisions internes au domaine, sans pour autant être partagées par d'autres emplois.

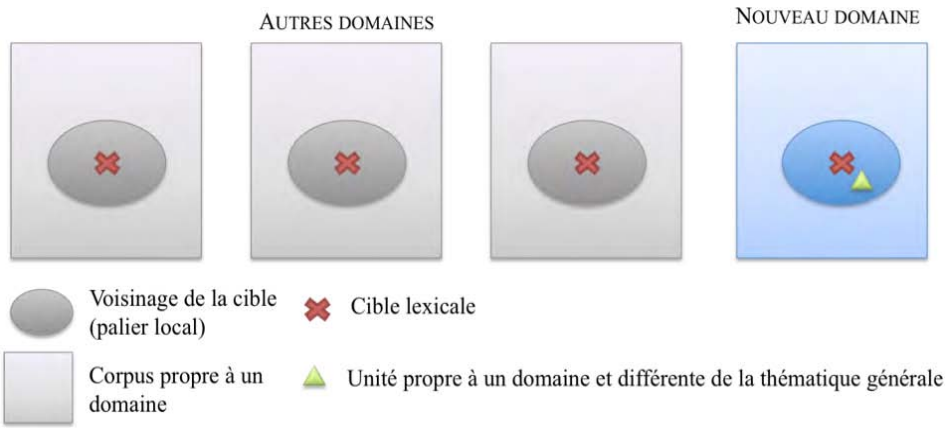


Figure III.1.22.c : Articulation d'une unité lexicale aux domaines – présence locale au sein du nouveau domaine

Ces unités lexicales participent à l'enrichissement sémantique. Lorsque ces unités seront rattachées aux traits sémantiques, les informations sémantiques qu'elles apportent pourront être associées à de nouveaux traits sémantiques qui correspondront à des sèmes spécifiques. Par exemple, il s'agira du nouveau lien sémantique de *toxique* avec des produits ou instruments financiers (*crédit, actif, titre*).

d- Bilan sur l'articulation aux domaines

Les unités lexicales apportent des informations de nature différente selon les croisements effectués à partir des critères suivants :

- critère du nombre de domaines d'études, avec une restriction à un domaine ou la comparaison de plusieurs domaines ;
- critère de voisinage local, avec un caractère distinctif ou non des voisinages locaux par rapport à un palier global.

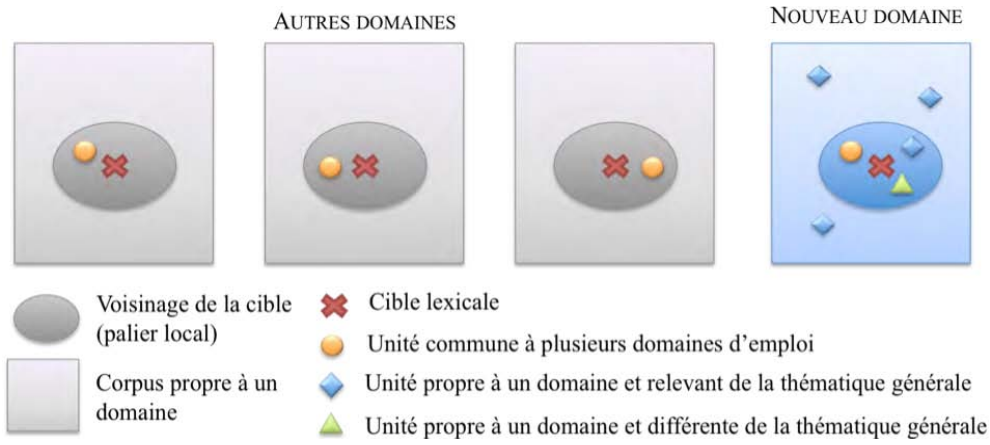


Figure III.1.22.d : Articulation d'une unité lexicale aux domaines – bilan

Le tableau ci-dessous illustre le type d'information apporté par chaque croisement à travers l'exemple de *toxique*.

Unités en affinité avec la cible (cooccurrents ou concurrents)		Critère du nombre de domaines d'étude	
		Propre au nouveau domaine d'emploi	**Communes aux autres domaines d'emploi**
Critère de voisinage local	**Propres à un palier local dans le nouveau domaine**	Affinité avec les produits ou instruments financiers (*crédit, actif, titre)*	Idée de diffusion et caractère négatif de *toxique*
	Non distinctives du palier local (en lien avec la thématique générale)	*crise* ou *financier* pour *toxique*	/

Tableau III.1.23 : Apports relatifs à chaque croisement entre l'axe d'analyse des domaines et l'axe d'analyse local-global

3.1.3 Des unités intermédiaires, supports de la récurrence de traits sémantiques

On cherche des unités lexicales qui distinguent l'environnement lexical par effet de masse, donc statistiquement. Ceci amène à exclure des unités lexicales rares. Mais, d'après la théorie de l'information, plus un événement est rare, plus il est informatif. L'hypothèse qui amène à exclure les unités rares est que les traits sémantiques pertinents exprimés par les unités rares s'expriment aussi à travers des unités lexicales plus fréquentes. Autrement dit, on s'appuie sur une hypothèse d'isotopie locale, c'est-à-dire de récurrence de traits sémantiques : un trait sémantique pertinent sera présent de façon répétée dans les voisinages ; cette répétition pourra être portée par des unités rares mais, pour qu'elle soit suffisamment marquée, elle devra aussi être portée par des unités plus fréquentes.

L'objectif n'est donc pas de retourner les unités lexicales les plus pertinentes, ni de retourner toutes les unités lexicales pertinentes, mais de retourner des unités qui lexicalisent des traits sémantiques suffisamment variés et appropriés pour donner une image précise et riche du nouveau sens. Autrement dit, à travers l'idée de lexicalisation, on suppose que les traits sémantiques 'prennent corps', ou encore s'expriment de façon privilégiée à travers des unités lexicales. Un ensemble d'unités lexicales sera donc l'expression d'un ou plusieurs traits sémantiques ; chaque trait sémantique s'exprimera de façon plus ou moins marquée (par exemple, pour les *tablettes* numériques, l'adjectif *numérique* évoque bien les nouvelles technologies, mais peu l'idée de tactile, et pas du tout l'idée de format réduit). On cherche un ensemble d'unités lexicales capable d'exprimer les différentes facettes sémantiques caractéristiques du nouveau sens (donc, pour les *tablettes* numériques, à la fois les idées du format réduit, mais aussi du lien avec les nouvelles technologies et du caractère tactile).

3.2 Sélection et pondération des cooccurrents

La sélection et la structuration de cooccurrents pertinents reposent sur l'hypothèse de distributionnalité de (Harris, 1968) , à savoir que les mots qui apparaissent dans les mêmes contextes tendent à avoir le même sens. L'affinité sémantique avec la cible, ou encore la significativité d'une cooccurrence, s'obtient en jouant sur différents contrastes en fonction des paramètres dont on dispose (domaines, localité/globalité et périodes de temps) : des contrastes entre voisinages relatifs aux anciens et aux nouveaux domaines d'emploi ; des contrastes au sein d'un domaine entre palier local (voisinage) et palier global ; des contrastes entre plusieurs périodes.

Le palier de cooccurrence retenu est le paragraphe, qui est communément utilisé dans les études de textométrie, par exemple, (Mayaffre, 2008), (Rossignol et Sébillot, 2002). L'ensemble initial de cooccurrents regroupe donc les unités lexicales présentes dans au moins un paragraphe contenant la cible.

3.2.1 Contrastes entre voisinages relatifs aux anciens et nouveaux domaines d'emploi

Les cooccurrents du ou des nouveaux domaines sont contrastés avec les cooccurrents des autres domaines d'emplois. Ce type de contraste est approprié pour faire émerger des saillances propres au nouveau domaine, notamment des unités qui relèvent de la thématique du nouveau domaine.

Considérons le cas de *toxique*. Ses emplois dans le corpus 'Crise financière' sont comparés à ceux du 'Monde Diplomatique'[82]. Le tableau ci-dessous présente la liste d'unités significatives dans le sous-corpus 'Crise', après un calcul de spécificités sur les formes lexicales, l'élimination des mots-outils (déterminants, pronoms, prépositions) puis l'application d'un seuil de spécificité de 3.

rg	Forme	Spéc.	rg	Forme	Spéc.	rg	Forme	Spéc.	rg	Forme	Spéc.
1	banques	32	7	financiers	9	14	taux	6	21	argent	4
2	actifs	23	8	financière	8	15	produits	5	22	économie	3
3	crise	15	9	hier	8	16	fonds	5	23	américain	3
-	toxiques	11	10	plan	8	17	système	5	24	situation	3
4	milliards	11	11	banque	7	18	avoir	4	25	a	3
5	Paulson	11	12	crédits	7	19	dollars	4	26	gouvernement	3
6	Trésor	9	13	sauvetage	7	20	marchés	4			

Tableau III.1.24 : Unités lexicales les plus spécifiques des voisinages de toxique du corpus 'Crise financière' par rapport aux voisinages du corpus du Monde Diplomatique

Il apparaît de nombreuses unités relevant de la thématique générale de la crise financière et qui évoquent le domaine de la finance : *crise, financier, financière, argent, économie, dollars, banque, milliards* ou encore *marchés*. Une unité plus spécifique du nouveau sens de *toxique*, *actifs,* apparaît également en tête de liste. Les unités *sauvetage* et *plan* sont indirectement liées au caractère néfaste de *toxique*, elles expriment une réaction face à un danger ou une catastrophe.

3.2.2 Contrastes entre palier local et palier global d'un domaine donné

Les voisinages de la cible dans le domaine ciblé (paragraphes contenant la cible) sont comparés à l'ensemble des textes du domaine. La thématique générale a tendance à moins ressortir. En revanche, cette démarche permet de faire ressortir les isotopies locales distinctes des isotopies globales. Des unités qui expriment des facettes du sens codé communes au nouvel emploi et aux anciens emplois auront aussi tendance à émerger.

À titre d'illustration, considérons l'exemple de *toxique* au sein du corpus 'Crise financière'. Le tableau ci-dessous présente les unités lexicales de spécificité supérieure à 3. Comme la taille du corpus est plus importante que précédemment, les valeurs des spécificités ont tendance à être plus importantes en valeur absolue, donc la liste d'unités spécifiques est plus importante. On s'est limité aux 50 premières unités, pour conserver une liste de taille modérée par rapport à la précédente.

[82] Le corpus des voisinages de *toxique* provenant du corpus 'Crise financière' et du *Monde Diplomatique* comte 21 956 formes. La taille du corpus des voisinages est réduite, de ce fait, le nombre d'unités significatives est relativement réduit, car la taille du corpus influe sur la gamme de spécificités. L'annotation sémique présentée dans la section 4 augmentera la quantité de données et, de ce fait, permettra d'obtenir un plus grand nombre d'unités saillantes.

rg	Forme	Sp	rg	Forme	Sp	rg	Forme	Sp	rg	Forme	Sp
	toxiques	***	13	Henry	6	26	fédéral	5	39	fédérale	4
1	actifs	38	14	transparence	6	27	subprime	5	40	élus	4
2	banques	15	15	crédits	6	28	américaines	4	41	comportement	4
3	produits	14	16	titres	6	29	milliards	4	42	Stiglitz	4
4	Paulson	11	17	collectivités	6	30	semblait	4	43	hier	4
5	Trésor	10	18	détenus	6	31	territoriales	4	44	taux	4
6	variables	10	19	garantir	5	32	diagnostic	4	45	apportée	4
7	bad	9	20	Seine-Saint-Denis	5	33	représentant	4	46	préférentielles	4
8	Citigroup	9	21	adossés	5	34	traitement	4	47	Atlantique	3
9	créances	8	22	sauvetage	5	35	instruments	4	48	douteux	3
10	financiers	7	23	racheter	5	36	rachat	4	49	jugés	3
11	bilans	7	24	initial	5	37	frais	4	50	permettent	3
12	emprunts	7	25	contribuable	5	38	nationaliser	4			

Tableau III.1.25 : Unités lexicales les plus spécifiques des voisinages de toxique du corpus 'Crise financière' par rapport au reste du corpus 'Crise financière'

Au niveau de la thématique générale, seules quelques unités relevant de la thématique générale et qui avaient émergé précédemment se retrouvent en tête de liste : on retrouve *financiers* (spécificité de 7 ; rang 10) et *banques* (spécificité de 15 ; rang 2), tandis que *financier, économie, crise, argent* ou *dollars* n'apparaissent pas dans les 50 premières unités les plus significatives. Plusieurs unités lexicales expriment des isotopies locales qui sont plus précises que la thématique générale. Ainsi, plusieurs substantifs associés aux produits et instruments financiers émergent en tête de liste, notamment *actifs*, *créances, emprunts, bilans, crédits* et *titres*. Les facettes du sens codé communes aux anciens et nouveaux emplois apparaissent de façon plus diffuse à travers des unités telles que *sauvetage* et *douteux*.

3.2.3 Contrastes entre différentes périodes de temps

Pour affiner les résultats précédents et pour différencier les unités lexicales saillantes en fonction de l'évolution dans le temps, on observe des séries chronologiques, à partir d'un découpage du corpus en périodes de temps. Ceci permet de voir s'il y a stratification du sens, c'est-à-dire si les nouveaux cooccurrents évoluent au cours du temps, ou à l'inverse, de repérer des cooccurrents stables, présents quelle que soit la période. Ceci permet également de regrouper les cooccurrents s'ils présentent des profils d'évolution chronologique similaires. À nouveau, les indices servant à comparer les périodes peuvent résulter de plusieurs contrastes : comparaison des paragraphes du domaine entre eux, période par période ; comparaison des saillances relatives à chaque période (à période fixée, saillance dans les paragraphes du domaine par rapport au reste du domaine ou saillance dans les paragraphes du domaine par rapport aux paragraphes des autres domaines).

Dans l'expérience sur le corpus 'Outreau', la comparaison des voisinages d'*Outreau* (paragraphes contenant *Outreau*) au reste du corpus, toutes périodes confondues, fait émerger les unités présentées en première colonne du tableau ci-dessous. Le découpage du corpus en cinq périodes permet de préciser les résultats précédents et d'observer l'évolution des unités lexicales au cours du temps. Les colonnes suivantes du tableau présentent les 20 unités les plus significatives à chacune des 5 périodes[83].

[83] La comparaison des différentes périodes repose sur une comparaison de rangs et non de valeurs de spécificité, car les sous-corpus propres à chaque période sont de taille disparate, ce qui influe sur la gamme de valeurs prises par les spécificités (les valeurs sont plus tassées pour des sous-corpus plus petits et plus étendues pour des sous-corpus pus grands).

Ensemble des périodes	Période 1	Période 2	Période 3	Période 4	Période 5
pédophilie	Outreau	Outreau	Karine	Perben	Outreau
Pas-de-Calais	examen	pédophilie	Outreau	Outreau	commission
Outreau	réseau	procès	Anthony	Sceaux	acquittés
procès	quartier	Pas-de-Calais	David	ministre	affaire
affaire	fouilles	affaire	soirée	verdict	Bot
assises	Boulogne-sur-Mer	assises	évènement	groupe	parlementaire
Saint-Omer	fillette	Saint-Omer	plaidoiries	garde	Yves
accusés	pédophile	cour	attendu	acquittés	députés
cour	pédophilie	accusés	verdict	justice	institution
pédophile	corps	spécial	procureurs	détention	paris
justice	jardin	mai	cauchemar	vendredi	justice
réseau	janvier	huis	inceste	affaire	Georges
spécial	mises	victimes	jeudi	provisoire	magistrats
judiciaire	ouvrier	présumés	Delplanque	condamné	réforme
hier	été	envoyé	avocat	procès	ministre
Perben	personnes	clos	distinguer	mois	détention
Boulogne-sur-Mer	ville	judiciaire	jury	sursis	provisoire
envoyé	meurtre	pédophiles	affaire	magistrats	excuses
examen	agressions	théâtre	dénoncer	juillet	procureur
quartier	mort	hier	âgé	provisoires	leçons

Tableau III.1.26 : Unités lexicales les plus spécifiques des voisinages d'Outreau pour chacune des 5 périodes du corpus 'Outreau'

Conformément à l'analyse manuelle (*cf.* 1.2.2), la dimension policière apparaît en période 1 (*fouilles, meurtres*), ainsi que la notion de réseau pédophile (*réseau, pédophilie*) et ce qui se rattache au lieu (*quartier, ville*). Le judiciaire apparaît nettement aux périodes suivantes. Par contre, l'émotion populaire est moins sensible. De même, l'idée de fiasco et d'erreur judiciaire n'apparaît quasiment pas en période 5, si ce n'est à travers *excuses.* Ce sont ces idées latentes qu'on cherchera à mettre en évidence à travers les analyses en traits sémantiques de la section 4.

3.3 Sélection et pondération du paradigme des cooccurrents d'ordre 2

Les cooccurrents d'ordre 2 sont susceptibles de définir des classes lexicales, sémantiquement cohérentes, auxquelles appartient la cible lexicale. Ils correspondent à des unités qui apparaissent avec les mêmes cooccurrents d'ordre 1 de la cible. Le palier de cooccurrence retenu pour sélectionner les cooccurrents d'ordre 2 est le syntagme. Les cooccurrents d'ordre 2 apparaissent donc dans des syntagmes qui partagent des composantes avec les syntagmes contenant la cible. Ils tendent à qualifier ou à être qualifiés par les mêmes corrélats. Comme ils interviennent dans les mêmes combinaisons, on peut considérer qu'ils répondent à un critère de substitution. Ils incluent les unités participant au phénomène de *foisonnement néologique* (chapitre I.2, 3.3*)*, c'est-à-dire l'ensemble des unités concurrentes d'une unité néologique.

Comme pour les cooccurrents d'ordre 1, les cooccurrents d'ordre 2 peuvent être définis par comparaison de domaines ou au sein d'un domaine[84]. Dans les expériences réalisées, les cooccurrents d'ordre 2 sont obtenus par comparaison de différents domaines. Les cooccurrents d'ordre 2 sont recherchés dans le corpus 'Crise financière', qui est contrasté avec un corpus de

[84] La définition des cooccurrents d'ordre 2 saillants au sein d'un domaine est la suivante : si A apparaît de façon significative dans les syntagmes contenant la cible relativement à toutes les occurrences de A dans le corpus, et si B apparaît de façon significative dans les syntagmes contenant A relativement à toutes les occurrences de B dans le corpus, alors B sera étroitement associés à la cible et sera un cooccurrent d'ordre 2 saillant.

référence, témoin d'emplois antérieurs. Le corpus de référence correspond à des articles variés du journal *Le Monde* de l'année 2002, constitué d'environ 28 millions d'occurrences. Il est intégré au logiciel TermoStat (Drouin, 2003), qui a permis d'établir le paradigme des concurrents associé à chaque cible lexicale (*cf. infra*). Le corpus 'Crise financière' est analogue en genre au corpus de référence, puisqu'il s'agit aussi de presse généraliste. Il s'en distingue thématiquement et chronologiquement. L'écart chronologique est pertinent pour étudier la néologie, de façon analogue à (Drouin *et al.*, 2006). L'écart thématique participe de l'émergence de tropes, sources de néologismes, par interaction entre le domaine de la finance et d'autres domaines éloignés. Les cibles lexicales sont *bouclier, pourri, toxique, tempête, tsunami* et *actif*. Ces cibles ont été choisies parce qu'elles apparaissent dans des syntagmes perçus comme métaphoriques, sauf *actif* qui sert de témoin. Certaines cibles ont un caractère néologique dans le corpus 'Crise', à savoir *toxique, tsunami, bouclier*, et dans une moindre mesure *pourri*. Par contre, l'emploi métaphorique de *tempête* n'est pas nouveau.

Dans le cadre expérimental, la cooccurrence est définie au sein de syntagmes noms-adjectifs. Les cooccurrents d'ordre 1 d'une cible substantif seront des adjectifs, et inversement. Les cooccurrents d'ordre 2 d'une cible substantif seront des substantifs qui partagent des cooccurrents adjectivaux avec la cible, et inversement si la cible est un adjectif. Les cooccurrents d'ordre 1 jouent le rôle de pivots : ils articulent les autres constituants des syntagmes auxquels ils appartiennent avec la cible. Considérons l'exemple de *toxique*. Dans le corpus 'Crise', cet adjectif apparaît dans les syntagmes *emprunts toxiques, crédits toxiques* et *titres toxiques*. *Emprunt, crédit* et *titre* sont donc cooccurrents d'ordre 1. *Emprunt* apparaît dans les syntagmes *emprunts européens, emprunts immobiliers, emprunts publics*, en plus d'*emprunts toxiques*. Ainsi, *européen, immobilier* et *public* sont cooccurrents d'ordre 2.

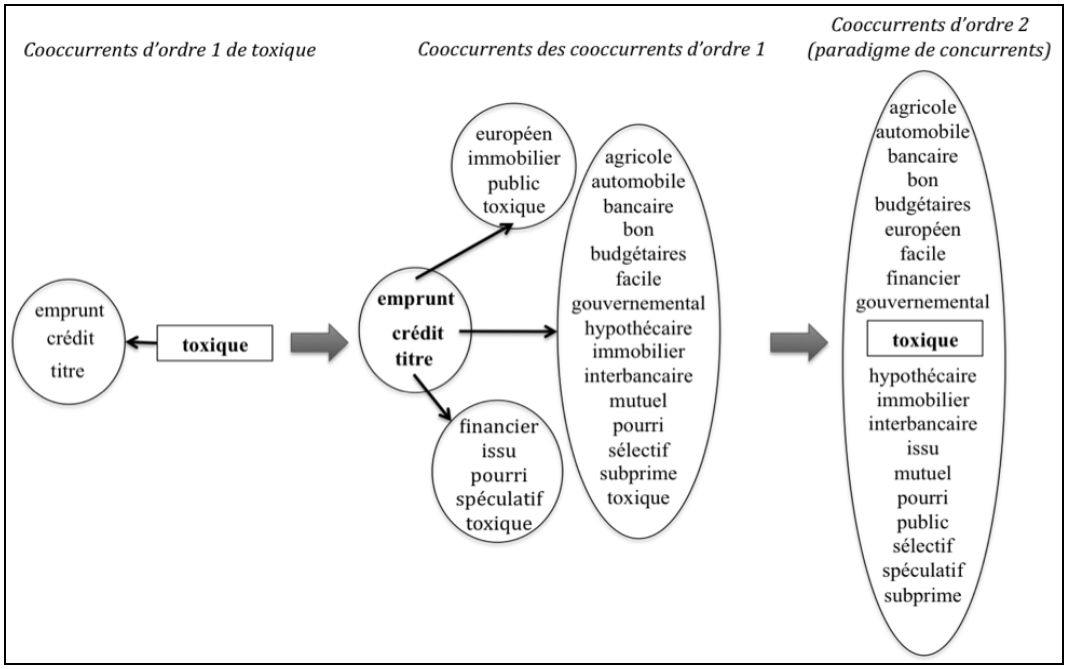


Figure III.1.27 : Construction d'un paradigme de cooccurrents d'ordre 2

Comme évoqué ci-dessus, la sélection de cooccurrents d'ordre 1 ou 2 se fait à l'aide de la plateforme TermoStat. Celle-ci extrait des unités lexicales simples ou complexes à partir de leur spécificité[85] dans un corpus d'analyse relativement à un corpus de référence. Le seuil

[85] Les spécificités ne sont pas celles utilisées dans les autres expériences, qui proviennent de (Lafon, 1984), même si elles proviennent aussi du modèle hypergéométrique. Elles sont calculées selon la méthode décrite par (Lebart et Salem, 1994) : elles correspondent à la valeur-test, indicateur statistique qui se rattache à l'écart-type

minimal de spécificité est de 3 et le seuil de fréquence de 2. Les unités complexes de type nom-adjectif sont utilisées pour construire le paradigme de cooccurrents d'ordre 2. Autrement dit, l'ensemble des cooccurrents d'ordre 2 comme dans l'exemple ci-dessus est d'emblée restreint aux unités lexicales spécifiques d'un pivot, lui-même spécifique de la cible.

Pondération des cooccurrents d'ordre 2. Un coefficient d'affinité avec la cible est affecté à chaque cooccurrent d'ordre 2 en fonction des distributions de spécificités affectées aux couples noms-adjectifs par TermoStat. Il est calculé à partir des spécificités relatives à l'ensemble des cooccurrents d'ordre 1, établies à la fois entre la cible et les cooccurrents d'ordre 1 et entre les cooccurrents d'ordre 1 et le cooccurrent d'ordre 2 considéré[86]. Les affinités calculées entre la cible lexicale et les éléments de son paradigme permettent d'ordonner les éléments du paradigme par affinité distributionnelle décroissante, comme illustré ci-dessous pour les cooccurrents d'ordre 2 de *tsunami*.

rg	cooc2	aff	rg	cooc2	aff	rg	cooc2	aff	rg	cooc2	aff
1	crise	1604	25	titre	125	48	instrument	70	73	autorité	60
2	système	761	26	produit	124	50	sommet	68	73	rendement	60
3	marché	566	27	capital	118	50	bourrasque	68	75	échange	58
4	capitalisme	464	28	spéculation	117	50	carte	68	76	aide	55
5	secteur	443	29	réunion	114	50	chapitre	68	76	puissance	55
6	institution	412	29	accumulation	114	50	choix	68	78	désorde	54
7	établissement	351	29	mathématicien	114	50	coup	68	78	chaos	54
8	rentabilité	333	32	entreprise	109	50	crash	68	80	notation	53
9	tempête	286	33	sauvetage	107	50	délinquant	68	81	soutien	52
10	tourmente	262	34	industrie	105	50	disposition	68	81	objectif	52
11	croissance	236	35	valeur	104	50	district	68	83	pôle	51
12	planète	222	35	organisme	104	50	édifice	68	84	globalisation	49
13	sphère	212	37	activité	102	50	front	68	84	technique	49
14	débâcle	192	38	placement	100	50	gonflement	68	86	risque	48
15	régulation	181	39	solidité	97	50	liquidité	68	87	domaine	46
16	turbulence	175	40	responsabilité	95	50	machinerie	68	88	profit	43
17	stabilité	170	41	cancer	93	50	mathématique	68	89	service	40
18	krach	168	41	élite	93	50	ouragan	68	89	assainissement	40
19	acteur	162	43	revenu	91	50	note	68	91	opération	39
20	place	157	44	bulle	82	68	circuit	67	91	matière	39
21	monde	155	45	assurance	81	69	mondialisation	64	93	investissement	38
22	tsunami	146	46	dérive	80	70	environnement	63	94	société	37
23	innovation	139	46	urgence	80	70	catastrophe	63			
24	séisme	134	48	libéralisme	70	70	juridiction	63			

Tableau III.1.28 : Cooccurrents d'ordre 2 de* tsunami *classés par degré d'affinité (aff) décroissant

3.4 Informations extraites de l'environnement lexical

Aussi bien pour les cooccurrents d'ordre 1 que pour les cooccurrents d'ordre 2, on n'étudie pas les unités lexicales, mais les unités lexicales sachant le domaine retenu, et éventuellement la période de temps. Autrement dit, les données correspondent à des paires (unité lexicale, domaine), voire à des triplets (unité lexicale, domaine, période), auxquels sont affectées des valeurs quantitatives, en l'occurrence des spécificités. Ces valeurs reflètent le degré d'affinité

d'une distribution observée par rapport à la distribution théorique, approximation normale d'une distribution hypergéométrique.

[86] Considérons l'ensemble des noms et adjectifs des syntagmes contenant la cible et des syntagmes à composante commune. Tout couple (nom – adjectif) se voit affecter un coefficient : la spécificité retournée par TermoStat, 0 sinon. Chaque nom peut se voir comme un vecteur sur l'espace des adjectifs, et inversement. L'affinité entre deux noms (resp. adjectifs) est obtenue par produit scalaire entre les vecteurs correspondants.

sémantique local ou paradigmatique, pour le domaine considéré et, le cas échéant, pour la période considérée.

Les spécificités permettent :

- d'extraire un ensemble d'unités lexicales saillantes propre au domaine, à partir d'un seuil minimal de spécificités et un seuil minimal de fréquence (pour l'effet de masse et la validité des traitements statistiques) ;
- de hiérarchiser les unités précédemment citées, par spécificité décroissante ;
- de regrouper ces unités à partir d'un découpage en plusieurs périodes de temps, lorsque les profils d'évolution sont similaires, et de répartir les groupes d'unités en fonction de leur apport : participation à une stratification du sens ou apport d'une composante sémantique stable dans le temps ;
- de visualiser l'évolution du sens ou les différences entre domaines d'emplois (enrichissement des représentations correspondant aux empreintes de fréquence domaniales).

L'information apportée par les traits sémantiques permet d'aller un cran plus loin dans l'analyse du contenu sémantique : elle permet de faire le lien avec le sens codé, d'enrichir l'information apportée par le plan lexical et de structurer les unités lexicales.

4. Niveau infra-lexical : préciser le niveau lexical et structurer les traits sémantiques

Tout comme pour les domaines, les informations issues du niveau lexical sont propres à la structure du corpus, elles reflètent des variations dans les textes. Il faut alors associer ces variations textuelles à la ressource lexicographique, afin de repositionner les variations lexicales par rapport au sens codé. Plus précisément, au niveau supra-lexical, les domaines textuels ne suffisaient pas pour établir le lien avec le sens codé, il était nécessaire de se ramener aux domaines lexicographiques. De même, au niveau inférieur, on ne se limitera pas aux unités lexicales, propres au texte, mais on se ramènera aux unités associées à la ressource lexicographique, c'est-à-dire aux traits sémantiques. Deux stratégies sont possibles :

- ***projeter les traits sémantiques dans les textes,*** puis utiliser la structure textuelle pour étudier les saillances et regroupements de traits sémantiques (sous-section 4.1) ;
- ***projeter les unités lexicales dans la ressource sémique***, puis utiliser la structure du dictionnaire pour organiser les unités lexicales, les réorganiser ou les analyser en termes de traits sémantiques (sous-section 4.2).

Quelle que soit la stratégie, la recherche d'informations sur le contenu sémantique est sous-tendue par les principes suivants :

- Un ***principe de lexicalisation des traits sémantiques :*** les unités lexicales peuvent être considérées comme une lexicalisation privilégiée de traits sémantiques. Les traits sémantiques peuvent être utilisés pour structurer un ensemble d'unités lexicales ou bien pour expliciter le lien sémantique implicite entre deux unités lexicales.
- Un ***principe d'isotopie locale :*** les unités lexicales expriment de façon récurrente un trait sémantique. La récurrence du trait sémantique peut prendre deux formes : elle peut être évaluée en fonction du nombre d'occurrences des unités qui expriment le trait sémantique ou en fonction de la diversité des unités lexicales porteuses du trait sémantique (nombre

d'unités distinctes exprimant le trait sémantique)[87]. Ces deux formes de récurrence permettront de définir l'importance du trait sémantique.

À nouveau, comme pour les domaines, l'organisation de l'information apportée par les traits sémantiques dépendra deux types de relations avec le sens codé : ce qui relève d'une reconfiguration du sémème et ce qui relève d'un enrichissement.

4.1 Projection de l'infra-lexical dans les discours et recoupements avec le niveau lexical

L'approche qui repose sur une projection des traits sémantiques dans les discours est abordée en plusieurs étapes : nous présentons d'abord comment le niveau infra-lexical, c'est-à-dire le niveau des traits sémantiques, est projeté dans les discours par annotation de corpus (4.1.1), puis comment cette annotation est exploitée (4.1.2 à 4.1.4). L'exploitation de l'annotation est décrite dans ses grands principes (4.1.2), elle est ensuite étudiée sous l'angle de la reconfiguration du sens codé (4.1.3) et enfin sous l'angle de l'enrichissement du sens codé (4.1.4).

4.1.1. Projection de l'infra-lexical par annotation du corpus en traits sémantiques

La projection des traits sémantiques dans les discours s'obtient en affectant aux unités lexicales leur contenu sémantique. On substitue à toute unité lexicale le sémème qui lui est associé.

Dans nos expériences, l'annotation de corpus en traits sémantiques a été réalisée à partir de la plate-forme Semy (Grzesitchak 2008 ; *cf.* chapitre II.1, 1.2.2). Seuls ont été enrichis en traits sémantiques les noms, verbes, adjectifs et adverbes, c'est-à-dire les unités lexicales considérées comme sémantiquement pleines. L'annotation repose sur une approche en sacs de traits sémantiques : on perd la structure syntaxique du texte, autrement dit, toute suite de mots sémantiquement pleins dans le texte devient un sac non structuré de traits sémantiques. La délimitation maintenue par Semy est celle des paragraphes. L'annotation s'accompagne d'un accroissement du nombre d'occurrences (taille multipliée par 25 environ). Le vocabulaire ne connaît pas un tel accroissement : il reste du même ordre de grandeur pour les deux plus grands corpus (4.10^5 formes et 1.10^6 formes) pour un vocabulaire de taille relativement proche, et il est multiplié par 2,5 pour le corpus des voisinages.

Corpus 'Crise financière'	Formes		Traits sémantiques	
	Nombre d'occurrences	920 551	Nombre d'occurrences	23 198 346
	Nombre de formes	35 147	Nombre de traits sémantiques	29 661
Corpus des voisinages (*Crise financière* et *Monde Diplomatique)*	Formes		Traits sémantiques	
	Nombre d'occurrences	21 956	Nombre d'occurrences	524 032
	Nombre de formes	5 642	Nombre de traits sémantiques	14 080
Corpus 'Outreau'	Formes		Traits sémantiques	
	Nombre d'occurrences	398 019	Nombre d'occurrences	9 691 627
	Nombre de formes	24 109	Nombre de traits sémantiques	24 581

Tableau III.1.29 : Taille des corpus avant et après annotation sémique

L'environnement de traits sémantiques peut être défini par rapport aux cooccurrents d'ordre 1 ou 2. L'annotation maintient la même structure pour l'environnement de traits que pour

[87] Cette distinction est de même nature que celle faite communément sur la taille du corpus et la taille du vocabulaire, sauf que le "corpus" et le "vocabulaire" sont extrêmement restreints : il se limitent aux unités qui expriment le trait sémantique étudié.

l'environnement lexical. Pour les cooccurrents d'ordre 1, les frontières des paragraphes contenant la cible définissent le voisinage de traits sémantiques et le complémentaire de ce voisinage. Pour les cooccurrents d'ordre 2, il faudrait introduire des frontières au niveau des syntagmes, afin de rester cohérent avec les choix faits jusque-là. De plus, l'annotation donne naissance à deux ensembles de traits sémantiques, un ensemble interne aux frontières du paradigme défini par les cooccurrents lexicaux d'ordre 2, l'autre externe à ce paradigme[88].

Dans les expériences, la projection d'information sémique dans le corpus n'a été réalisée que pour les cooccurrents d'ordre 1. Les cooccurrents d'ordre 2 sont étudiés ultérieurement, pour la projection des unités lexicales dans le dictionnaire (*cf.* sous-section 4.2). Les développements de cette sous-section porteront donc sur la cooccurrence de traits sémantiques d'ordre 1, ou cooccurrence sémique, même s'ils peuvent être généralisés aux traits sémantiques associés aux cooccurrents d'ordre 2.

4.1.2. Un jeu de contrastes similaire à celui de l'environnement lexical

Les paramètres sur lesquels il est possible de jouer pour faire ressortir des contrastes au niveau des traits sémantiques sont les mêmes que ceux du niveau lexical : paramètre de voisinage (présence des traits sémantiques dans le même paragraphe que la cible) ; paramètre de domaine (approche au sein d'un domaine ou confrontant plusieurs domaines) ; paramètre temporel (découpage en deux ou plusieurs périodes).

Au niveau des techniques utilisées, on recourt au même type de traitements que pour les unités lexicales :

- Un calcul d'indices de significativité en contrastant l'ensemble des traits sémantiques des cooccurrents et des traits sémantiques du paradigme de substitution soit avec leurs analogues dans d'autres domaines, soit au reste du corpus domanialisé. Ces indices permettent de sélectionner des traits sémantiques saillants et de hiérarchiser les traits.
- L'utilisation d'une décomposition en périodes multiples pour affiner l'analyse : on compare les saillances aux différentes périodes pour voir la stabilité du comportement (accroissement stable, saillance ponctuelle, etc.) et pour regrouper les traits sémantiques par similarité de profils d'évolution.

Ce jeu de contrastes prépare l'analyse de l'évolution du sémème, à savoir la reconfiguration du sémème et son enrichissement.

4.1.3 Étude de la reconfiguration du sémème

Pour étudier la reconfiguration du sémème, les observables se réduisent aux traits sémantiques associés au sens codé de la cible. On observe dans quelle mesure les traits sémantiques sont surreprésentés ou sous-représentés au voisinage de la cible lexicale. Un trait surreprésenté de façon significative sera considéré comme activé dans les nouveaux emplois, il participera à la définition du nouveau sens et constituera un lien avec les définitions existantes. Un trait sous-représenté sera considéré comme inhibé dans les nouveaux emplois, il sera exclu de la définition du nouveau sens. La perte d'information sémantique usuellement associée à la cible lexicale témoignera du caractère nouveau du sens émergent. La structure

[88] Par exemple, *crédit* est cooccurrent d'ordre 1 de *toxique,* les cooccurrents d'ordre 2 provenant de *crédit* appartiennent à l'ensemble des syntagmes contenant *crédits*. /banque/ est trait sémantique de *crédit*, donc il est un cooccurrent d'ordre 1 de *toxique* sur le plan des traits sémantiques. /banque/ apparaît dans tous les syntagmes où *crédit* apparaît, mais, comme /banque/ est aussi trait sémantique de *paiement*, il apparaît dans les syntagmes contenant *paiement* et pas *crédit*. On sort ainsi des frontières du paradigme de cooccurrents d'ordre 2 induit par *crédit*.

textuelle module ainsi l'ancien sémème, elle le reconfigure sous forme d'information en lien ou, inversement, en rupture avec le nouveau sens.

a- **Toxique** – ***contraste au sein d'un domaine et entre différents domaines, puis recoupements***

L'étude de la reconfiguration du sémème de *toxique* s'appuie sur des traitements préalables similaires à ceux des domaines. Une annotation en traits sémantiques est effectuée par la plateforme Semy. Les spécificités associées à chaque trait sémantique sont calculées par le logiciel Lexico3, par le double jeu de contrastes au sein d'un domaine (au sein du corpus 'Crise financière') et entre plusieurs domaines (entre les voisinages du corpus 'Monde Diplomatique' et du corpus 'Crise financière'). On réduit les observables aux traits sémantiques correspondant aux sèmes microgénériques ou spécifiques du sens codé, à savoir l'ensemble suivant :

{/toxique/, /produit/, /minéral/, /animal/, /végétal/, /provoquer/, /intoxication/, /destruction/, /organisme/, /vivre/, /agir/, /négativement/, /individu/, /physique/, /psychisme/, /être/, /contenir/, /poison/, /toxine/, /distiller/, /ouvertement/, /propos/, /médisant/, /dépréciateur/, /empoisonnement/, /quantité/, /devenir/, /devoir/, /présence/, /produire/, /action/, /organe/, /tissu/, /origine/, /cause/, /pollution/}

Lorsque le sous-corpus des voisinages du corpus 'Crise' est contrasté avec le reste de ce corpus ou aux voisinages du *Monde Diplomatique*, la majeure partie des traits sémantiques n'émerge significativement dans aucun cas, c'est-à-dire que leur spécificité est inférieure au seuil de 2. Les traits saillants de façon positive ou négative[89] sont présentés dans le tableau III.1.30.

Corpus	Corpus 'Crise financière'		Corpus des voisinages de *toxique* des corpus 'Crise' et 'Monde Diplomatique'	
Sous-corpus	Voisinages de *toxique*		Voisinages dans le corpus 'Crise'	
	Trait sémantique	Spécificité	Trait sémantique	Spécificité
Traits sémantiques surreprésentés	/devoir (v)/	5	/devoir (v)/	11
	/agir/	3	/agir/	2
	/intoxication/	2		
Traits sémantiques sous-représentés			/devenir (v)/	-2
			/végétal (adj)/	-3
			/mot/	-3
	/individu/	-2	/individu/	-3
	/action/	-4	/minéral (adj)/	-3
	/produire/	-4	/produire/	-4
			/vivre/	-4
			/organisme/	-4
			/pollution/	-4
			/quantité/	-7

Tableau III.1.30 : Spécificités des traits sémantiques de la définition de toxique respectant un seuil de 2

Les traits surrreprésentés /devoir/ et /agir/, sont ininterprétables :

- /devoir/ est à valeur résultative dans la définition de *toxique* ("qui est dû à la présence d'un poison dans l'organisme). Ses occurrences en corpus proviennent du métalangage lexicographique (par exemple, dans *banque*, on a "Réserve de parties organiques vivantes devant servir en chirurgie") ou de définitions où sa valeur est autre que résultative (par exemple, une valeur d'obligation dans la définition de *dette*) ;

[89] On fixe un seuil arbitraire de 2 pour les spécificités positives et de -2 pour les spécificités négatives.

- /agir/ ressort essentiellement parce qu'il est trait sémantique d'*actifs*, unité lexicale fortement surreprésentée dans les voisinages lexicaux de *toxique*. Mais /agir/ provient d'une acception d'*actif* non pertinente dans les contextes d'emploi. Il n'est pas éliminé faute de procédure de désambiguïsation, à même d'exclure des acceptions relevant de domaines étrangers au contexte de crise financière.

Quant à /intoxication/, sa surreprésentation au demeurant peu marquée est due à une unique occurrence dans le sous-corpus, pour 16 au total dans le corpus 'Crise'. En imposant un seuil de fréquence minimale, /intoxication/ est éliminé.

Parmi les traits sémantiques sous-représentés, trois ensembles méritent attention :

- Un certain nombre d'éléments rattachés au vivant, /végétal/, /minéral/, /vivre/ et /organisme/, apparaissent en déficit dans les nouveaux emplois par rapport aux anciens emplois que représentent les voisinages du *Monde Diplomatique*.
- Il en est de même pour /pollution/ et /mot/, deux représentants de sens figurés de la définition de *toxique* et à propos desquels l'analyse des domaines ne donnait pas d'information. Ces sens figurés ne caractériseraient donc pas le sens de *toxique* dans le corpus 'Crise'.
- Deux candidats, /quantité/ et /individu/ semblent témoigner d'un autre type de changement, pour lequel on propose l'interprétation suivante : *toxique* s'implante dans un univers virtuel, où la victime n'est plus un individu mais le système économique dans son ensemble, et où la substance et sa nature particulaire quantifiable en chimie disparaissent : le toxique des produits et instruments financiers perd son caractère mesurable.

In fine, l'ensemble des résultats obtenus invite à voir non une restructuration alternant saillances et effacements, mais une inhibition massive du sémème de *toxique* dans ses nouveaux emplois, tout comme pour les domaines. Cette reconfiguration n'est pas tout à fait conforme à une analyse manuelle des occurrences de *toxique* : la connotation négative de *toxique* et l'idée de dégât sont présentes en corpus. Au niveau du sémème, ces facettes sémantiques sont exprimées par les traits sémantiques /négativement/ et /destruction/ ; pourtant, les spécificités ne montrent pas que ces traits sémantiques émergent significativement.

b- Outreau – *modulation dans le temps*

Dans l'analyse d'*Outreau*, on cherche à observer la modulation du sémème sur plusieurs périodes de temps. Pour cela, on s'appuie sur l'entrée lexicographique virtuelle, proposée par M. Lecolle à partir de ses connaissances du corpus.

Cette entrée comporte deux définitions : une correspondant à l'ancien sens, l'autre au nouveau sens. On sort quelque peu du cadre fixé, puisque le nouveau sens est intégré au sens codé. Cependant, ce cas de figure n'est pas totalement étranger à notre démarche. L'idée est qu'on dispose des unités de sens récupérées à partir d'un contraste de deux périodes, ou encore récupérées d'un contraste sur l'ensemble des voisinages par rapport au reste du corpus, toutes périodes confondues. On cherche à observer la modulation dans le temps des traits sémantiques, afin d'identifier ce qui relève de l'ancien sens, d'un nouveau sens intermédiaire et temporaire et d'un nouveau sens durable.

Une fois transformées en un sac de traits sémantiques, les définitions donnent le sémème suivant :

> { /ville/, /français/, /pas-de-calais/, /erreur/, /judiciaire/, /découverte/, /croyance/, /existence/, /réseau/, /pédophile/, /réfutation/, /publique/ }

La méthodologie est similaire à celle présentée pour *toxique :* on calcule les spécificités associées à tout trait sémantique du sens codé dans les voisinages (paragraphes contenant *Outreau*) par rapport à la totalité du corpus. Les spécificités sont calculées pour chaque période.

Nous avons retenu trois axes d'observation :

- l'étude de la pertinence des traits sémantiques une fois la définition déstructurée ;
- l'analyse de l'activation par période de chaque trait. Celle-ci est établie d'une part à partir des résultats numériques, d'autre part manuellement, sans connaissance préalable des résultats numériques. Les listes établies manuellement et automatiquement sont ensuite confrontées ;
- la validation de l'allure, sous forme d'histogrammes, de l'évolution observée sur les cinq périodes à trait sémantique fixé.

Après déstructuration de la définition sous forme de sac de traits sémantiques, la pertinence de certains traits est remise en question. L'analyse se heurte au caractère prédicatif de certains traits, à savoir /découverte/, /existence/, /croyance/ et /réfutation/ : si ces traits sémantiques sont traités de façon isolée, leur analyse est ambiguë et délicate, voire impossible.

Pour évaluer l'activation des traits sémantiques, des listes de données qualitatives sont constituées manuellement, où les traits sont classés avec les valeurs "activé", "non-activé" ou "indécidable" pour chaque période, puis confrontées aux listes correspondantes de spécificités calculées automatiquement. Afin de mettre en parallèle les résultats obtenus manuellement et automatiquement, on considère que les valeurs de spécificités négatives ou faibles (inférieures à 2), correspondent à une non-activation du trait, et les spécificités supérieures à 2, à son activation.

	période				
Trait sémantique	**1**	**2**	**3**	**4**	**5**
ville	oui	non	non	non	non
français (adj)	oui	non	non	non	non
pas-de-calais	oui	oui	oui	oui	non
erreur	non	non	-	-	oui
judiciaire	non	oui	oui	oui	oui
découverte	oui	-	-	-	oui
croyance	-	-	-	-	oui
réfutation	non	-	-	-	oui
existence	oui	oui	oui	oui	oui
réseau	oui	oui	oui	oui	oui
pédophile	oui	oui	oui	oui	oui

Légende :
"oui" : activation du trait sémantique
"non" : non-activation du trait sémantique
"-" : trait inclassable (cas ambigus)
case grise : convergence avec les spécificités

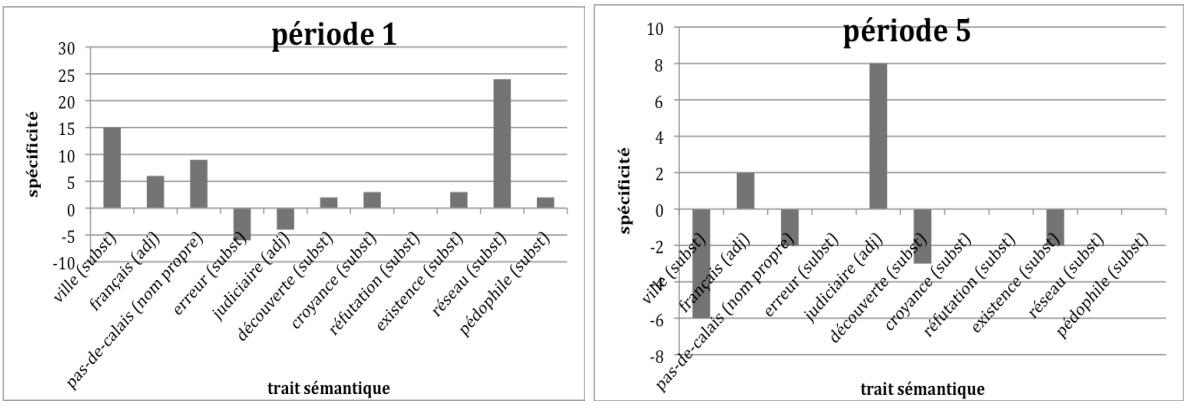
	période				
Trait sémantique	**1**	**2**	**3**	**4**	**5**
ville	15	8	0	-7	-6
français (adj)	6	0	0	0	2
pas-de-calais	9	50	4	0	-2
erreur	-6	0	0	0	0
judiciaire	-4	15	-2	23	8
découverte	2	-4	0	0	-3
croyance	3	0	0	0	0
réfutation	0	-2	0	0	0
existence	3	-3	-2	0	-2
réseau	24	2	0	-5	0
pédophile	2	8	0	0	0

Légende :
spécificité ≥ 2 : activation du trait sémantique
spéc. ≤ -2 : non-activation du trait sémantique
0 : non-activation (valeur faible)
case grise : convergence avec l'analyse manuelle

Figure III.1.31 : Activation des traits sémantiques du sémème d'Outreau. Confrontation de l'analyse manuelle et du calcul de spécificités.

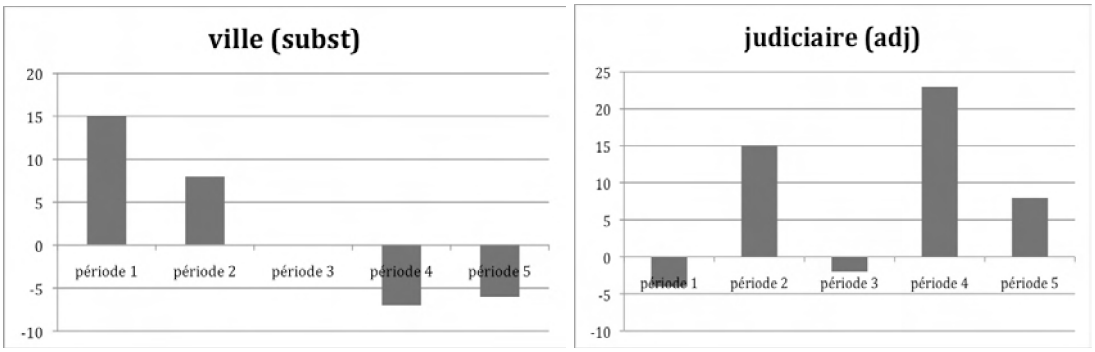
Hors cas ambigus mentionnés précédemment, on constate une convergence parfaite entre évaluation humaine et résultats numériques en période 1 et sur l'essentiel de la période 2. En revanche, le taux de convergence est médiocre aux périodes 3 à 5. On remarque cependant que, dans les cas tranchés, c'est-à-dire sur les spécificités les plus fortes en valeur absolue, les listes manuelle et automatique s'accordent généralement. Ces résultats sont en partie dus à la taille des sous-corpus des périodes 3 à 5, plus réduites. De ce fait, la gamme de spécificité est

plus réduite et le seuil de spécificité est rarement franchi. Les histogrammes ci-dessous détaillent les résultats obtenus en période 1 et 5.



Figures 6.32 a et b : Spécificités du sémème de toxique en périodes 1 et 5.

Enfin, les histogrammes d'évolution par période des traits sémantiques ont été globalement jugés cohérents avec l'analyse manuelle, à l'exception de /pédophile/, dont l'évolution, non couplée à celle de /réseau/, est en désaccord avec la connaissance du corpus, et hors cas ambigus. Les exemples de /ville/ et /judiciaire/ sont illustrés ci-dessous par des graphiques d'évolution des spécificités.



Figures 6.33.a et b : Évolution des spécificités de /ville/ et /judiciaire/ au cours du temps.

La reconfiguration progressive du sémème au cours du temps, identifiée manuellement, est donc également sensible à travers les résultats numériques de la procédure. Conformément à l'analyse manuelle, les résultats numériques font apparaître que les traits sémantiques qui renvoient au lieu s'effacent progressivement. Le caractère judiciaire s'impose (trait sémantique /judiciaire/). Par contre, si /erreur/ est bien sous-représenté en première période, il n'évolue pas de façon suffisamment marquée vers une saillance positive.

Les résultats restent imparfaits, notamment à cause du choix de représentation du sémème (déstructuration en unités qui, isolément, sont ambiguës) et de tailles de sous-corpus (l'échelle de spécificités dépend de la taille du sous-corpus ; pour un petit sous-corpus, les valeurs de spécificité tendront à être plus faibles, et les saillances n'émergeront pas).

4.1.4. Étude de l'enrichissement du sémème

Pour enrichir le sémème, le critère de sélection de candidats à l'enrichissement est celui de la saillance maximale dans les emplois textuels : on recherche les traits sémantiques qui ressortent le plus significativement dans le voisinage de la cible lexicale.

Comme pour les unités lexicales, les traits sémantiques vraiment nouveaux, c'est-à-dire différents des traits sémantiques du sens codé, sont susceptibles d'apparaître lors d'une comparaison confrontant différents domaines : on fait ainsi ressortir des traits sémantiques

spécifiques au nouveau domaine d'emploi, par contraste avec les domaines d'emplois usuels. Ce type de contraste permet de dégager des sèmes spécifiques, en cela qu'ils se distinguent des autres acceptions. Cependant, les contrastes internes à un domaine, obtenus en opposant le voisinage de la cible au reste du corpus, ont aussi leur intérêt pour l'enrichissement. Ils permettent de distinguer une thématique omniprésente dans le corpus, pas particulièrement caractéristique du voisinage de la cible lexicale, d'une thématique qui tend à se renforcer au voisinage de la cible. Ils permettent également de faire ressortir des facettes sémantiques proches des sens existants. Ces facettes viendraient renforcer d'anciennes traits sémantiques du sens codé déjà identifiés comme activés lors de la reconfiguration du sémème, ou encore elles permettraient d'exhiber un contenu sémantique activé mais jusque-là non identifié comme tel (par exemple, des traits sémantiques représentatifs du sémème peuvent ne pas ressortir faute d'analogues formels associés à d'autres unités lexicales ; par exemple, /négativement/ définit *toxique* mais n'est présent que dans 9 définitions du dictionnaire, alors que le trait /négatif/, sémantiquement proche mais formellement distinct, est présent dans 186 définitions).

L'annotation en traits sémantiques peut améliorer l'apport du plan lexical de deux façons :

- ***Renforcement sémique :*** elle fait émerger des informations déjà présentes et relativement explicites sur le plan lexical. Ces informations sont reprises à même niveau, voire renforcées. Comme on le détaillera, ce que nous appelons renforcement est, d'une certaine façon l'expression d'isotopies locales : cela correspond à une reprise d'un trait sémantique associé à une unité lexicale saillante par d'autres unités lexicales, autrement dit, à une récurrence de trait sémantique sur une unité lexicale saillante et également sur d'autres unités.
- ***Innovation sémique :*** elle fait émerger du contenu sémantique déjà sensible au niveau lexical, mais de façon sous-jacente. Autrement dit, l'annotation permet d'expliciter des idées exprimées de façon diffuse par les unités lexicales.

Les trois expériences que nous présentons ont pour objectif de valider l'existence d'un apport de l'annotation (paragraphe a), d'exploiter le renforcement sémique (paragraphe b), puis de faire émerger l'innovation sémique (paragraphe c).

a- Double apport de l'annotation en traits sémantiques : une validation à forte composante manuelle

La double contribution de l'annotation en traits sémantiques (renforcement et innovation sémiques) est mise en évidence à travers une étude sur le sens *d'économie réelle* dans le corpus 'Crise financière'. Cette étude est sous-tendue par une analyse manuelle importante, destinée à valider l'annotation en traits sémantiques et à identifier ses apports et limites.

Deux versions du corpus sont établies, une version lexicale et une version annotée en traits sémantiques. Les spécificités sont calculées sur l'ensemble des paragraphes contenant *économie réelle* par rapport à l'ensemble du corpus 'Crise financière'. Les unités lexicales (respectivement traits sémantiques) saillantes sont conservées. Elles sont sélectionnées en fonction d'un seuil de spécificité minimale de 2 et d'un seuil de fréquence minimale de 10.

Deux approches sont ensuite mises en place :

- Une approche où sont observées les unités les plus spécifiques (unités lexicales et traits sémantiques), qui favorise une saillance marquée avec le choix d'un seuil de spécificité élevé. Les unités sont étudiées de façon isolée.
- Une approche où plus d'unités sont sélectionnées (fixation d'un seuil de spécificité bas), mais où ces unités sont regroupées en classes sémantiques.

a1) Observation des unités les plus spécifiques

Considérons les listes de formes et de traits sémantiques les plus spécifiques du voisinage *d'économie réelle,* disponibles en figures (6.34.a) et (6.34.b).

Rang	Forme	Spéc.	Rang	Forme	Spéc.	Rang	Forme	Spéc.
1	l	20	6	conséquences	7	11	effets	6
2	financière	15	7	financier	7	12	revenus	6
3	impact	11	8	richesses	7	13	dite	6
4	crise	9	9	profits	7	14	contagion	6
5	récession	9	10	salaires	6			

Tableau III.1.34.a : Formes lexicales les plus spécifiques du voisinage d'économie réelle

Rg	Forme	Spé	Rg	Forme	Spé	Rg	Forme	Spé
1	/diminution (subst)/	11	16	/budget (subst)/	21	31	/surproduction (subst)/	9
2	/roi (subst)/	11	17	/ressource (subst)/	16	32	/sous-production (subst)/	9
3	/dysfonctionnement (subst)/	10	18	/particulier (subst)/	16	33	/D=dramaturgie/	9
4	/contagion (subst)/	10	19	/régir (v)/	15	34	/rupture (subst)/	9
5	/profond (subst)/	10	20	/argent (subst)/	14	35	/craindre (v)/	9
6	/capitaliste (adj)/	10	21	/particulier (adv)/	14	36	/boursier (adj)/	9
7	/subit (adj)/	10	22	/répercussion (subst)/	14	37	/noeud (subst)/	9
8	/appréciable (adj)/	10	23	/théâtre (subst)/	13	38	/enthousiasme (subst)/	9
9	/époux (subst)/	10	24	/bien (subst)/	12	39	/effondrement (subst)/	9
10	/décisif (adj)/	10	25	/chômage (subst)/	12	40	/pathologique (adj)/	9
11	/économie (subst)/	10	26	/ralentissement (subst)/	11	41	/économique (adj)/	9
12	/collision (subst)/	10	27	/déterminant (adj)/	11	42	/retentissement (subst)/	9
13	/financier (adj)/	9	28	/néfaste (adj)/	11	43	/galaxie (subst)/	9
14	/intense (adj)/	9	29	/en/	11	44	/développement (subst)/	9
15	/progressif (adj)/	9	30	/phénomène (subst)/	9			

Tableau III.1.34.b : Traits sémantiques les plus spécifiques du voisinage d'économie réelle

Ces listes font ressortir une **dimension économique et financière**, par exemple à travers des traits sémantiques tels que /budget/, /argent/, /capitaliste/, /économie/ et des unités lexicales *financière, financier, profit.* De même, la **sphère réelle** apparaît à travers les unités les plus spécifiques, surtout sur le plan des traits sémantiques, à travers des traits sémantiques comme /chômage/, /bien/, /ressource/, /surproduction/. La notion de **choc** est également présente (/collision/, /répercussion/, /effondrement/ sur le plan des traits sémantiques ; *impact* sur le plan lexical), de même que celle de propagation ou même de maladie, avec l'unité lexicale *contagion* t e les traits sémantiques /contagion/, /dysfonctionnement/ et /pathologique/. Les idées sensibles à la lecture se retrouvent ainsi au niveau des unités les plus spécifiques, sur le plan lexical et de façon encore plus marquée sur le plan des traits sémantiques.

Cependant, le nombre d'unités lexicales ou de traits sémantiques associés à une idée donnée reste relativement limité, du fait de la taille volontairement réduite de la liste d'unités les plus spécifiques, d'où la mise en place de la seconde approche.

a2) Regroupement en catégories

Les catégories sont établies à partir d'idées dégagées de la lecture et de l'observation des unités les plus spécifiques. Les formes lexicales et les traits sémantiques sont obtenus par application d'un seuil bas (spécificité positive, supérieure à 2), puis affectés manuellement aux catégories considérées comme pertinentes, par parcours exhaustif des listes d'unités

spécifiques. Enfin, les formes lexicales et les traits sémantiques d'une même catégorie sont confrontés.

Les principales catégories choisies manuellement correspondent aux idées suivantes : la maladie ; le cataclysme ; le choc ou la brutalité ; la réalité ou, par opposition, la virtualité ; l'économie dans sa dimension matérielle. Les classes définies ont un degré de généralité variable. De plus, elles ne forment pas une partition : elles se superposent parfois et ne couvrent pas toutes les facettes sémantiques présentes dans l' "*économie réelle*". Certains traits sémantiques sont donc affectés à plusieurs classes, tandis que d'autres ne rejoignent pas de classe particulière.

À titre d'exemple, considérons les catégories suivantes : la catégorie 'maladie' (tableau III.1.35.a) et la catégorie 'choc, brutalité' (tableau III.1.35.b).

Traits sémantiques de la catégorie 'maladie'						**Formes de la catégorie 'maladie'**	
Trait	**Spé**	**Trait**	**Spé**	**Trait**	**Spé**	**Forme**	**Spé**
néfaste (adj)	11	mal (subst.)	7	saignée (subst)	4	crise	9
contagion (subst.)	10	physiologique (adj)	6	affecter (v)	3	contagion	6
dysfonctionnement (subst)	10	épidémie (subst.)	5	bistouri (subst)	3	affectée	4
pathologique (adj)	9	maladie (subst)	5	défaillir (v)	3	injectés	3
dépression (subst)	8	infection (subst)	5	nuisible (adj)	3	affecter	3
trouble (subst.)	8	contagieux (adj)	4	psychose (subst)	3	aggravée	3
crise (subst.)	7	remédier (v)	4	soigner (v)	3		

Tableau III.1.35.a : Traits sémantiques saillants associés à la catégorie 'maladie'

Traits sémantiques de la catégorie 'choc, brutalité'				**Formes de la catégorie 'choc, brutalité'**	
Trait	**Spé**	**Trait**	**Spé**	**Forme**	**Spé**
effondrement (subst.)	9	débris (subst)	4	choc	4
tarissement (subst.)	8	déferler (v)	3	dommages	3
décrue (subst.)	8	secousse (subst)	3	éclate	3
dépression (subst.)	8	cataclysme (subst.)	3	onde	3
choc (subst..)	7	tempête (subst)	3	fumée	3
violemment (adv)	5	inondation (subst)	3	tempête	3
désagréger (v)	4				

Tableau III.1.35.b : Traits sémantiques saillants associés à la catégorie 'choc, brutalité'

Les classes possèdent des représentants aussi bien lexicaux que sémiques, les mêmes idées tendent donc à s'exprimer sur les deux plans. Cependant, le nombre d'unités par catégorie est plus important sur le plan des traits sémantiques que sur le plan lexical. De plus, certaines idées sensibles à la lecture mais sous-jacentes au niveau des formes lexicales apparaissent explicitement au niveau des traits. Par exemple, la maladie prend un caractère beaucoup plus prégnant et tangible avec des traits tels que /pathologique/, /trouble/, /infection/, /épidémie/ ou encore /maladie/. De même, l'ébranlement et la violence liés à la crise, que seul *impact* reflète assez explicitement sur le plan lexical, s'imposent avec force sur le plan sémique, avec des traits tels que /effondrement/, /heurt/, /brusque/, /violemment/ ou encore /secousse/. De façon générale, les catégories sont plus riches sur le plan sémique sur le plan lexical, parce qu'elles contiennent plus de traits sémantiques que de formes lexicales mais aussi, et surtout, parce que des idées perçues à la lecture sont exprimées clairement par les représentants sémiques alors qu'elles sont seulement sous-jacentes à travers les représentants lexicaux.

L'étude montre ainsi une convergence sur le sens *d'économie réelle* entre la lecture, le plan lexical et le plan sémique. Elle montre également un enrichissement apporté à une approche lexicale par l'annotation en traits sémantiques : des idées latentes sont explicitées par les traits.

*b- Renforcement sémique : caractérisation du sens d'***Outreau** *en intégrant la progression diachronique*

La procédure présentée ici a pour objectif de retourner des candidats à l'enrichissement en exploitant les convergences entre plan lexical et plan sémique. Autrement dit, notre démarche est centrée sur le renforcement sémique.

L'étude porte sur l'exemple *d'Outreau*, elle s'appuie sur un recoupement des spécificités obtenues sur le plan lexical et sur le plan sémique. Ces recoupements sont effectués sur chacune des cinq périodes caractéristiques de l'évolution de sens d'*Outreau,* la progression diachronique est donc intégrée à cette étude afin de faire émerger la stratification de sens.

Le calcul de spécificités appliqué au corpus non annoté et au corpus annoté permet d'obtenir deux listes en sortie : une liste de formes lexicales et une liste de traits sémantiques. Chaque unité est affectée d'un coefficient, sa spécificité.

Les traits sémantiques témoignant d'un renforcement sémique sont obtenus en confrontant les deux listes, restreintes aux items de spécificité positive supérieure à 2 : seuls sont conservés les traits possédant un analogue sur le plan lexical. Plus précisément, pour un trait sémantique donné, nous choisissons comme référence sur le plan lexical les formes morphologiquement proches de son signifiant, c'est-à-dire les formes auxquelles il est associé de façon immédiate, mais pour lesquelles il apparaît comme le plus générique (par exemple, /magistrat/ pour *magistrats* ou *magistrature*).

Le trait sémantique sera considéré comme pertinent et retenu s'il respecte deux critères :

- Un critère de coactualisation qui implique :
    - l'existence d'une forme de référence parmi les unités de spécificité positive : un des représentants lexicaux "immédiats" du candidat, c'est-à-dire proche morphologiquement, est de spécificité positive dans le sous-corpus considéré (seuil de spécificité fixé à 2) ;
    - l'existence du trait sémantique lui-même comme unité spécifique positivement : le trait est lui aussi surreprésenté dans le sous-corpus considéré (seuil de spécificité fixé à 5 ou 3 selon l'approche).
- Un critère de renforcement : le trait sémantique a une spécificité strictement supérieure à celle de la forme de référence de plus grande spécificité.

Le premier critère revient à sélectionner les traits qui se "manifestent" de façon significative en tant que formes.

Le critère de renforcement exploite le nombre d'occurrences des unités lexicales à l'origine d'un trait sémantique dans la procédure d'annotation. Il amène à sélectionner des traits qui proviennent d'autres formes que la forme de référence la plus spécifique, morphologiquement liées ou non à celle-ci.

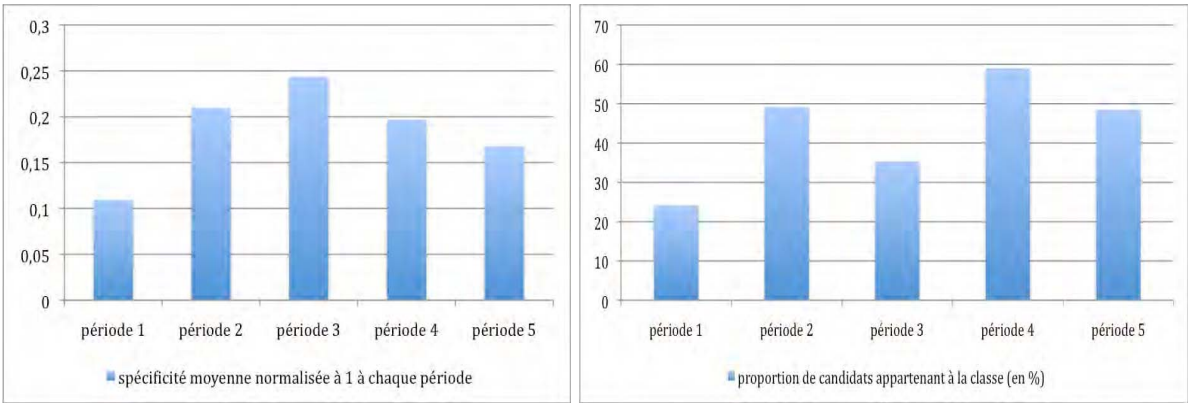
Pour observer l'évolution diachronique d'*Outreau*, on génère pour chaque période un ensemble de traits sémantiques témoins d'un renforcement sémique. Ils sont obtenus à partir des spécificités calculées pour les sous-corpus des cinq périodes. On dispose donc de cinq listes de traits, une par période, dont les unités constitutives varient d'une période à l'autre.

À partir des listes de traits par période, des classes sont définies manuellement. Ces classes correspondent à des regroupements sémantiques réalisés à des fins d'observation sur la base d'intuitions sémantiques. Par exemple, en période 1, émerge une classe JUDICIAIRE qui comporte notamment les traits /police/, /procureur/, /écrouer/. Ces classes forment une partition sur les traits retenus pour une période donnée et ne sont pas nécessairement

communes à toutes les périodes. Pour chaque période, l'importance des classes est représentée par deux indicateurs :

- (i) l'un prend en compte la taille de la classe, par calcul de la proportion de traits de la liste appartenant à la classe considérée ;
- (ii) l'autre est destiné à refléter la significativité des traits sémantiques affectés à la classe, par calcul de la moyenne des spécificités des traits de la classe, puis, pour permettre une comparaison entre différentes périodes, homogénéisation des tailles des vecteurs-périodes des moyennes de spécificités. Les conclusions avancées s'appuient sur des tendances communes aux deux indicateurs.

Le JUDICIAIRE est présent à chaque période (*cf.* figure III.1.36). La proportion de traits sémantiques de même que la spécificité moyenne sont plus faibles en période 1, période à laquelle le procès n'est pas encore entamé.



Figures 6.36.a et 6.36.b : Évolution par période de la classe JUDICIAIRE d'après (a) la spécificité moyenne des traits sémantiques de la classe et (b) la proportion de traits sémantiques de la période appartenant à la classe

Au niveau qualitatif, les traits sémantiques de cette classe renvoient aux notions émergeant à chaque période. Par exemple, en période 1, l'ensemble de traits {/écrouer/, /police/, /arrestation/, /incarcération/, /incarcérer/, /prévenu/}, soit près de la moitié des éléments associés à la sphère judiciaire, renvoient à l'idée d'arrestation.

Le POLITIQUE apparaît à la période 4 et se renforce à la période 5, ce qui est en accord avec la mise en place d'une commission d'enquête parlementaire et la volonté politique de se pencher sur les dysfonctionnements du système judiciaire.

Inversement, la classe du CRIME (évoquée par des candidats tels que /pédophilie/, /meurtre/, /viol/, /délinquant/), incluse dans la dimension judiciaire et policière, est présente en périodes 1 à 3, puis cette présence chute en période 4 et enfin disparaît en période 5 (*cf.* figure III.1.37*)*. Cette évolution s'explique doublement : d'une part, par l'ampleur accordée à l'émotionnel populaire dans un premier temps, puis le recul de son influence avec la prise de conscience de l'erreur judiciaire; d'autre part, par le glissement du scandale de l'affaire de pédophilie vers l'erreur.

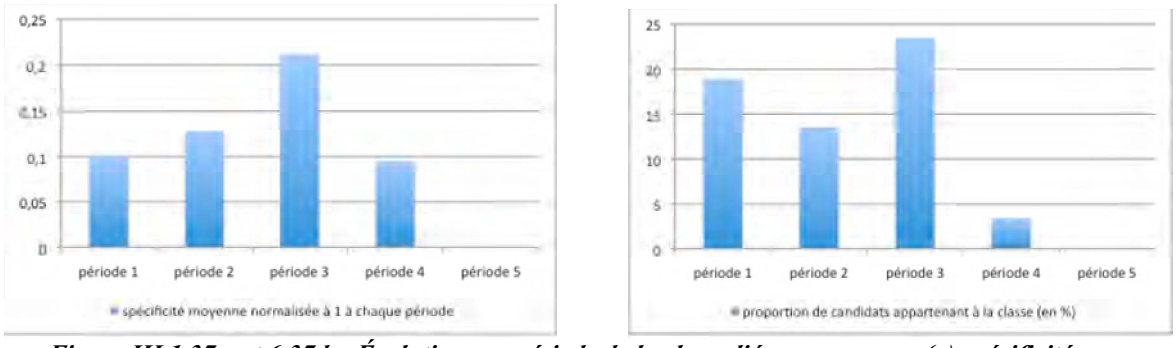


Figure III.1.37.a et 6.37.b : Évolution par période de la classe liée au CRIME en (a) spécificité moyenne et (b) proportion de traits sémantiques affectés à la classe

Les classes rattachées au LIEU (LIEU D'HABITATION, LIEU GEOGRAPHIQUE) ne sont représentées significativement qu'à la période 1, ce qui rejoint l'évolution du sens locatif d'*Outreau*, initialement dominant voire exclusif, puis supplanté par d'autres sens. Inversement, la classe du FIASCO (constituée de /naufrage/, /drame/, /faillite/, /faute/) s'impose comme représentative de la période 5.

En résumé, les résultats expérimentaux sont en accord avec les analyses manuelles : ils font émerger des tendances saillantes qui mettent en évidence une évolution diachronique conforme à l'étude manuelle. Les traits sémantiques obtenus permettent de nuancer les classes saillantes. Néanmoins, ces résultats concluants sont tributaires d'une identification manuelle de classes, extraites à la fois de traitements de résultats issus de la procédure semi-automatique et d'un regard orienté par l'étude linguistique. Cette émergence de classes, par exemple celle correspondant à la dimension judiciaire et policière, est le résultat d'un processus d'interprétation complexe et que l'on ne sait, à ce jour, pas encore décrire précisément. Cependant, la constitution de classes pourrait s'appuyer sur la macrostructure du dictionnaire (*cf.* chapitre III.1, 3.2) et des techniques telles que celles évoquées au (chapitre II.2, 4).

c- Innovation sémique : filtrage par dépendance à un domaine commun

L'innovation sémique est étudiée à travers l'exemple de *toxique*. Le principe qui sous-tend la procédure mise en place est de même nature que celui qui guidait l'analyse d'*économie réelle* : les traits sémantiques contribuant à l'innovation sémique sont regroupés dans une même classe sémantique, dotée d'un dénominateur commun.

Pour *toxique*, le dénominateur commun est le domaine émergent, à savoir FINANCES. La recherche de candidats à l'enrichissement du sens de *toxique* s'appuie sur l'hypothèse qu'ils forment une isotopie locale et qu'ils dépendent du nouveau domaine. Ces hypothèses se traduisent par un critère de spécificité élevée et par un filtrage en fonction du domaine considéré.

Les candidats retenus (tableau III.1.38) vérifient les points suivants :

- leur spécificité est supérieure à un seuil, fixé arbitrairement à +6, dans chacune des confrontations (interne au corpus crise et relativement au *Monde Diplomatique*) ;
- ils apparaissent dans au moins une définition du dictionnaire dont l'étiquette de domaine est FINANCES.

Pour obtenir une hiérarchie globale des candidats qui prenne en compte leur hiérarchie dans chaque confrontation (au sein d'un domaine et entre plusieurs domaines), un indicateur

synthétique est calculé. Cet indicateur est une moyenne pondérée entre les deux valeurs de spécificité obtenues pour chaque trait sémantique[90].

	Corpus Crise	Corpus des voisinages	Synthèse		Corpus Crise	Corpus des voisinages	Synthèse
Trait sémantique	**Spéc.**	**Spéc.**	**Ind.**	**Trait sémantique**	**Spéc.**	**Spéc.**	**Ind.**
/banque/	21	50	30,2	/payer/	7	26	13,0
/bilan/	27	22	25,4	/négociable/	9	21	12,8
/prêt (subst.)/	16	43	24,6	/budget/	6	27	12,7
/banquier/	13	37	20,6	/bénéfice/	10	15	11,5
/actif (adj)/	20	20	20	/pur/	10	15	11,5
/financier (adj)/	13	29	18,1	/dépenser/	10	14	11,2
/tirelire/	12	31	18,0	/dépôt/	8	18	11,1
/servir/	12	30	17,7	/équipe/	8	17	10,8
/escroc/	11	32	17,7	/rue/	7	18	10,5
/productif/	19	14	17,4	/dette/	10	11	10,3
/boniment/	11	31	17,3	/ouvrier (subst)/	7	17	10,1
/quinzaine/	11	31	17,3	/réussir/	7	17	10,1
/voleur/	11	31	17,3	/commerce/	7	16	9,8
/escroquerie/	11	30	17,0	/priver/	8	13	9,5
/succursale/	10	29	16,0	/réserve/	7	15	9,5
/réclamer/	17	11	15,0	/installer/	9	9	9
/semaine/	6	32	14,3	/crédit/	6	14	8,5
/D=finances/	8	27	14,0	/opération/	8	8	8
/argent/	6	30	13,6	/public (subst)/	6	11	7,5
/verser/	10	21	13,5				

Tableau III.1.38 : Liste ordonnée de traits sémantiques surreprésentés au voisinage de toxique dans le sous-corpus 'Crise financière' relativement au corpus 'Crise' et au corpus des voisinages

Certes, certaines unités comme /tirelire/, /rue/, /servir/ ou encore /installer/ manquent de pertinence ou sont ininterprétables. Cependant, les premiers traits sémantiques de la liste, /banque/, /bilan/, /prêt/, /banquier/ ou /financier/, s'accordent avec ce qui se dégage à la lecture des cotextes. De même, plus loin dans la liste, émerge la thématique de l'illiquidité (/réclamer/, /dette/, /payer/, /priver/, /réserve/) et la nocivité transparaît à travers des éléments comme /escroc/, /boniment/, /voleur/ ou /dette/. Une interprétation, que le lecteur est libre de juger trop audacieuse, amène à voir en filigrane la nouvelle nature de ce qui est *toxique*, qui vient remplacer la substance (/argent/, /opération/) et les nouvelles victimes qui se substituent aux êtres vivants (/public/, /succursale/, /ouvrier/).

Cette expérience guide vers un enrichissement qui, contrairement à la démarche de renforcement sémique, ne joue pas sur une forte dépendance au plan lexical. L'enrichissement repose sur l'existence d'une étiquette de domaine commune. Les traits sémantiques obtenus permettent de préciser les nuances sémantiques au sein du nouveau domaine. L'ensemble des

[90] Chaque coefficient est rapporté à la moyenne des spécificités calculées pour un corpus de confrontation donné, puis, pour chaque trait sémantique, la moyenne entre les deux nouveaux coefficients est calculée et multipliée par un facteur d'ajustement. Plus précisément, soient m_1 la moyenne de spécificités dans le corpus 'Crise', m_2 celle du corpus des voisinages. Soit un trait sémantique de spécificité $spec_1$ dans le corpus 'Crise' et $spec_2$ dans le corpus des voisinages. L'indicateur (Ind) associé au trait sémantique se calcule comme suit : $$Ind = \frac{\frac{spec_1}{m_1} + \frac{spec_2}{m_2}}{\frac{1}{m_1} + \frac{1}{m_2}}$$

traits proposés reste incomplet. La recherche de traits complémentaires pourrait être élargie à des traits sémantiques qui ne relèvent pas nécessairement de la finance, soit issus de la reconfiguration du sémème, soit associés à d'autres dénominateurs communs qui ne soient pas nécessairement des domaines, comme l'idée de destruction.

4.2 Projection du lexical dans la ressource lexicographique

4.2.1 Articulation du lexical à l'infra-lexical à partir de la ressource lexicographique

Plutôt que de projeter l'information sémique dans le corpus et d'utiliser la structure du corpus pour moduler l'information sémique, on peut projeter l'information lexicale dans la ressource lexicographique, autrement dit dans le réseau de traits sémantiques, puis se servir de la structure du dictionnaire soit pour restructurer l'information lexicale (les traits sémantiques sont des clés de structuration implicites), soit pour expliciter des liens entre unités lexicales saillantes.

Suite à l'analyse des unités lexicales en corpus, on dispose d'un ensemble d'unités lexicales associées à un ou des domaine(s), affectées d'une pondération, éventuellement hiérarchisées ou regroupées sur un critère distributionnel.

Pour les unités lexicales pondérées mais non regroupées en classes : on peut chercher (1) à exhiber les traits sémantiques les plus significatifs ou (2) à utiliser le réseau de traits pour effectuer des regroupements.

Dans le premier cas, la sélection de traits sémantiques significatifs dépend de leur distribution dans les définitions associées aux unités lexicales retenues. Le degré de saillance dépend de la présence des traits sémantiques (des traits récurrents seront considérés comme isotopiques, donc pertinents), mais aussi d'un principe lié à la loi de Zipf : un sème présent dans un grand nombre de définitions lexicographiques est peu informatif. Pour trouver un équilibre entre isotopies (récurrence informative) et traits sémantiques omniprésents (récurrence non informative), il faut disposer de coefficients d'ajustement, qui reflètent la présence d'un trait sémantique dans la totalité de la ressource lexicographique ou dans un espace considéré comme représentatif.

Dans le second cas, les traits sémantiques peuvent être vus comme des clés de structuration des unités lexicales. Ils sont considérés dans leur ensemble. On peut considérer qu'ils définissent des structures multidimensionnelles ou encore des structures de graphes. La structuration des unités lexicales repose sur l'application de techniques de clusterisation sur le réseau construit à partir des traits sémantiques partagés.

Pour les regroupements obtenus en fonction de la distribution des unités lexicales en corpus : les traits sémantiques permettent de nommer les regroupements obtenus, ou encore d'exhiber un dénominateur commun sur le plan sémantique. Un trait sera considéré comme pertinent s'il est partagé par les différents éléments du groupe.

4.2.2 Structuration des cooccurrents d'ordre 2 d'un ensemble de cibles

a- Cibles et cooccurrents d'ordre 2 issus du corpus

La sélection des cooccurrents d'ordre 2 présentée à la sous-section 3.2 a permis de constituer les listes présentées dans le tableau ci-dessous pour les cibles *actif, bouclier, pourri, tempête, toxique* et *tsunami*.

Cible	Cooc. d'ordre 1	Nb cooc. d'ordre 2	Cooccurrents d'ordre 2
actif	bancaire douteux bancaire	44	actif, activité, bilan, commission, créance, crédit, crise, débâcle, défaillance, dépôt, dette, domino, établissement, faillite, filiale, financement, garantie, géant, groupe, immobilier, industrie, institution, krach, marché, métier, nationalisation, organisme, panique, paysage, perte, plan, pôle, public, pratique, prêt, régulation, sauvetage, secret, secteur, situation, souveraineté, syndicat, système, valeur
bouclier	fiscal social	57	accession, acquis, agitation, allégement, baromètre, besoin, bouclier, budget, cadeau, casse, catastrophe, combat, concurrence, conséquence, contenu, cotisation, crise, dépense, développement, droit, dumping, économie, efficacité, exonération, explosion, force|forces, harmonisation, hécatombe, information, justice, largesse, logement, lutte, mécontentement, médias, mesure, minima, niche, norme, paquet, paradis, plan, protection, question, récession, recette, relance, réseau, restructuration, riposte, ristourne, situation, stabilité, tension, traitement, union, urgence
pourri	créance crédit titre	33	accorder, adosser, américain, automobile, bancaire, bon, bonifier, budgétaire, compromettre, consentir, créditeur, détenir, douteur|douteux, émettre, facile, financier, gouvernemental, hypothécaire, immobilier, interbancaire, issu, lier, meilleur, mutuel, nécessaire, négocier, pourri, relais, sélectif, spéculatif, subprime, titrisées, toxique
tempête	économique financier	154	absurdité, accumulation, acteur, activité, affaire, agent, aide, analyse, assainissement, assurance, autorité, bourrasque, bulle, cancer, capital, capitalisme, carte, catastrophe, chaos, chapitre, choix, chroniqueur, circuit, climat, confiance, conjoncture, conseiller, contexte, coup, crash, crise, croissance, culture, cycle, débâcle, délinquant, demande, démocratie, dérégulation, dérive, désordre, développement, difficulté, disposition, district, domaine, domination, échange, édifice, élite, entreprise, environnement, équipe, espace, établissement, étude, expert, financier, force, front, globalisation, gonflement, gouvernement, guerre, impact, industrie, innovation, institut, institution, instrument, investissement, jeu, juridiction, krach, libéralisme, licencié, liquidité, logique, machine, machinerie, marasme, marché, mathématicien, mathématique, matière, modèle, modèle, monde, mondialisation, nationalisation, notation, note, nouvelle, objectif, opération, organisation, organisme, ouragan, période, perspective, philosophie, place, placement, plan, planète, pôle, politique, prévision, prix, produit, profit, puissance, question, quotidien, ralentissement, réalisme, récession, recherche, régulation, relance, rendement, rentabilité, responsabilité, réunion, revenu, risque, sauvetage, secousse, secteur, séisme, service, situation, société, solidité, sommet, soutien, spéculation, sphère, stabilité, statistique, système, technique, tempête, tension, titre, tourmente, tragédie, tsunami, turbulence, urgence, utilité, valeur, vie, zone
toxique	crédit emprunt titre	36	accorder, adosser, américain, automobile, bancaire, banque, bon, bonifier, budgétaire, consentir, créditeur, détenir, émettre, emprunteur, européen, facile, financier, gouvernemental, heure, hypothécaire, immobilier, impôt, interbancaire, issu, lier, logement, meilleur, mutuel, nécessaire, négocier, pourri, pourrir, public, relais, sélectif, spéculatif
tsunami	financier	96	accumulation, acteur, activité, aide, assainissement, assurance, autorité, bourrasque, bulle, cancer, capital, capitalisme, carte, catastrophe, chaos, chapitre, choix, circuit, coup, crash, crise, croissance, débâcle, délinquant, dérive, désordre, disposition, district, domaine, échange, édifice, élite, entreprise, environnement, établissement, financier, front, globalisation, gonflement, industrie, innovation, institut, institution, instrument, investissement, juridiction, krach, libéralisme, liquidité, machinerie, marché, mathématicien, mathématique, matière, monde, mondialisation, notation, note, objectif, opération, organisme, ouragan, place, placement, planète, pôle, produit, profit, puissance, régulation, rendement, rentabilité, responsabilité, réunion, revenu, risque, sauvetage, secteur, séisme, service, société, solidité, sommet, soutien, spéculation, sphère, stabilité, système, technique, tempête, titre, tourmente, tsunami, turbulence, urgence, valeur

Tableau III.1.39 : Cooccurrents d'ordre 1 et 2 des cibles lexicales actif, bouclier, pourri, tempête, toxique et tsunami

On cherche à faire émerger des regroupements sémantiquement cohérents dans ces listes à partir de la structure du dictionnaire. Les regroupements seront générés en fonction des traits sémantiques partagés.

b- Constitution de regroupements sémantiques à partir du dictionnaire de référence

Les unités appartenant au paradigme de cooccurrents d'ordre 2 de la cible lexicale constituent un réseau dans le dictionnaire de référence. Elles peuvent être reliées les unes aux autres en fonction du nombre de traits sémantiques que leurs définitions partagent et du degré d'information apporté par chaque trait sémantique. On cherche à dégager du réseau d'unités lexicales ainsi généré des groupes sémantiquement cohérents, c'est-à-dire comportant des unités par rapport auxquelles les traits sémantiques se distribuent de la même façon.

À l'aide de Semy, un ensemble *T(u)* de traits sémantiques est associé à toute unité lexicale *u* du paradigme de cooccurrents d'ordre 2. Par exemple :

- *T(tsunami)* = { /chute/, /côte/, /D=sciences de la terre/, /déferler/, /dégât/, /éruption/, /falaise/, /glacier/, /grand/, /grave/, /immense/, /mer/, /océanique/, /onde/, /origine/, /pan/, /provoquer/, /solitaire/, /sous-marin/, /terre/, /tremblement/, /tsunami/, /vague/, /volcanique/ }
- *T(séisme)* = { /allure/, /brusque/, /catastrophe/, /D=sciences de la terre/, /écorce/, /émotion/, /fort/, /globe/, /large/, /modification/, /nature/, /phénomène/, /plus ou moins/, /population/, /pouvoir/, /prendre/, /provoquer/, /relief/, /ressentir/, /secousse/, /séisme/, /série/, /soulever/, /surface/, /terrestre/, /toucher/, /zone/ }

Un coefficient d'affinité *c(u,t)* est associé à tout couple unité lexicale - trait sémantique *(u,t)* :

$$\begin{cases} c(u,t) = 0 \quad si \quad t \notin T(u) \\ c(u,t) = -\log\left(\dfrac{\left|T^{-1}(t)\right|}{n_U}\right) \quad si \quad t \in T(u) \end{cases}$$

, où $T^{-1}(t)$ désigne l'ensemble des unités lexicales dont *t* est trait sémantique et où n_U désigne le nombre total d'unités lexicales. Par exemple :

- *tsunami* a 96 cooccurrents d'ordre 2 (n_U=96), dont *séisme* et *risque* ;
- T^{-1}*(/phénomène/)* = {accumulation, crise, environnement, produit, séisme, sphère, stabilité, système, turbulence} ; $\left|T^{-1}(phénomène)\right| = 9$;
- *c(risque, /phénomène/)* = 0 car */phénomène/* n'est pas trait sémantique de *risque* ;
- $c(séisme, /phénomène/) = -\log\left(\dfrac{\left|T^{-1}(/phénomène/)\right|}{n_U}\right) = -\log\left(\dfrac{9}{96}\right) = 1,03$.

Ce calcul s'inspire de la méthode tf-idf et cherche à refléter le degré d'informativité du trait sémantique : des traits partagés par de nombreuses définitions, tels que /être/, /chose/ ou /objet/, auront un faible coefficient et seront considérés comme faiblement informatifs. Des traits sémantiques qui ne sont présents que dans quelques définitions auront un coefficient élevé et seront considérés comme informatifs. Chaque unité lexicale correspond alors à un vecteur de traits sémantiques, avec un coefficient d'affinité sémantique par trait.

Un coefficient d'affinité sémantique entre deux unités lexicales est obtenu par produit scalaire des vecteurs de traits sémantiques associé à chacune des deux unités. Si les deux unités partagent un grand nombre de traits sémantiques ou des traits sémantiques informatifs,

le coefficient est élevé ; si les deux unités partagent peu de traits sémantiques ou des traits sémantiques peu informatifs, le coefficient est faible.

Une unité lexicale se définit ainsi comme un vecteur d'affinités lexicales avec les autres unités lexicales. Une classification hiérarchique ascendante est effectuée sur les unités lexicales à partir des coefficients d'affinité sémantique, par moyenne pondérée des groupes associés (méthode implémentée par le logiciel PermutMatrix (Caraux et Pinloche, 2005), qualifiée de méthode *WPGMA*). La CAH obtenue pour *tsunami* est présentée ci-dessous.

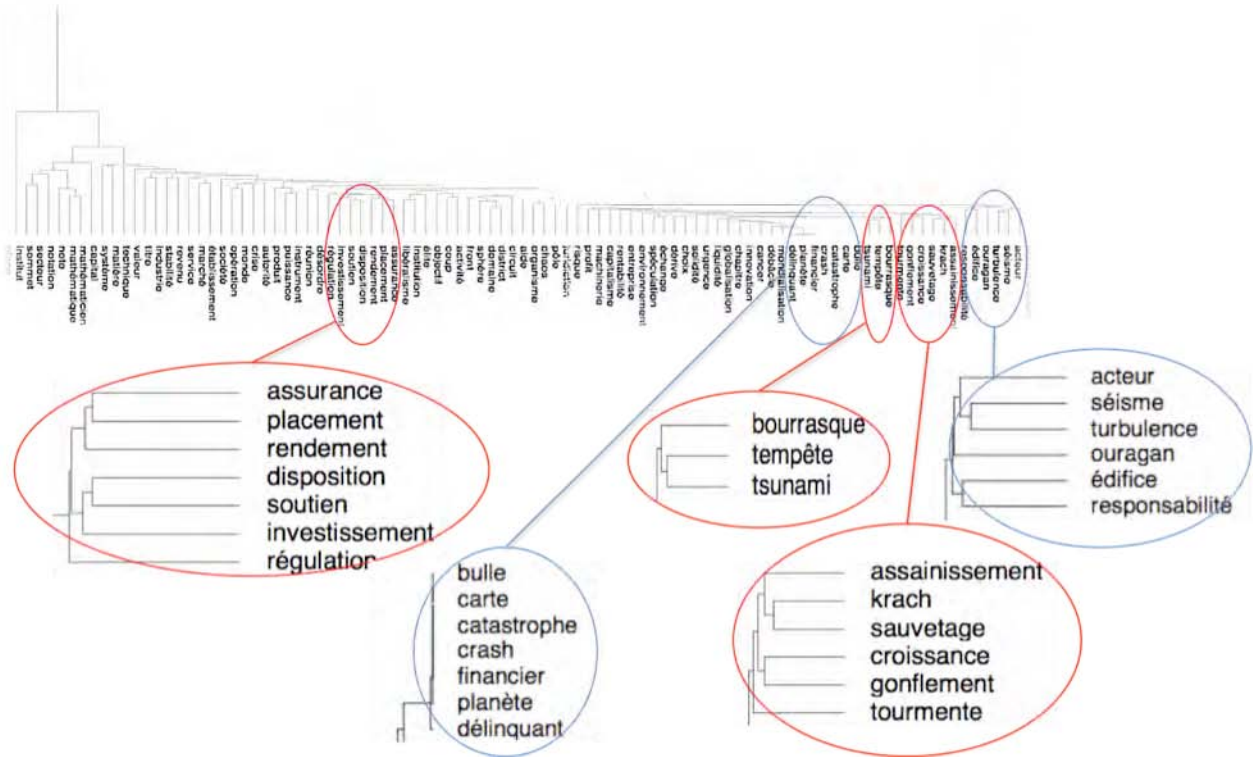


Figure III.1.40 : Classification ascendante hiérarchique réalisée sur les cooccurrents d'ordre 2 de tsunami

Une exploitation rigoureuse des CAH obtenues (c'est-à-dire en traçant une ligne transversale à une hauteur donnée de l'arbre et en récupérant les sous-groupes issus de cette section) n'a pas permis de dégager des regroupements qui soient systématiquement cohérents d'un point de vue sémantique. Certains groupes aberrants sont dus à des anomalies au niveau de la procédure. Par exemple, le groupe {*bulle, carte, catastrophe, crash, financier, planète, délinquant*} est constitué d'éléments qui n'ont que des coefficients nuls, car SEMY n'a pas réussi à leur affecter de signifié (orthographe de *crash* non identifiée par exemple). D'autres groupes discutables, tels que {*acteur, séisme, turbulence, ouragan, édifice, responsabilité*} comportent des éléments en affinité sémantique ({*séisme, turbulence, ouragan*}), mais ils sont composés de sous-ensembles hétérogènes, ou ils comportent des éléments sans lien avec les autres (par exemple *acteur*). Cependant, des groupes présentant une forte cohérence sémantique apparaissent également, comme {*bourrasque, tempête, tsunami*}.

Comme aucune manière systématique d'exploiter les résultats n'a été trouvée, une sélection manuelle de classes a été effectuée. Quelques groupes ont été sélectionnés selon les critères suivants : les éléments sont des descendants d'un même nœud ; la taille du groupe reste limitée et les groupes sont ressentis comme sémantiquement cohérents. Les résultats des CAH ont donc été utilisés comme guides pour une constitution manuelle de classes. Nous avons procédé ainsi pour pouvoir proposer une méthodologie complète et réaliser la dernière étape

de confrontation des groupes aux listes ordonnées obtenues à partir de la distribution des unités lexicales en corpus.

c- Comportement des groupes sémantiques dans la liste distributionnelle

Les groupes sémantiques extraits de la CAH, fondés sur les liens lexicographiques correspondants, sont comparés à la liste de cooccurrents d'ordre 2 pondérés en fonction de leur distribution en corpus et classés par spécificité décroissante (pour les pondérations affectées aux cooccurrents d'ordre 2 à partir des distributions en corpus, *cf.* 3.2).

On cherche à observer dans quelle mesure la distribution témoigne de l'émergence de groupes sémantiquement cohérents, puis à interpréter les saillances en terme d'activation ou d'enrichissement du sens. L'articulation des groupes au sens codé n'a pas fait l'objet d'une procédure automatique : l'analyse en terme d'activation ou d'enrichissement est effectuée manuellement[91].

Trois types de comportements des groupes sémantiquement structurés sont observés : une émergence massive en tête de la liste issue des distributions en corpus, une émergence massive mais dispersée/médiane, une émergence en fin de liste.

Émergence massive en tête de liste : Si le groupe émerge massivement en tête de liste, deux cas de figure se présentent :

- Le groupe émergent reflète un phénomène d'activation d'une partie du sens de la cible lexicale si la facette sémantique qu'elle évoque s'exprimait déjà dans le sens répertorié de la cible.
- Le groupe émergent reflète un phénomène d'enrichissement si la facette sémantique qu'on peut lui associer est étrangère au sens répertorié de la cible

Le tableau ci-dessous met en évidence des groupes cohérents sémantiquement et saillants distributionnellement pour les différentes cibles lexicales.

Cible lexicale	**Groupe sémantique extrait de la CAH**	**Proportion d'unités apparaissant dans la première moitié / dans le premier quart de la liste distributionnelle**	**Qualification de l'activation ou de l'enrichissement par rapport au sens de la ressource lexicographique**
bouclier[92]	cadeau, besoin, ristourne, concurrence, minima, exonération, largesse	100% / 57%	Facette sémantique nouvelle, correspondant à l'idée d'allègement ou de cadeau financier

[91] Une telle articulation aurait nécessité un travail complémentaire au niveau de la ressource lexicographique. Il faudrait pour cela établir une mesure sur le graphe de dictionnaire pour évaluer la distance de la cible lexicale aux différents regroupements.

[92] Les deux regroupements associés à *bouclier* proviennent en fait d'un seul regroupement de la CAH, hétérogène, constitué de {cadeau, besoin, ristourne, concurrence, minima, exonération, largesse, médias, efficacité, urgence, force, riposte, catastrophe}. Si on exclut *médias*, on distingue deux sous-groupes sémantiquement homogènes, l'un qui renvoie à l'idée d'allègement financier ({cadeau, besoin, ristourne, concurrence, minima, exonération, largesse}), l'autre à l'idée de danger et de réaction face à ce danger ({efficacité, urgence, force, riposte, catastrophe}). Les deux sous-ensembles ont été séparés manuellement car, au niveau de la CAH, ils sont étroitement imbriqués et aucun critère ne permet de les dissocier. Ces deux sous-groupes ont des positionnements opposés dans la liste distributionnelle : la facette liée à l'allègement est en tête de liste, tandis que la facette liée à la réaction face à une menace est totatement absente du premier quart de la liste distributionnelle et se concentre plutôt en milieu de liste.

pourri	hypothécaire, bancaire, interbancaire, automobile, issu, subprime, spéculatif, gouvernemental, créditeur	66% / 33%	Facette sémantique nouvelle, rattachée à la finance et l'économie réelle
toxique	hypothécaire, immobilier, interbancaire, bancaire, automobile, subprime, emprunteur, créditeur	87,5% / 50%	Facette sémantique nouvelle, rattachée à la finance et l'économie réelle
actif	créance, dépôt, institution, crédit, public, dette, souveraineté	71% / 71%	Activation d'un sens existant, rattaché aux produits et instruments financiers
tsunami	turbulence, acteur, séisme, responsabilité, ouragan, édifice	67% / 50%	Activation de l'idée de catastrophe naturelle
	tourmente, croissance, krach, sauvetage, gonflement, assainissement	67%/50%	Activation de l'idée de perturbation, de danger, d'effondrement, couplée à une connotation financière qui participe d'un enrichissement
tempête	capitalisme, rentabilité, spéculation, nationalisation	75%/75%	Facette sémantique nouvelle, associée principalement à l'idée de finance

Tableau III.1.41 : Activation de facettes sémantiques portées par les groupes extraits des CAH

Émergence massive dispersée ou médiane : si le groupe est dispersé ou en zone médiane, on considérera qu'il a un effet sémantique s'il est suffisamment massif, puisqu'il ne se distingue pas par un positionnement remarquable dans la liste issue des distributions en corpus. Un groupe de grande taille sera le reflet de la présence d'un fond sémantique qui viendra soit activer une facette sémantique déjà existante pour la cible lexicale, soit connoter son sens. Cette facette sémantique ne sera cependant pas considérée comme dominante dans le sens de la cible lexicale.

Cette dispersion à tous les niveaux de la liste est présente pour le grand groupe sémantique {*tempête, planète, débâcle, innovation, cancer, bulle, bourrasque, carte, chapitre, crash, délinquant, liquidité, mondialisation, catastrophe, globalisation*} associé à *tsunami*, dont les unités évoquent l'idée de catastrophe naturelle planétaire. Il s'agit d'un fond sémantique qui fait écho à une facette sémantique dont *tsunami* est porteur.

De même, pour *actif*, le groupe {*débâcle, panique, filiale, financement, sauvetage*} est présent de façon compacte en zone médiane de la liste distributionnelle. Il ajoute au sens *d'actif* une connotation, pas dominante mais qui s'exprime de façon latente, connotation associée à l'idée d'effondrement et de gestion de cet effondrement.

Émergence en fin de liste : si le groupe est en fin de liste, le sens auquel il correspond ne s'associe pas à celui de la cible lexicale. L'inhibition peut être observée à travers des groupes plus réduits, dont la faible taille témoigne en elle-même : le petit nombre d'éléments renvoie à la faible expression de la facette sémantique correspondante. Deux cas de figure se présentent :

- Si le contenu sémantique véhiculé par ce groupe appartient aux potentiels de sens de la cible lexicale, il y aura inhibition de facettes sémantiques correspondantes ; ce phénomène apparaît de façon frappante sur un petit groupe associé à *bouclier* qui, du fait même de sa petite taille, témoigne de l'inhibition de la facette sémantique correspondante : le groupe {*combat, lutte*}, qui évoque le sens de conflit associé à celui de bouclier, est très distant des autres groupes dans la CAH et se retrouve parmi les cooccurrents en queue de liste dans la hiérarchie distributionnelle. Le sens de conflit semble donc inhibé dans les emplois de *bouclier* en contexte de crise financière.
- Si le contenu sémantique n'est pas associé aux potentiels de sens de la cible lexicale, il correspondra à un sens associé aux cooccurrents d'ordre 1, mais qui restera distinct de celui de la cible et ne participera pas à son enrichissement. Ainsi, pour *tsunami*, le groupe {*front, sphère, domaine, circuit, district*}, n'a que *sphère* qui émerge dans la première moitié distributionnelle. On peut considérer que l'idée de champ ou d'ensemble auquel ce groupe renvoie ne se rattache pas au sens de *tsunami*.

d- Bilan

L'intérêt de l'étude sur les cooccurrents d'ordre 2 est principalement méthodologique. L'expérience est exploratoire et les techniques utilisées demandent à être consolidées et affinées. L'idée qui sous-tend la démarche est d'analyser les saillances en corpus à l'aide de regroupements sémantiquement cohérents obtenus à partir des relations lexicographiques du TLFi, puis mis en relation avec le sens de la cible lexicale. La constitution de classes est une piste pour faire face au problème de la dispersion des données. Même si la qualité des résultats est loin d'être optimale, une caractérisation sémantique du sens des cibles se dessine, avec l'émergence de facettes sémantiques déjà rattachées à l'ancien sens ou, pour les cibles néologiques, des facettes sémantiques nouvelles, susceptibles de participer à un enrichissement.

5. Récapitulation des résultats d'expérience

Nous avons proposé une procédure étayée par des expériences illustratives, qui caractérise l'évolution de sens à partir des variations significatives en discours, puis à partir d'une confrontation avec le sens codé de ces variations en discours. Le schéma générique de traitement est décliné successivement selon trois niveaux de granularité : supra-lexical (domaines), lexical et infra-lexical (traits sémantiques autres que les domaines) (figure III.1.42.a ci-dessous).

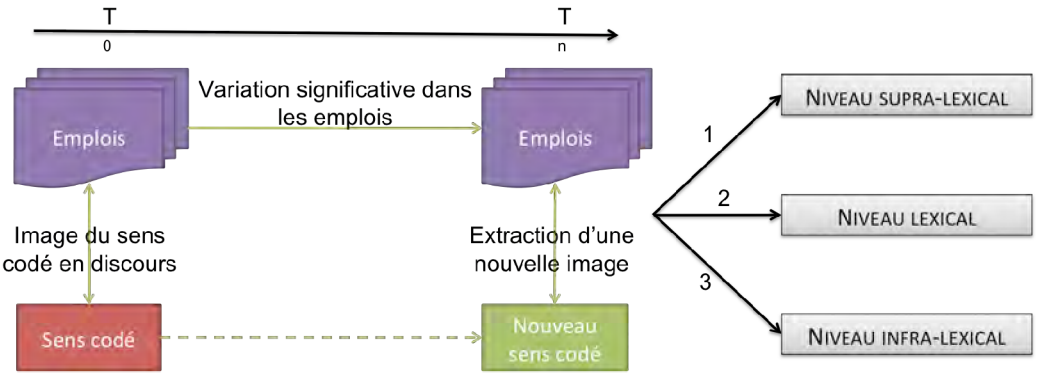


Figure III.1.42.a : Schéma générique de l'analyse par niveaux de granularité

Les résultats sont repositionnés par rapport aux traitements annoncés en début de chapitre (tableau III.1.4, sous-section 1.3) et récapitulés par niveau ci-dessous.

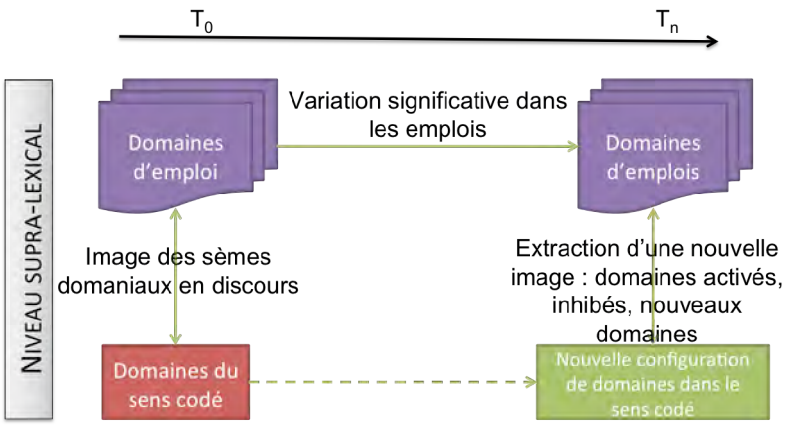


Figure III.1.42.b : Schéma générique appliqué au niveau supra-lexical

Niveau	Sémème de la cible	Observables	Analyse	Résultats sur les cibles
Supra-lexical	Domaines (sèmes macrogénériques)	Domaines textuels (domaines donnés)	Existence d'une évolution	*tsunami toxique tablette* ~~*moléculaire*~~ *numérique* ~~*dangereux délétère*~~ *tempête* ~~*raz-de-marée*~~
			Domaines émergents	*toxique* : ECO *tsunami* : ART, SOC *numérique* : multiples (ART, MOD, POL, SAN, SCI, SOC) *tablette* : multiples (ART, ECO, POL, SCI) *tempête* : ECO, POL, SOC
			Régularité d'évolution	*toxique* : accroissement d'ECO, avec pic en 2009 *tsunami* : pic en 2005 *numérique* : croissance progressive *tablette* : pic en dernière période (2010) *tempête* : évolution variable ; accroissement des domaines émergents
			Structuration	*toxique* : ECO≈(ART, POL), ECO≠ENV *tsunami* : ART≈ECO, SOC≈(SCI,ENV), (ART, SOC)≠(MOD, POL, SAN) *numérique* : les domaines évoluent tous ensemble de façon similaire, sans que des distinctions plus précises se dessinent nettement *tablette*: SAN à part, autres domaines liés *tempête* : pas de structure nette, comportement des domaines dissociés
		Dom. relatifs au sens codé	Reconfiguration	*toxique :* inhibition de tous les domaines du sens codé (BIOLOGIE, CHIMIE, MEDECINE, PHARMACOLOGIE, PHYSIOLOGIE) ; pas d'activation
			Enrichissement	*toxique* : FINANCES

Tableau III.1.42.c : Résultats aux différentes étapes de l'analyse supra-lexicale[93]

Les domaines permettent d'appréhender le changement de sens et l'émergence d'un nouveau sens de façon précise et structurée. Dans les expériences, leur analyse donne des clés pour qualifier la progression de la diffusion, pour coupler ou opposer certains sous-ensembles de domaines. L'articulation des domaines au sens codé d'une cible lexicale apporte des informations pertinentes sur la reconfiguration du sémème ou son enrichissement.

[93] ***Légende du tableau III.1.42.c*** – ART : ARTS ET SPECTACLES ; ECO : INFORMATIONS ECONOMIQUES ; ENV : ENVIRONNEMENT ; MOD : MODE DE VIE ; POL : POLITIQUE / RELATIONS INTERNATIONALES ; SCI : SCIENCES ET TECHNOLOGIES ; SOC : SOCIETE / COMMUNAUTE / TRAVAIL ; ≈ : forte affinité entre domaines ; ≠ : opposition entre domaines.

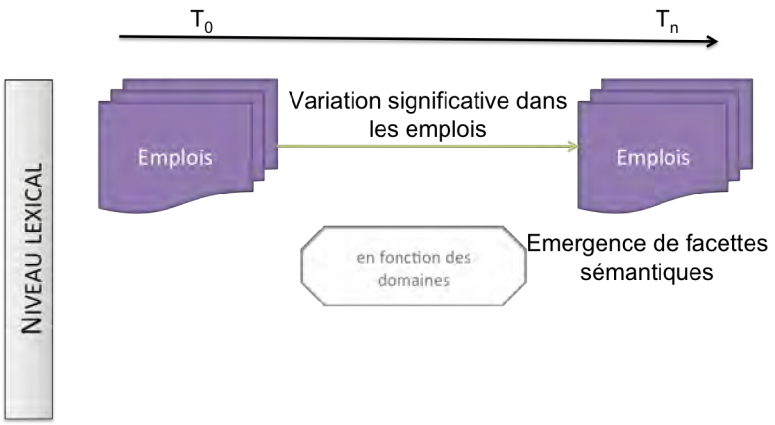


Figure III.1.42.d : Schéma générique appliqué au niveau lexical

Niveau	**Sémème de la cible**	**Observables**	**Analyse**		**Résultats sur les cibles**
Lexical		Unités lexicales	Sélection et pondération	confrontation de domaines	Pour *toxique*, saillance d'un petit nombre d'unités lexicales évoquant principalement la thématique de la finance et de la crise.
				au sein d'un domaine	Pour *toxique*, la thématique générale cède le pas à des notions économiques plus précises. Le caractère négatif de *toxique* se perçoit, mais de façon très diffuse.
				par période	Certaines facettes sémantiques d'*Outreau* identifiées dans l'analyse manuelle de M. Lecolle et caractéristiques de l'émergence par strate d'un nouveau sens sont sensibles au niveau des unités lexicales saillantes aux différentes périodes.
			Sélection et pondération (cooc. d'ordre 2)		Constitution de listes hiérarchiques en fonction de la distribution en corpus, par rapport auxquelles on cherchera à positionner des facettes sémantiques communes portées par plusieurs unités lors de l'analyse infra-lexicale.

Tableau III.1.42.e : Résultats aux différentes étapes de l'analyse lexicale

Le sens des cibles lexicales apparaît à travers les unités lexicales saillantes. Les différents axes d'observation du corpus apportent des informations complémentaires. Certaines facettes sémantiques associées au sens des cibles s'expriment parfois de façon diffuse. Le lien avec le sens codé est établi à travers des parcours interprétatifs. Les unités lexicales constituent ainsi un terrain préparatoire pour l'analyse en traits sémantiques, destinés à établir le lien avec le sens codé de façon plus directe et plus automatisée.

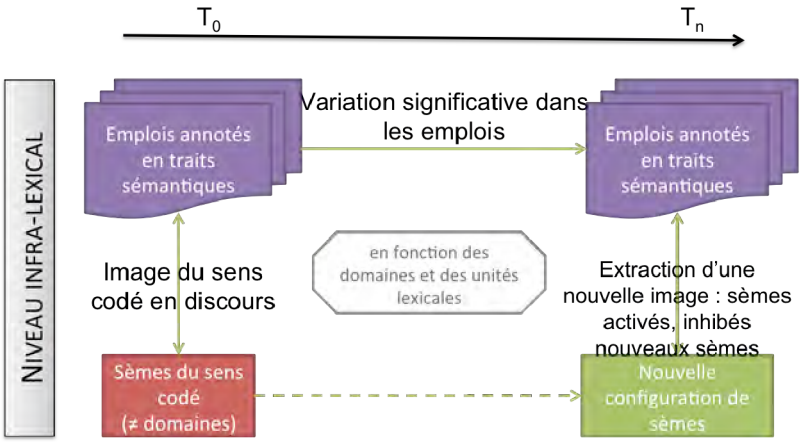


Figure III.1.42.f : Schéma générique appliqué au niveau infra-lexical

Niveau	Sémème de la cible	Observables	Analyse		Résultats sur les cibles
Infra-lexical	Sèmes microgénériques ou spécifiques	Traits sémantiques	Reconfiguration du sémème		*toxique* : inhibition de traits non pertinents ; pas d'activation (traits pertinents non identifiés) *Outreau* : modulation dans le temps globalement conforme à l'analyse manuelle (mais problème d'interprétabilité des traits sémantiques et de saillances pas suffisamment marquées)
			Enrichissement	Apport des sèmes	convergence avec le plan lexical explicitation de facettes sémantiques diffuses
				Renforcement sém/lex	*toxique, Outreau :* renforcement et explicitation de facettes sémantiques sensibles sur le plan lexical
				Enrichissement sèm/lex	*toxique* : récupération de traits relatifs au domaine émergent FINANCES (/bilan/, /banque/, /prêt/), mais bruit
			Constitution de classes (cooc. d'ordre 2)		Émergence de regroupements correspondant à des facettes sémantiques nouvelles ou à l'activation de facettes existantes, mais manque de qualité dans les regroupements faute de récurrence suffisante au niveau des traits sémantiques du dictionnaire.

Tableau III.1.42.g : Résultats aux différentes étapes de l'analyse infra-lexicale

Les traits sémantiques interviennent soit pour l'annotation de corpus, soit pour structurer les unités lexicales en groupes sémantiquement cohérents. Leur utilisation est beaucoup plus délicate que celle des domaines et certains résultats sont discutables. Ceci s'explique notamment par la méthodologie choisie pour représenter les traits sémantiques et par la dispersion des traits sémantiques dans le dictionnaire, à l'origine d'un manque de récurrence. L'application de filtres permet d'améliorer la qualité et l'interprétabilité des résultats. Les filtres appliqués sont l'appartenance au sémème de la cible, la dépendance à un domaine ou le recoupement avec les unités lexicales. Les relations entre traits sémantiques et entrées dans le dictionnaire pour structurer les unités lexicales et observer l'émergence de classes sémantiquement cohérentes ont fait l'objet d'une expérience exploratoire qui reste à consolider, mais dont les résultats présentent des tendances encourageantes.

Le chapitre qui s'ouvre a pour objectif de tirer un bilan plus général des apports et limites du protocole et de définir des pistes complémentaires pour consolider le modèle et l'enrichir.

Chapitre III.2
Bilan et perspectives

1. Apports et limites des premières expériences

Le cœur d'un processus d'allocation de signifié a été établi à travers la définition d'une chaîne de traitements et des expériences illustratives pour étayer chaque étape. Ce processus s'applique à un type de cibles lexicales, relevant de la néologie sémantique. Les autres types de néologie n'ont pas été traités expérimentalement, mais leur articulation avec la néologie sémantique a été précisée (*cf.* chapitre I.3, 3.2.2). Les étapes du processus sont les suivantes :

- **Génération d'une représentation du contenu sémantique des unités lexicales**. Cette représentation est construite à partir du sens codé, elle se présente comme un ensemble de sèmes extraits des définitions lexicographiques. En pratique, le sémème est un sac de sèmes, regroupés de façon indifférenciée quelle que soit la définition dans laquelle le sème apparaît. Deux types de sèmes sont présents, associés à deux niveaux de granularité de l'information : les domaines, associés à des sèmes mésogénériques, et les sèmes définitoires, considérés comme microgénériques ou spécifiques. La représentation sémique utilisée dans le cadre expérimental est moins structurée que la représentation théorique sous forme de molécule sémique décrite au (chapitre I.3, 3.1.2). Ce format de représentation plus grossier que le format idéal a suffi pour faire émerger des unités saillantes pertinentes, malgré un certain bruit.
- **Extraction de représentations du sens des cibles en corpus**. Le sens en corpus a été construit à partir d'approches cooccurrentielles d'ordre 1 ou 2, dans un espace textuel structuré relativement à la cible lexicale (décomposition de l'espace en fonction du voisinage de la cible). Les traitements statistiques appliqués aux cooccurrences peuvent se concevoir comme une série de schémas d'urnes. Derrière cette apparente simplicité, il existe une réelle complexité qui ne réside pas dans la structure élémentaire, de type "sac de", mais dans la multiplicité des découpages de l'espace textuel et dans la combinaison des sources, des observables et des structures associées à chaque découpage. Trois paramètres ont structuré l'espace textuel : le temps, les domaines et le voisinage de la cible. L'analyse textuelle a été réalisée par strates, à partir d'observables de granularité décroissante (domaines, unités lexicales, sèmes).
- **Confrontation des représentations du sens en corpus à la représentation du sens codé**. La confrontation s'est appuyée sur une distinction entre les unités de sens déjà associées à la cible dans le dictionnaire (perspective d'activation – inhibition de sèmes) et les nouvelles unités (perspective d'enrichissement du sémème). Les nouveaux sèmes ont été regroupés manuellement en classes.

Le noyau dur de la procédure a été défini et testé, il reste à étendre sa portée applicative et à consolider ou améliorer certains points, en intégrant des traitements intermédiaires et en automatisant certains tests réalisés de façon fragmentée à petite échelle.

Une attention particulière doit être portée à la résolution des points suivants :

- L'**ambiguïté des unités lexicales et sémiques**. Cette ambiguïté est due au formalisme adopté pour représenter les sèmes, à la déstructuration des définitions au sein d'une entrée et à une représentation des unités lexicales en corpus qui s'apparente plus à celle de *vocable* (unité sous forme lemmatisée qui correspond à un signifiant et plusieurs sens liés associés à ce signifiant, d'après (Polguère, 2008:59)) que de *lexème* (unité simple associée à un sens donné, sous forme lemmatisée ; d'après (Polguère, 2008:50)).
- La **dispersion des données**. Ce problème est récurrent dans le traitement de données langagières. Dans notre cadre, il est démultiplié par l'articulation des ressources lexicographique et textuelle, chacune étant affectée isolément par le problème d'éparpillement des données.
- La structuration et la validation essentiellement manuelles des résultats.

Des élargissements sont possibles sur plusieurs plans :

- pour le **type de cibles** traitées, qui pourraient inclure des néologies de forme ou des phrasèmes néologiques ;
- pour des **représentations du sens codé plus riches**. Par exemple, la représentation du sens codé manque de nuances : seule compte la relation d'inclusion, il n'y a pas de distinction selon qu'il s'agit de relation synonymique, hyperonymique, métaphorique, etc. ;
- pour **intégrer des paramètres textuels** susceptibles d'influer sur l'émergence d'unités de sens significatives, comme la présence d'indices conscients de néologie ;
- pour **compléter l'approche isotopique**. Nous avons fait le choix de rechercher le nouveau sens à travers des contrastes entre fonds et formes sémantiques (recherche d'isotopies et de faisceaux d'isotopies) . Ce choix a amené à privilégier des liens paradigmatiques au détriment des relations syntagmatiques ; il a également mis au premier plan la cooccurrence et légitimé le cadre statistique. La représentation du sens propre à cette approche pourrait être complétée par une approche répondant une perspective différente, fonction des liens syntagmatiques.

Plutôt que d'aborder les perspectives à travers des axes transversaux, nous préférons les organiser en fonction des grandes composantes du processus, à savoir les cibles (section 2), la représentation des sens codés (section 3) et la représentation des sens en discours (section 4). Enfin, des développements seront proposés pour le système dans son ensemble.

2. Élargissement du protocole à d'autres cibles

Le protocole peut être étendu à d'autres types de néologies que la néologie sémantique : emprunts, néologismes issus de procédés morphologiques, nouveaux phrasèmes (*cf*. chapitre I.2, 3.4). La particularité de ces autres cas de néologie est que les cibles n'ont pas d'entrée correspondante dans le dictionnaire. De ce fait, l'élargissement du protocole à ce type de cibles suppose un traitement supplémentaire destiné à fournir un sémème analogue à celui tiré des sens codés. On qualifiera ce sémème de *sémème candidat*. Les traitements proposés pour la néologie sémantique permettent de tester la validité des sémèmes candidats et d'enrichir des sémèmes s'ils s'avèrent incomplets.

Pour obtenir un sémème candidat, il faut mettre en place des prétraitements qui varient selon le type de néologie.

Pour les ***emprunts***, un sémème candidat peut être obtenu à partir de ressources externes. Les outils et ressources intermédiaires pourraient avoir les fonctions suivantes :

- identifier la langue d'origine de l'emprunt. Cette tâche répond à une volonté d'exhaustivité, qui intègre une multiplicité de langues sources. Dans la suite des traitements, cela risque d'avoir un coût en termes de ressources à disposition : il faut disposer de ressources multilingues ;
- disposer d'une ressource lexicographique externe dans la (ou les) langue(s) retenue(s) ;
- établir l'analogie entre la cible lexicale en français et l'entrée en langue étrangère, par traduction ou analogie formelle (notamment pour compléter une approche sur des emprunts tels que les calques) ;
- si la ressource étrangère est dotée d'une base analogue à la base SEMEME (*cf.* chapitre II.1, 1.1.2), en extraire les sèmes et adapter leur représentation formelle à la langue d'origine (traduction) ; sinon, traduire le contenu de l'entrée et extraire de l'entrée traduite un sémème de format analogue aux sémèmes habituellement générés. Dans ce dernier cas, en plus des outils de traduction, un outil spécifique d'extraction de sèmes est nécessaire.

Pour les ***néologismes issus de procédés morphologiques***, des outils tels que DériF (Namer, 2009) permettraient de décomposer le néologisme en deux types de composantes : une racine et des opérations sur cette racine. Par exemple, dans les résultats ci-dessous, DériF décompose *titrisation* en la racine *titre (NOM)* et deux opérations successives sur cette racine, celle rattachée au suffixe –ion (Xion = action de X) et celle rattachée au suffixe –iser (Xiser = transformer en X), comme :

> « titrisation/NOM==> [[[titre NOM] iser VERBE] ion NOM] (titrisation/NOM, titriser/VERBE, titre/NOM) " (Action - résultat de l'action) de titriser" »
>
> « titriser/VERBE==> [[titre NOM] iser VERBE] (titriser/VERBE, titre/NOM) " Faire (de - un(e)) - transformer en - soumettre à l'action de (un(e)) - mettre dans (un(e)) titre; faire le titre" »
>
> (DériF, requêtes 'titrisation (substantif)' et 'titriser' (verbe), http://www.cnrtl.fr/outils/DeriF/requete.php)

La racine peut être associée à une entrée du dictionnaire, donc des sèmes peuvent lui être affectés. Le rapport entre les opérations sur la racine et les sèmes est moins immédiat. Un travail d'articulation des opérations sur la racine aux sèmes est nécessaire. Deux options se présentent :

(1) ramener les opérations morphologiques aux sèmes. Par exemple, le suffixe *-ion* traduit en "action de – résultat de l'action de" dans *titrisation* deviendrait le trait sémantique /action/. De même, le suffixe *dé-*, qui exprime l'opération "enlever – (faire) sortir de", pourrait être associé au trait sémantique /enlever/, par exemple pour *déconventionner*. Convertir de la sorte les formulations explicites des opérations morphologiques en traits sémantiques peut générer des problèmes : les traits sémantiques produits ainsi n'entretiennent pas de dépendance les uns aux autres et leur interprétation peut devenir délicate, notamment en raison de leur caractère prédicatif ;

(2) changer le format de représentation des sèmes pour les ramener aux opérations morphologiques. Dans ce dernier cas, il conviendrait d'introduire une distinction entre deux types de sèmes : des sèmes qui opèreraient des transformations sur d'autres sèmes, comme des

fonctions, et des sèmes de contenu[94]. Les sèmes fonctionnels pourraient être définis à partir d'un rapprochement avec les fonctions lexicales. Quelle que soit la solution retenue, la démarche exige de mettre en place tout un système d'analogie entre les opérations de construction morphologique et les sèmes.

Pour les ***nouveaux phrasèmes***, leur étude nécessiterait d'une part d'intégrer les lexies complexes aux représentations lexicographiques et textuelles, d'autre part de mettre en place des outils d'analyse spécifiques. Dans le TLFi, il est actuellement possible d'élargir l'ensemble des entrées aux lexies complexes que sont les syntagmes définis dans le TLFi – avec les limites que présentent ceux-ci (*cf.* chapitre II.1, 2.3.1). Un projet plus ambitieux, mais aussi garant d'une plus grande qualité, serait de coupler un dictionnaire de collocations tel que le DiCo (Mel'Xuk et Polguère, 2006) au TLFi. Pour que les lexies complexes du dictionnaire soient utilisables, elles doivent pouvoir se retrouver au niveau des ressources textuelles. Ceci exige une segmentation des textes appropriée, incluant notamment l'extraction d'unités complexes analogues à celles du dictionnaire. Il faudrait également des outils pour relier les unités complexes des discours à celles des ressources lexicographiques. Une fois qu'on dispose de représentations adaptées (textuelles et lexicographiques), l'étude des nouveaux phrasèmes nécessite deux modules d'analyse supplémentaires. Le premier module serait destiné à la détection du figement. Le second module servirait pour la construction d'un sémème de même nature que le sens codé. Un sémème candidat peut par exemple être construit par réunion des sémèmes associés à chaque composante du syntagme. Par exemple, pour *économie réelle* ou respectivement *cuisine moléculaire*, le sémème candidat s'obtiendrait par réunion du sémème d'*économie* et de *réelle* (respectivement de *cuisine* et de *moléculaire*). Le sémème obtenu par réunion peut ensuite être retravaillé selon les informations fournies par les corpus.

3. Enrichir la représentation des sens codés

La représentation des sens codés peut être améliorée de deux façons. D'une part, pour aller plus loin, il conviendrait d'intégrer plus d'informations propres à la ressource lexicographique, dont la richesse est sous-exploitée. D'autre part, le formalisme choisi génère du bruit ou une perte d'informations, ce qui nécessite de retravailler certains aspects de la représentation du sens. Les améliorations à apporter portent sur la microstructure et sur la macrostructure du dictionnaire.

3.1 Microstructure : améliorer les sémèmes

3.1.1 Résoudre le problème d'interprétabilité des sèmes

Dans notre procédure, les définitions lexicographiques d'une entrée sont déstructurées pour générer un sac de sèmes. Certaines unités se retrouvent isolées et leur interprétation pose problème lorsqu'elles sont affranchies des apports sémantico-syntaxiques des définitions rédigées.

Cette situation se présente lorsque les sèmes prennent la forme d'unités à caractère prédicatif telles que /existence/ (de quoi ?) ou /constitution/ (de quoi ?). Pour y remédier, il faudrait réviser de façon assez radicale le formalisme choisi afin d'introduire des dépendances entre

[94] Cette distinction est dans l'esprit de celle proposée par (Bastuji, 1974). L'auteure se situe à un palier supérieur au mot. Elle distingue les traits inhérents, qui définissent le contenu sémantique d'une unité, et les traits de sélection, qui régissent la combinatoire d'unités lexicales. Elle se positionne donc au niveau de la syntaxe. Pour se placer au niveau de la morphologie, comme nous le proposons, il faut transposer cette distinction à un palier inférieur au mot.

unités ou en distinguant les unités de contenu et les opérations appliquées à ces unités. L'introduction de dépendances demanderait une analyse syntaxique des définitions. L'adoption d'un formalisme distinguant les unités de contenu et les opérations appliquées à ces unités demanderait par exemple de recourir à des fonctions lexicales, comme évoqué précédemment. Cette solution n'est envisageable qu'après une reconfiguration générale des descriptions lexicographiques, telle qu'amorcée par le projet RLF (Lux-Pogodalla et Polguère, 2011).

Un problème du même ordre se rencontre lorsque la représentation des sèmes prête à ambiguïté, comme pour /titre/, qui intervient aussi bien dans des définitions financières (*action, coupon, taxe, régulariser*) que sportives (*champion, challenger*) ou encore relatives à des distinctions honorifiques (*duc, führer, marquis*). Pour lever l'ambiguïté, une désambiguïsation préalable des unités composant les définitions serait nécessaire, ainsi que l'utilisation de représentations non ambiguës. Par exemple, il faudrait distinguer $titre_1$ (désignation honorifique) de $titre_2$ (vainqueurs sportifs) et de $titre_3$ (produit boursier).

Par ailleurs, les sèmes issus du métalangage lexicographique introduisent de la confusion. Par exemple, l'unité lexicale *foie* est définie comme suit :

> 'Organe vital (au même titre que le cœur, les poumons)'.

Le sème /titre/ lui est associé, à cause de la présence de *au même titre que* et faute d'un filtrage suffisamment sélectif : le critère de sélection des sèmes est la catégorie grammaticale (noms, verbes, adjectifs et adverbes) et les syntagmes prépositionnels comme *au même titre que* ne sont pas identifiés comme tels ni éliminés. Une solution serait d'établir une liste d'exclusion (stop-list) sur les expressions du métalangage et les locutions grammaticales (prépositionnelles, adverbiales et conjonctives), puis de filtrer la base SEMEME à partir de cette liste. Cette solution est de moindre envergure que les solutions proposées pour les autres problèmes, mais elle ne dépend pas d'un formalisme complexe, elle pourrait donc être a priori plus facile et plus rapide à mettre en œuvre. Un risque est qu'elle génère du silence. Une solution alternative consiste à voir dans les expressions du métalangage et les locutions grammaticales des opérations ou des relations sémantiques entre sèmes de contenu. Comme précédemment, l'enjeu serait de convertir le métalangage en fonctions lexicales.

3.1.2 Maintenir la division en acceptions et les liens de dépendance

Une description théorique du sémème sous forme de molécule sémique avait été proposée au (chapitre I.3, 3.1.2).

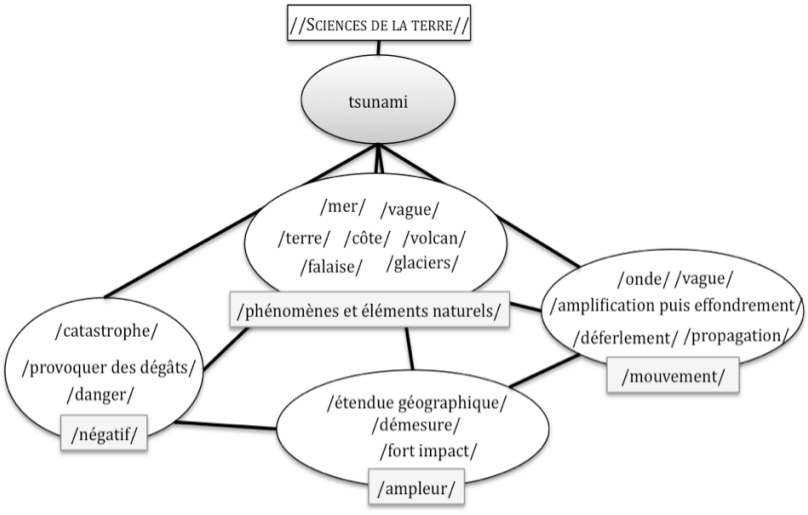


Figure III.2.1 : Sémème structuré de tsunami

Le sémème se présentait comme un ensemble structuré, composé de sous-ensembles, dont les éléments étaient de degré de généricité variable. De fait, dans les expériences réalisées, la représentation sémique est moins structurée : les sèmes sont regroupés en un seul ensemble, la seule distinction maintenue est celle de la granularité, avec une distinction entre sèmes mésogénériques (domaines) et sèmes microgénériques ou spécifiques (sèmes définitoires).

La structure actuelle de la base SEMEME permettrait de récupérer les subdivisions en définitions. L'ensemble de sèmes serait constitué de sous-ensembles et cette structure serait adaptée à une désambiguïsation en contexte, notamment par élimination de sous-ensembles non pertinents (associés par exemple au domaine BIOLOGIE lorsque le contexte renvoie à de l'ÉCONOMIE). Ces subdivisions permettraient également de maintenir les dépendances de sèmes définitoires à des domaines, ce qui contribuerait à résoudre l'ambiguïté induite par le format de représentation des sèmes. Ainsi, dans les définitions, la dépendance de /titre/ au domaine du SPORT, de l'ECONOMIE ou de l'HISTOIRE contribue à lever l'ambiguïté de /titre/ considéré isolément. La représentation d'un sème pourrait être alors envisagée comme un couple (sème, domaine).

3.1.3 Conserver les différents types de liens sémantiques

La représentation sémique ne s'appuie que sur un type de relation entre sème et mot-vedette, à savoir la relation d'inclusion. Dans les données lexicographiques, les relations sémiques sont plus nuancées, elles peuvent être de nature différente. La qualité et la précision du sémème seraient renforcées si l'on conservait les informations sur les différents types de liens.

Il est possible d'utiliser les informations sur les liens sémantiques de plusieurs façons :

- pour pondérer les sèmes. Par exemple, un sème obtenu à partir d'une relation de synonymie pourrait être considéré comme plus proche du mot-vedette qu'un sème provenant d'un emploi par extension ou qu'un sème provenant d'un emploi rare ;
- pour filtrer ou décomposer l'information sémique. Par exemple, pour analyser l'activation ou l'inhibition de sèmes, les relations sémiques permettraient d'établir un distinguo entre des sèmes provenant d'emplois métaphoriques de sèmes sans lien particulier ;
- pour servir de brique de base à des réseaux construits au niveau de la macrostructure du dictionnaire. On y reviendra en sous-section 3.2.

Un certain nombre de relations sémantiques sont actuellement codées (indicateurs d'emplois métaphoriques, de domaines, de synonymie, etc.), mais toutes les relations n'apparaissent pas systématiquement et d'autres ne sont pas encodées. Il existe notamment des liens implicites, accessibles à un interprétant ou lors d'une analyse linguistique manuelle, mais pas à un traitement automatique en l'état actuel. De plus, la rédaction du TLF s'est étendue sur 30 ans et les normes de rédaction ont évolué. De ce fait, les informations encodées ne sont pas homogènes entre les différentes définitions. Les liens sémantiques encodés sont susceptibles d'avoir évolué. La mise en évidence de tels liens nécessite de vérifier le système dans son intégralité et mettre à jour les relations sémantiques lorsque c'est nécessaire. Le projet RELIEF mettra probablement en place ce type de révision et d'amendement global du système.

3.1.4 Pondérer les sèmes

Dans la représentation des sémèmes, tous les sèmes sont à égalité : ils ont même poids, quelle que soit leur nature et leur emplacement dans l'entrée lexicographique. Seule compte la présence ou l'absence du sème dans l'entrée. Cette parité par défaut n'est pas nécessairement légitime. Par exemple, on peut se demander si /vague (subst)/ et /côte (subst)/ sont de même importance dans la définition de *tsunami*. De plus, ce nivellement est en décalage avec les

résultats des traitements textuels, qui vont au-delà d'une simple représentation binaire : les indices obtenus en corpus permettent des pondérations plus fines, définies sur des échelles de valeurs discrètes ou continues. L'analyse des résultats textuels pourrait permettre des conclusions plus précises et de meilleure qualité si on ramenait les valeurs d'activation ou d'inhibition des sèmes à un ensemble où les sèmes sont d'importance variable plutôt que de même importance. Par exemple, *moléculaire* possède une définition dans le domaine de la PHILOSOPHIE et de la CHIMIE. Si, en contexte, les sèmes domaniaux PHILOSOPHIE et CHIMIE sont inhibés, avec un degré d'inhibition du même ordre de grandeur, il n'est pas aberrant de voir l'inhibition de CHIMIE comme plus significative que celle de PHILOSOPHIE, partant de l'hypothèse que le poids de CHIMIE est a priori plus important que celui de PHILOSOPHIE.

La pondération des sèmes dans les définitions constitue un travail à part entière. Différents paramètres mériteraient d'être testés, à savoir :

- ***La position dans la définition***. On peut supposer comme (Muller *et al*, 2006:68 ; *cf.* chapitre II.2, 1.2.3) que des sèmes en début de définition ont plus d'importance que ceux en fin de définition, moyennant un filtrage du métalangage lexicographique (par exemple, le verbe *qualifier* dans la définition d'*alexandrin : "***Qualifie** le vers de douze syllabes (…)"), des locutions grammaticales (par exemple, *au même titre que* dans la définition de *foie* évoquée précédemment) et autres unités qui ne sont pas sémantiquement pleines (tels que les pronoms et les déterminants). La position peut être aussi considérée de façon relative à certaines composantes, par exemple le genre prochain et les différences spécifiques. Pour définir ce type de position, une analyse des composantes des définitions accompagnée du balisage correspondant est nécessaire. Ce type de balisage fait actuellement l'objet du projet Definiens (Barque *et al.,* 2010).
- ***La définition dans laquelle le sème apparaît.*** Les poids seraient relatifs à des sous-ensembles de sèmes, où un sous-ensemble correspond à une définition constitutive de l'entrée. Ces poids permettraient de hiérarchiser ces définitions et de définir leur importance relative (par exemple, une définition rare ou vieillie serait considérée comme moins importante que les autres ; les sèmes provenant d'une telle définition auraient un moindre poids que ceux d'autres définitions). La base SEMEME ne fournit pas d'information sur l'importance relative des définitions. Au niveau de l'importance relative des définitions, il serait judicieux de pondérer les définitions en s'appuyant sur les emplois en discours. Des bases textuelles pourraient servir à évaluer l'importance relative des définitions en fonction des emplois discursifs. Le poids relatif des définitions peut jouer dans certains cas, comme celui de *tablette* : la néologie n'est pas tant due à l'émergence d'un nouveau sens qu'à un changement de l'importance relative des définitions. Avec les *tablettes numériques*, le sens de *tablette* en informatique devient prépondérant par rapport aux autres sens sur l'ensemble des emplois, et la hiérarchie des définitions est reconfigurée.
- ***Le nombre d'occurrences*** d'un sème dans l'entrée et le nombre total d'occurrences d'un sème dans le dictionnaire. La pondération pourrait être établie par exemple selon la méthode tf-idf. Une telle pondération renvoie à l'idée qu'une unité fortement partagée est moins informative qu'une unité rare, distribuée de façon déséquilibrée.

3.1.5 Réduire la diversité sémique

En corpus, l'annotation sémique est utilisée pour mettre en évidence des effets de récurrence, c'est-à-dire des isotopies. Elle doit augmenter la redondance qu'on pourrait observer sur le plan lexical.

En termes de récurrence, l'apport spécifique de l'annotation sémique a été notable pour les domaines, bien que non optimal (1) ; par contre, pour les sèmes définitoires, c'est-à-dire autres que les domaines, l'apport potentiel de l'annotation a été largement sous-exploité (2).

(1) Pour les domaines, la récurrence n'apparaît pas de façon aussi marquée qu'elle l'est en réalité, étant donné l'ensemble des domaines associés aux définitions. Un premier facteur explicatif est que les domaines ne sont pas tous du même niveau : certains domaines sont des sous-domaines d'un domaine plus général par exemple. Un domaine donné et un de ses sous-domaines sont représentés formellement comme deux sèmes différents, ils sont donc considérés comme des entités distinctes et aucune forme de récurrence n'est identifiée. Sur ce plan, il faudrait soit établir une table de correspondance et définir des règles à appliquer lors de traitements, soit réaliser une harmonisation des niveaux dans la ressource lexicographique, à reporter ensuite dans la base SEMEME, de façon à ce que, dans les traitements automatiques, la présence d'un sous-domaine ne soit pas considérée comme l'absence du domaine dont il dépend. Un autre problème est lié à l'étalement dans le temps du TLFi. La variation des étiquettes de domaines entre les premières et dernières lettres de l'alphabet est à étudier, pour vérifier qu'il n'y a pas de biais à ce niveau. De plus, les domaines n'ont pas été systématiquement précisés dans les définitions, d'où une absence qui réduit la récurrence domaniale en corpus. Pour résoudre ces problèmes, il faudrait un travail de grande ampleur, avec une analyse et une mise à jour systématiques des définitions, réalisé dans le cadre des projets vivants autour du TLFi. Dans un premier temps, il faudrait acquérir des connaissances précises sur la situation, à travers une analyse approfondie de la structure des domaines, de leur ventilation et des lacunes éventuelles. Dans un second temps, il faudrait proposer des modifications et les intégrer à la ressource, afin d'harmoniser les niveaux, combler les lacunes et rééquilibrer les éventuels changements survenus au fur et à mesure de la rédaction.

(2) Pour les sèmes définitoires, l'effet de récurrence a été nettement atténué par le fait que beaucoup de sèmes sont très peu partagés. Les récurrences observées en corpus provenaient plus de la récurrence d'unités lexicales que de la récurrence de sèmes d'une entrée à l'autre du dictionnaire, alors que les deux sources de récurrence (lexicale en corpus, sémique dans les entrées du dictionnaire) auraient dû avoir un poids plus équilibré, voire même déséquilibré de façon inverse, pour légitimer complètement l'annotation sémique (c'est-à-dire avec une récurrence due plus à la répétition de sèmes d'une définition à l'autre qu'à la répétition d'unités lexicales dans les textes). Ce phénomène est dû à l'éparpillement des sèmes dans la ressource lexicographique. Pour améliorer ce point, deux solutions sont envisageables :

- Augmenter le nombre de sèmes associés à chaque définition. Plusieurs techniques sont possibles. On pourrait en particulier inverser la relation d'inclusion pour associer d'autres sèmes au sémème. Plus précisément, au lieu de considérer seulement que x est sème de la cible s'il appartient à sa définition, on peut aussi considérer que x est sème de la cible si la cible appartient à une de ses définitions, ou encore s'il cooccurre avec la cible dans une définition, alternatives qui rejoignent celles évoquées par (Loiseau *et al.*, 2010) pour décrire les relations entre unités. Par exemple, pour *toxique*, le sémème obtenu par relation d'inclusion normale et inversée est le suivant :

Sémème élargi					
sémème par inclusion normale (unités contenues dans la définition de *toxique)*			sémème par inclusion inverse (entrées contenant *toxique(s)*[95]*)*		
/toxique/ /produit/ /minéral/ /animal/ /végétal/ /provoquer/ /intoxication/ /destruction/ /organisme/ /vivre/ /agir/ /négativement/	/individu/ /physique/ /psychisme/ /être/ /contenir/ /poison/ /toxine/ /distiller/ /ouvertement/ /propos/ /médisant/ /dépréciateur/	/empoisonnement/ /quantité/ /devenir/ /devoir/ /présence/ /produire/ /action/ /organe/ /tissu/ /origine/ /cause/ /pollution/	/association/ /tableau/ /tabac/ /ivre/ /diffusion/ /carbone/ /essai/ /sérum/ /résolution/ /ivresse/ /alcaloïde/ /intoxication/ /uranium/ /masque/ /néfaste/ /obus/ /neutraliser/	/toxine/ /phosphore/ /stérile/ /venimeux/ /absinthe/ /phénol/ /stupéfiant/ /arsenic/ /venin/ /putréfaction/ /épuration/ /asphyxie/ /urée/ /renoncule/ /brome/ /arséniate/ /intoxiquer/	/adonis/ /ozone/ /accoutumance/ /inhalation/ /digitale/ /intoxiqué/ /toxicomanie/ /sublimé/ /empoisonnement/ /arsénieux/ /panthère/ /dialyse/ /fixateur/ /pollution/ /béryllium/ /cyanhydrique/

Tableau III.2.2 : Sémème de toxique enrichi par une autre relation lexicographique

- Réduire la diversité sémique, autrement dit, avoir un vocabulaire sémique plus petit. La réduction du vocabulaire sémique peut être effectuée en deux étapes, à savoir définir l'ensemble d'unités qui constituent le vocabulaire, puis y ramener les autres unités. Elle pourrait être réalisée à partir d'une représentation du dictionnaire sous forme d'un graphe établissant des liens entre les entrées. Nous proposons ici deux pistes. La première serait de s'appuyer complètement sur le graphe de dictionnaire : le vocabulaire serait défini par des sommets correspondant à un certain niveau de confluence (on ne conserverait que les sommets dont part un nombre d'arcs supérieur à un certain seuil) ; les sommets trop peu connectés, *i.e.* les sèmes trop rares, seraient ramenés à ces sommets : ils seraient remplacés par les sommets suffisamment connectés les plus proches. La deuxième piste serait de partir d'une liste d'unités préétablie de taille réduite, dans l'esprit des primitives de (Wierzbicka, 1972) ou de la liste réduite utilisée pour les définitions du *Longman Dictionary of Contemporary English*, un dictionnaire anglais pour allophones (Béjoint, 2009:136 ; *cf.* chapitre II.1, 1.3.2,).

3.2 Macrostructure : d'une partition en entrées à un réseau de relations

La macrostructure du dictionnaire a été faiblement exploitée dans le cadre expérimental : les entrées sont considérées comme disjointes et les liens entre entrées sont établis seulement à partir des unités lexicales saillantes en corpus ; de plus, les sèmes ont été regroupés en classes de façon manuelle. Un travail sur la macrostructure du dictionnaire permettrait d'établir des liens entre sèmes et de mettre à jour des classes.

[95] Liste restreinte des unités définies par *toxique*. Seules ont été conservées des unités partagées par d'autres unités lexicales. Ici, seules sont présentées les unités ayant au moins 7 occurrences dans l'ensemble des définitions du TLFi. La liste est triée par nombre d'occurrences décroissant dans les définitions du TLFi.

3.2.1 Structures imbriquées en fonction des domaines

Le TLFi possède les métadonnées nécessaires pour générer des regroupements d'entrées en fonction des domaines d'emploi. Des regroupements par domaines sont déjà accessibles indirectement, par l'intermédiaire du moteur de recherche intégré au dictionnaire électronique.

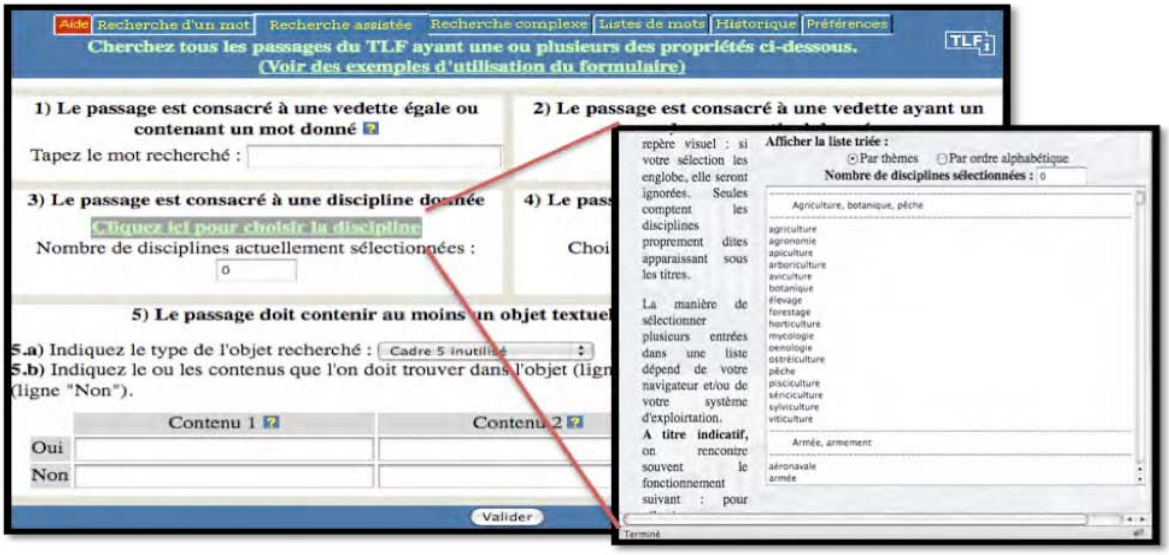


Figure III.2.3 : Capture d'écran du moteur de recherche assistée du TLFi

Le moteur de recherche propose une hiérarchie à deux niveaux, avec des grands domaines (*centres d'intérêt*) et des sous-domaines (*disciplines*). Parallèlement à la hiérarchie intégrée au moteur de recherche, il existe la nomenclature des domaines ayant servi de support à la rédaction qui permettrait des subdivisions encore plus fines : les grands domaines sont subdivisés en sous-domaines, qui eux-mêmes se ramifient en sous-domaines plus précis, jusqu'à un niveau d'ordre 5. Ces subdivisions en domaines et sous-domaines offrent la possibilité de générer des structures imbriquées, avec des classes et sous-classes thématiquement homogènes (propres à un domaine, à un sous-domaine, etc.). Ce type de structure pourrait être utilisé pour des annotations en corpus de plus en plus précises, permettant des affinements successifs de la granularité de l'information, plus nombreux que ceux déjà mis en œuvre.

Cependant, un travail amont sur la ressource lexicographique est nécessaire pour garantir la cohérence et la qualité de ces structures imbriquées, ainsi que leur accessibilité à un traitement informatique. Il existe des écarts entre la nomenclature originelle et les domaines intégrés au moteur de recherche du TLFi. De plus, le moteur de recherche intègre deux niveaux, alors que la nomenclature en propose plus. Ajoutons qu'il serait judicieux de vérifier l'actualité de la nomenclature, des subdivisions et de certaines affectations de domaines aux entrées, pour éviter certains problèmes dus à l'étalement dans le temps de la rédaction et à l'émergence de nouveaux domaines depuis la fin de la rédaction (*cf.* chapitre II.1, 1.1.3).

3.2.2 Graphe en fonction des liens lexicaux et extraction de classes

La structure exploitée dans le cadre expérimental est une structure partitionnée, avec des unités discrètes indépendantes – autrement dit, une structure en "sacs de sèmes". Nous avons montré qu'au niveau de la microstructure, donc au sein des définitions, il serait possible d'intégrer des relations sémantiques. Ces liens sémantiques internes aux entrées peuvent servir à générer des relations entre différentes entrées, c'est-à-dire un graphe de dictionnaire. De tels graphes ont fait l'objet ou font actuellement l'objet de divers projets (Gaume, 2006 ; Loiseau *et al.*, 2010 ; Lux-Pogodalla et Polguère, 2011).

Ces graphes amèneraient à redéfinir le sémème associé à une entrée. Le sémème serait établi à partir d'un continuum. La navigation dans le graphe permettrait d'associer un plus grand nombre de sèmes à la cible et les sèmes pourraient être pondérés en fonction de distances dans le graphe, ou encore de la configuration locale (nombre d'arêtes associées à un nœud, zones de confluence, unité isolée, etc.).

L'existence d'un graphe permettrait de constituer de nouvelles classes de sèmes, en fonction de certains types de liens et en fonction de la densité locale (zones fortement connectées ou faiblement connectées), par extraction de cliques comme celles de (Ploux et Victorri, 1998). Pour un graphe construit à partir de relations qualifiées (lien de synonymie, lien métaphorique, etc.), il serait alors possible d'affecter des ensembles de sèmes de nature différente à une même unité, autrement dit, on pourrait subdiviser le sémème en sous-sémèmes propres à un certain type de relation (sous-sémème correspondant à un réseau synonymique, sous-sémème obtenu à partir de dépendances métaphoriques, etc.).

4. Affiner la représentation des ressources textuelles

4.1 Désambiguïser au préalable pour annoter de façon ciblée

La désambiguïsation des unités lexicales en discours permettrait une annotation sémique plus précise. Par exemple, l'absence de désambiguïsation d'*actifs* dans le corpus sur la crise financière génère du bruit au niveau des domaines saillants : sa forte présence au voisinage de *toxique* a entraîné une saillance des domaines de sa définition, en particulier de CARACTEROLOGIE, qui n'est pas pertinent dans le contexte de la crise financière.

Diverses procédures de désambiguïsation sont possibles (Agirre et Edmonds, 2006). Elles peuvent reposer sur des analyses syntaxiques ou sur l'appariement des unités lexicales à des thèmes ou des domaines. La désambiguïsation permettrait une annotation ciblée, telle que les sèmes non pertinents seraient éliminés. Ainsi, *banque* et *financier* possèdent des définitions dans le domaine du théâtre, l'annotation sans désambiguïsation préalable introduit indûment les sèmes de ces définitions, tels que /saltimbanques/ ou /personnage/. Ce type de bruit devrait pouvoir être filtré en éliminant les sèmes dépendant de domaines absents ou sous-représentés dans les textes. En amont, il serait nécessaire d'établir un parallèle entre les domaines ou thèmes des ressources textuelles et les domaines des ressources lexicographiques.

4.2 Réduire l'éparpillement des données

Le problème d'éparpillement des données a déjà été évoqué pour les sèmes, mais il affecte aussi les unités lexicales en corpus. Un grand nombre d'unités lexicales sont absentes, mais parmi ces unités absentes, certaines auraient pu apparaître de façon légitime, ne serait-ce qu'à travers des mécanismes de paraphrases, de synonymes, etc. Ces absences jouent notamment sur la taille des sous-corpus et, par conséquent, sur les valeurs retournées par des indices statistiques, comme abordé au (chapitre II.2, 2.1.3.d ; l'échelle des coefficients de significativité pourra dans certains cas être plus tassée). De plus, le nombre d'unités différentes influe sur la récurrence sémique ; cette récurrence sémique aura tendance à être renforcée si on introduit des unités qui auraient pu être employées de façon alternative à celles effectivement présentes.

Des techniques telles que les techniques de lissage (*smoothing techniques*) permettraient d'introduire des unités lexicales artificiellement. Ce type de méthodes peut avoir une influence sur les traitements statistiques appliqués au corpus des voisinages. Les techniques de lissage affectent aux unités lexicales des coefficients qui ne sont pas forcément des entiers.

L'utilisation de modèles statistiques discrets comme les spécificités de (Lafon, 1984) devient délicate. L'application de techniques de lissage devrait donc s'accompagner d'une réflexion sur les modèles mathématiques adaptés à la reconfiguration induite.

La pondération obtenue par lissage doit aussi être prise en compte lors de l'annotation sémique : les sèmes sont alors affectés d'un poids qui dépend de celui attribué à l'unité lexicale qu'ils caractérisent. Si comme évoqué précédemment, la représentation sémique est elle aussi pondérée, il devient nécessaire de prendre en compte une double pondération, sur les unités lexicales et sur les sèmes qui définissent ces unités lexicales. La pondération résultante peut être obtenue par multiplication de coefficients, solution assez immédiate, mais le calcul des coefficients mériterait une réflexion plus approfondie.

4.3 Analyser l'influence de différents paramètres

Un certain nombre de paramètres ont été fixés par défaut dans les expériences réalisées. Une voie d'amélioration des résultats serait d'étudier l'influence de ces paramètres sur la qualité des informations obtenues. Il s'agirait de faire varier ces paramètres afin de déterminer où se trouve l'information sémique la plus pertinente et quel réglage des paramètres est optimal.

4.3.1 Influence des indices conscients

Un paramètre à étudier est l'influence des indices conscients de néologies, tels que les guillemets, les expressions introductives ("qualifié de", "dit(e)") ou explicatives ("c'est-à-dire"). Ces indices témoignent d'un sentiment néologique. On peut supposer que le locuteur s'efforcera d'introduire dans le cotexte des éléments de définition du sens ressenti comme anormal, donc que les unités lexicales ou sèmes pertinents seront plus concentrés dans les voisinages comportant des indices conscients. À l'inverse, lorsque le nouveau sens se stabilise et qu'il commence à être intégré à l'unité lexicale, le locuteur ressentira probablement moins le besoin d'expliciter ce qu'est ce nouveau sens. Il faudrait une étude spécifique pour valider cette hypothèse et mesurer l'influence des indices de détection sur les sèmes participant à l'enrichissement sémantique.

4.3.2 Influence de la distance linéaire (syntaxique ou taille du palier de cooccurrence)

Dans les analyses réalisées, le palier de cooccurrence a été fixé au paragraphe. Le choix de ce palier était déterminé par la volonté de se situer à la frontière du global et du local et par une recherche focalisée sur les isotopies. Des paliers plus réduits sont susceptibles d'apporter des informations complémentaires, répondant à une perspective différente de celle retenue, mais qui aurait également son intérêt. Ces informations relèveraient moins de contrastes entre fonds et formes sémantiques que de relations syntactico-sémantiques et de propagation locale de sèmes. Pour accéder à ces informations, le paramètre à faire varier serait la distance linéaire, c'est-à-dire la taille du palier de cooccurrence et une distance définie par un arbre de dépendances syntaxiques.

Pour intégrer les dépendances syntaxiques à l'analyse sémique, il conviendrait de distinguer actants et prédicats, puis d'utiliser les prédicats comme modificateurs du contenu sémique des actants. Analyser l'effet des relations syntactico-sémantiques sur les sémèmes constitue un travail à part entière.

Au niveau des techniques utilisées, le recours à des modèles statistiques serait moins approprié, en raison de la taille réduite du sous-corpus textuel. Les dépendances entre unités orientent vers d'autres types d'outils d'analyse, plutôt associés aux graphes, notamment aux cheminements dans les graphes. Pour construire des graphes ou réseaux, des arbres de dépendance peuvent être obtenus à partir des relations syntaxiques pour des paliers locaux,

c'est-à-dire inférieurs au paragraphe, à partir des relations au sein d'une phrase, mais aussi entre phrases (par exemple, des relations créées par des anaphores).

4.3.3 Influence de la catégorie grammaticale

La question de l'influence de la catégorie grammaticale sur des résultats lexicométriques est une question ouverte. (Mayaffre, 2006) invite à effectuer des traitements statistiques en fonction des catégories grammaticales, sous l'hypothèse que différencier les unités par catégorie permet de faire émerger des informations plus précises. L'étude lexicométrique qu'il réalise semble aller dans le sens de l'hypothèse émise[96].

Dans les traitements effectués, l'élimination des mots-outils et la conservation des unités sémantiquement pleines (noms, verbes, adjectifs et adverbes) établissent une dichotomie de premier niveau. Il conviendrait d'étudier deux questions :

- Si certaines catégories d'unités apportent plus d'information sur le nouveau sens que d'autres. Par exemple, dans l'analyse d'un cotexte d'emploi de *tsunami* au (chapitre I.3, 3.3, figures 3.12 et 3.14), les unités lexicales ont été annotées manuellement en sèmes. Les sèmes destinés à enrichir le sens de *toxique* proviennent presque exclusivement de substantifs. Cette situation peut être accidentelle, elle peut résulter du choix de l'exemple ou d'une analyse manuelle trop superficielle, mais à défaut d'être une quelconque preuve, elle légitime qu'on s'interroge sur la question.
- Si le type d'information sémantique apporté dépend de la catégorie grammaticale. Par exemple, on pourrait étudier si les adjectifs ont plus tendance à exprimer des sèmes génériques que les substantifs, dont l'information est de même nature que celle des domaines, ou encore si les verbes expriment plutôt des différences spécifiques. De même, on pourrait étudier si telle ou telle catégorie grammaticale participe plus à l'enrichissement ou à la reconfiguration du sens codé.

4.3.4 Importance relative de l'activation et de l'enrichissement

Dans les expériences, des analyses distinctes ont été effectuées sur les sèmes selon qu'ils appartenaient ou non au sémème de la cible lexicale. On peut s'interroger sur l'importance relative des sèmes qui proviennent d'une activation d'anciens sèmes et ceux qui proviennent d'un enrichissement. Pour ce faire, il faudrait effectuer des tests avec un paramètre à régler, afin d'évaluer le poids à attribuer aux sèmes activés et aux sèmes d'enrichissement. Une telle démarche permettrait d'obtenir une représentation unifiée du nouveau sens.

4.4 Augmenter le nombre d'affinements successifs

Le protocole proposé procède par affinements successifs avec des observables de granularité de plus en plus fine (domaines, unités lexicales puis sèmes). Le fait de qualifier le contenu sémantique à partir d'affinements successifs a un impact positif sur la qualité des résultats. Utiliser des descripteurs de granularité de plus en plus fine permet de structurer les résultats (en fonction des domaines dans nos applications), d'interpréter plus facilement des unités de granularité fine et de jouer sur la précision de l'information. Ce principe mériterait d'être

[96] (Mayaffre, 2006) applique un calcul de spécificités à un corpus de discours présidentiels de deux façons : toutes catégories confondues et en dissociant les catégories grammaticales. Il souligne que des réalités plus sensibles émergent lorsque les catégories sont dissociées, par exemple au niveau des verbes dans le discours de Giscard : « Ainsi, apparaissent plus clairement des événements lexicaux (ici de nature verbale) qui nous semblent caractériser fortement le discours giscardien et qui risquaient de passer jusqu'ici inaperçus. "Considérer" (désormais 61ème rang, remontée de +14), "observer" (66ème rang, +15) ou "concerner" (69ème rang, +14)) par exemple, illustrent parfaitement le discours didactique du pédagogue Giscard qui se comporte plus en professeur qu'en président »

poussé plus loin, avec des niveaux de granularité intermédiaires, donc des affinements plus nombreux. Ces affinements successifs peuvent être définis en fonction de différents types de sèmes :

- ***Affinements thématiques***. Les affinements thématiques sont dans le sillage de l'annotation en domaines. Pour cela, il faut disposer de regroupements thématiques de plus en plus en précis, issus soit de la ressource lexicographique, soit des ressources textuelles. Au niveau lexicographique, les affinements successifs pourraient être obtenus à partir de la macrostructure du dictionnaire relative aux imbrications de domaines et sous-domaines. En corpus, les unités lexicales seraient annotées successivement en fonction des domaines et sous-domaines, avec une précision croissante. Pour *toxique*, on pourrait avoir comme unités thématiques saillantes ECONOMIE, puis FINANCES (comme sous-domaine d'ECONOMIE). Au niveau de thématiques issues du corpus, une caractérisation pourrait être obtenue par des techniques d'extraction de mots-clés. les affinements seraient définis par des n-uplets de mots-clés de taille croissante (un mot-clé, deux mots-clés, etc. ; par exemple, pour *toxique*, on pourrait avoir *crise*, puis *crise* + *marché*, puis *crise* + *marché* +*subprime*, etc.).
- ***Affinements connotatifs***. Les connotations pourraient être obtenues à partir de sèmes dimensionnels de plus en plus précis (/animé/ – /non animé/, puis, pour /animé/, /humain/ – /non humain/, puis, pour /humain/, /homme/-/femme/, etc.). À l'heure actuelle, les sèmes dimensionnels n'ont pas été identifiés et balisés comme tels dans le TLFi. Cependant, il serait possible de retrouver des équivalents de ces sèmes dimensionnels à travers le formalisme que doit introduire le projet RLF. En particulier, certaines fonctions lexicales pourraient être considérées comme analogues à des sèmes dimensionnels. Par exemple, les sèmes dimensionnels /positif/-/négatif/ pourraient être assimilés aux fonctions lexicales Bon(x) – Antibon(x). Les affinements d'Antibon permettraient de préciser comment se manifeste le caractère positif ou négatif, par exemple à travers l'idée d'ampleur (BonMagn(x) - AntibonMagn(x)). En corpus, l'annotation pourrait alors se limiter aux fonctions lexicales assimilées à des sèmes connotatifs, avec des fonctions de plus en plus précises pour permettre d'affiner les nuances portées par les connotations.
- ***Affinement de l'ensemble des proxèmes.*** L'annotation sémique pourrait être effectuée par familles de proxèmes, tels que définis par (Gaume, 2004). De même que, dans l'application PROX, l'ensemble des proxèmes peut être élagué progressivement, de façon à faire apparaître des sous-ensembles plus précis de proxèmes, de même, une annotation peut d'abord comporter l'ensemble des proxèmes, puis se restreindre à des sous-groupes de plus en plus précis, obtenus par les élagages successifs de PROX. Il serait possible de savoir quel ensemble de synonymes caractérise la cible, puis, au sein de cet ensemble, quel sous-ensemble lui est le plus étroitement associé, et ainsi de suite.

5. Vers un système complet

Nous proposons trois pistes pour faire évoluer le modèle proposé vers un système complet :

- un renforcement théorique du modèle, avec l'articulation du modèle principal à un modèle complémentaire qui décrit plus précisément les relations sémantiques locales ;
- la réalisation d'une plateforme réalisant un traitement intégral ;
- l'adaptation des traitements à une évolution dynamique des ressources textuelles.

5.1 Extension théorique : un modèle complémentaire fondé sur des graphes

Dans le modèle proposé, les méthodes ont une cohérence avec la perspective choisie, focalisée sur l'influence du global sur le local et sur les contrastes entre fonds et formes sémantiques. Ce modèle peut être enrichi par un autre modèle, qui relèverait d'une perspective axée sur des relations plus locales, définies par des dépendances et des liens syntagmatiques. Ce modèle complémentaire peut se décrire par transposition d'un certain nombre de caractéristiques du modèle proposé.

	Modèle proposé	**Modèle complémentaire**
Niveau d'influence dominant (global ou local)	global	local
Relation entre unité lexicale et sème	inclusion	dépendance
Phénomène décrit	contrastes entre fonds et formes sémantiques	propagation de sèmes
Façon de repérer l'information sémantique	saillances (contrastes)	chemins (navigation)
Cadre mathématique	statistiques	graphes

Figure III.2.5 : Modèle complémentaire du modèle proposé pour décrire le comportement des traits sémantiques

Dans notre approche, les croisements et combinaisons étaient destinés à faire émerger des faisceaux d'isotopies dans les nouveaux contextes d'emploi d'une cible lexicale. Cette approche est adaptée pour des contrastes entre le global et le local et pour des phénomènes récurrents, c'est-à-dire pour une diffusion suffisamment importante de nouveaux emplois.

Pour intégrer les phénomènes sémantiques locaux avec plus de finesse et pour tendre vers une diffusion minimale, c'est-à-dire des variations sémantiques ponctuelles, il devient nécessaire de quitter le cadre statistique et de considérer des alternatives au schéma d'urne. Une solution est de substituer aux unités isolées des unités connectées, et de construire des graphes sémantiques. Le passage à une représentation sous forme de graphe a été évoqué dans les points précédents de ce chapitre aussi bien au niveau de la ressource lexicographique (sous-section, 3.2.2.) que des ressources textuelles (sous-section 4.3.2). À partir des deux réseaux, lexicographique et textuel, un double défi est à relever :

- articuler les deux réseaux ;
- en extraire un sous-graphe relatif à une cible lexicale, destiné à représenter son sens.

De même qu'au niveau des cibles lexicales, on avait cherché à élargir le processus à d'autres types de néologies, de même, pour la variation lexicale, la représentation sous forme de graphe ouvrirait des possibilités pour s'affranchir du critère de diffusion, c'est-à-dire de la répétition dans le temps. Les limites du modèle seraient repoussées pour tendre vers une caractérisation précoce du nouveau sens.

Une représentation sous forme de graphe présente un autre intérêt. Cette représentation est susceptible de modéliser plus finement des relations locales, à partir d'un nombre de cotextes plus réduits. Cela peut aussi servir si, au lieu d'essayer d'extraire l'information d'un ensemble de cotextes, on choisit un petit nombre de cotextes jugés représentatifs dont on souhaite extraire le nouveau sens.

5.2 Extension applicative : implémentation intégrale du modèle

Sur le plan théorique, un système cohérent a été proposé, avec un enchaînement d'étapes précis. Sur le plan applicatif, la mise en œuvre reste incomplète et fragmentée pour trois raisons :

- les modules de traitement s'appuient sur des outils différents, qui ne sont pas intégrés dans un tout cohérent ;
- la validation ne repose pas sur une expérience filée du début à la fin du processus, mais sur une série d'expériences illustratives, appliquées à des cibles et des corpus différents ;
- la portée des cas d'étude est limitée, elle reste celle de tests ponctuels relatifs aux différentes étapes. Pour donner une portée plus large, adaptée à la perspective de veille lexicale et d'acquisition sémantique réactive, il faut prévoir des adaptations du modèle pour pouvoir appliquer la procédure à un corpus dynamique.

5.2.1 Intégrer les modules dans une plateforme globale

La chaîne de traitements que nous avons proposée n'est pas implémentée dans une plateforme qui effectue l'ensemble des traitements. Les différentes étapes ont été traitées à partir de moyens ou outils fragmentés, soit par des plateformes ou logiciels (Lexico3 (Salem *et al.*, 2003), Semy (Grzesitchak, 2008), PermutMatrix (Caraux et Pinloche, 2005)), soit par des approches semi-automatiques parcellaires et fortement contrôlées (lignes de commandes isolées ; programmes spécifiques adaptés à des tests), soit par une analyse manuelle. Le tableau III.2.6 (*cf.* infra) récapitule les différentes étapes et le type de traitement utilisé.

Pour une mise en œuvre effective, il faudrait réaliser une plateforme qui automatise les traitements qui correspondent à des approches-tests et qui intègre les différents logiciels dans un tout cohérent. Un point plus délicat est d'identifier ou d'élaborer des traitements qui génèrent des sorties analogues aux résultats des analyses manuelles.

Étape	Observable ou paramètre d'étude		Ajout d'information	Traitements
I	Choix de la période de temps et de la taille de cotexte		Balisage en temps Découpage du corpus (f(temps)) **[ProcSpé]**	Détection de cibles **[Manu]**
II	Cible lexicale		Balisage des voisinages Découpages : f(temps, voisinage) **[ProcSpé]**	
III.A	Alternatives	Domaines	Annotation en domaines issus du dictionnaire **[SEMY]**	1. Calcul de spécificités **[LEXICO3]** 2. Distinction entre domaines associés au sens codé de la cible (a) / autres domaines (b) **[ProcSpé]** 3.a) Tri entre domaines activés et inhibés **[LEXICO3 + ProcSpé]** 3.b) Extraction de domaines saillants **[LEXICO3 + ProcSpé]**
III.B			Annotation en domaines du corpus (domaines donnés) **[ProcSpé]**	1. Calcul de spécificités **[SEMY]** 2. Extraction de domaines saillants **[SEMY + ProcSpé]** 3. Tri entre domaines nouveaux et domaines associés au sens codé de la cible **[Manu]**
IV			Balisage en domaines Découpages : f(temps, voisinage, domaines) **[ProcSpé]**	
V	Unités lexicales			1. Calcul de spécificités **[LEXICO3]** 2. Extraction d'unités lexicales saillantes **[LEXICO3 + ProcSpé]**
VI.A	Sèmes		Distinction entre sèmes du sens codé de la cible (activation) et autres sèmes (enrichissement) **[ProcSpé]**	
VI.A.1	Complémentaire	Sèmes d'activation	Annotation du corpus en sèmes d'activation **[SEMY + filtre ProcSpé]**	1. Calcul des spécificités **[LEXICO3]** 2. Tri activation / inhibition (saillance positive, pas de saillance, saillance négative) **[ProcSpé]**
VI.A.2		Sèmes d'enrichissement	Annotation du corpus en sèmes **[SEMY]** après neutralisation de la cible **[ProcSpé]**	1. Calcul des spécificités **[LEXICO3]** 2. Extraction de sèmes saillants **[LEXICO3 + Manu]** 3. Filtre : recoupement avec les unités lexicales saillantes **[LEXICO3 + ProcSpé]** 4. Classes **[Manu]**
VI.B (Alter-native à VI.A)	Sèmes		Liste d'unités lexicales saillantes	Dans le dictionnaire : 1. Constitution d'un réseau entrées (unités lexicales) – sèmes (matrice) **[SEMY + ProcSpé]** 2. Extraction de classes **[PERMUTMATRIX + Manu]**
Légende. [ProcSpé] : procédure spécifique ; [Manu] : analyse manuelle				

Tableau III.2.6 : Outils d'analyse utilisés aux différentes étapes de la procédure

5.2.2 Des adaptations à prévoir pour intégrer la dynamique temporelle

L'évolution de sens est un phénomène continu dans le temps. Les traitements réalisés s'appliquent à des corpus découpés en tranches de temps. Il est possible d'évoluer d'un temps discret vers un temps continu par des discrétisations de plus en plus fines, mais un découpage trop fin du corpus risque de poser des problèmes pour les traitements statistiques : plus le sous-corpus d'étude est réduit, plus on s'approche des limites d'application des modèles

statistiques. Pour s'adapter à une incrémentation de corpus pratiquement continue, il faudrait mettre en place des techniques d'analyse de flux continus (*cf.* (Allan *et al.*, 1998), ou plus récemment (Forestier et Velcin, 2009), par exemple) et y adapter les traitements.

5.3 Mise en place d'une validation robuste

5.3.1 Réaliser une expérience de validation complète

Les expériences réalisées pour valider les différentes étapes du traitement sont elles aussi fragmentées. La fragmentation est due à plusieurs facteurs :

- La démarche pour élaborer le processus était inductive. La chaîne de traitements a été définie à partir des apports de chaque expérience. Ces expériences ont permis de faire émerger des parties du traitement, l'intégration de leurs apports dans un système cohérent n'a été effectuée que dans un second temps.
- Les corpus étaient variés, avec des structures propres à différents axes d'observation. La constitution de plusieurs corpus répondait à divers objectifs : profilage domanial (corpus 'Factiva'), évolution temporelle (corpus 'Outreau') ou focalisation sur les contrastes entre local et global à domaine et période fixés (corpus 'Crise financière').
- Plusieurs cibles ont été étudiées, chacune sous un angle qui lui était propre. De ce fait, nous n'avons pas utilisé un exemple filé, ni un panel de cibles qu'il aurait été possible de comparer pour chaque type de traitement.

La validation est donc marquée par un manque de systématicité. De plus, cette validation a été essentiellement manuelle.

Pour consolider la validité de la procédure, il faudrait mettre en place une expérience de validation à plus grande échelle, qui reposerait sur l'observation d'un ensemble de cibles présentes à chaque étape, et ce dans un seul corpus. L'expérience pourrait s'appuyer sur les éléments suivants :

- La sélection d'un ensemble de cibles de référence identifiées à partir de travaux lexicographiques externes. On pourrait par exemple s'appuyer sur la liste des mots pour lesquels un nouveau sens a été proposé dans le *Nouveau Petit Robert* ou le *Petit Larousse*. (Martinez, 2011) a relevé une liste de mots dont le sens a été amendé dans les deux dictionnaires depuis 2008. Il resterait à vérifier la validité de ces sources (par exemple, pour contrôler s'il ne s'agit pas d'un sens éliminé quelques années auparavant et réintroduit dans la foulée).
- La constitution d'un corpus où les observer. Ce corpus doit comporter une variété de domaines (avec notamment la présence des domaines rattachés aux anciens et aux nouveaux sens des cibles identifiées), avoir une étendue en temps adaptée aux néologismes observés (de 2000 à 2010 par exemple, si l'on part du principe que la durée d'implantation d'un néologisme est d'environ de 5 ans, même si cette durée n'est pas systématique (*cf.* chapitre I.2, 1.3.3)) et être doté d'un balisage qui permette d'identifier les paliers de cooccurrence qui définissent les voisinages (paragraphes par exemple).
- Une validation à partir de sémèmes extraits des dictionnaires. Une fois les anciens et nouveaux sens récupérés à partir de ressources telles que le *Nouveau Petit Robert* ou le *Petit Larousse*, il faudrait extraire des sémèmes avant changement de sens et après changement de sens. Le résultat des traitements peut alors être validé à partir de ces sémèmes correspondant à l'ancien et au nouveau sens.

5.3.2 Intégrer des techniques de validation mathématiques à la validation manuelle

Les méthodes de validation utilisées dans le cadre des expériences étaient essentiellement manuelles. Pour consolider la validation, il conviendrait d'introduire un autre type de validation. On peut notamment s'appuyer sur des techniques mathématiques de validation pour compléter et renforcer les techniques de validation manuelle. Une de ces techniques mathématiques est présentée brièvement.

D'un point de vue statistique, les résultats dépendent de l'échantillon sur lequel les traitements sont effectués. Le statut d'échantillon affecte tout autant le corpus que ses sous-corpus (ensemble des cotextes d'emploi de la cible lexicale) . Des techniques permettent de tester la stabilité des résultats. Dans le cas de l'AFC, des techniques de rééchantillonnage permettent de valider les structures qui se dégagent des représentations (Lebart *et al.*, 2003:58). Ces techniques dites de *bootstrap* testent la stabilité des résultats et appartiennent aux techniques de validation interne, pour lesquelles il n'est pas besoin d'introduire de nouveaux éléments de contenu informationnel. Le bootstrap consiste à générer de nouveaux échantillons de même taille que l'échantillon initial, obtenus à partir de tirages avec remise au sein de l'échantillon initial. Le bootstrap est total lorsqu'à chaque nouvel échantillon, une nouvelle représentation est produite avec des axes de projection susceptibles de varier à chaque fois. Le bootstrap est partiel lorsqu'on conserve la structure obtenue à partir de l'échantillon initial (conservation des axes en particulier) et qu'on projette la structure des nouveaux échantillons dans l'espace défini par la structure initiale. Ceci permet de définir des ellipses de confiance, *i.e.* des zones de variation potentielle de chaque point. Cette méthode de *bootstrap partiel* est particulièrement efficace pour vérifier si les résultats sont stables (Lebart *et al*., 2006).

Dans le cadre de l'allocation de signifié, ce type de technique permettrait de vérifier si le nouveau sens qui se dégage des résultats est fiable. Il pourrait aussi être utilisé pour tester la stabilité dans le temps du nouveau sens : un échantillonnage sur différentes tranches de temps montrera probablement une instabilité en début de diffusion, par exemple lors d'un pic événementiel, puis une stabilisation progressive au cours des périodes ultérieures.

Conclusion

Nous avons montré que, lors de l'émergence d'un nouveau sens en discours, il est possible d'acquérir de façon semi-automatique une représentation nuancée du nouveau sens et de l'articuler au sens codé dans une ressource de référence. Pour cela, nous nous sommes appuyée sur une modélisation de la néologie sémantique inspirée de principes de la sémantique textuelle et centrée sur une perspective applicative. Nous avons défini les grandes lignes d'une procédure d'allocation de signifié, qui ont été étayées par des expériences illustratives.

Dans une première partie, nous avons proposé un modèle théorique pour allouer un nouveau sens à une unité lexicale ayant un sens codé dans une ressource de référence.

Notre objet d'étude a été défini à l'intersection de deux phénomènes : la variation sémantique et la néologie.

Dans le chapitre I.1, nous avons abordé la variation sémantique comme un écart entre le sens codé dans une ressource de référence et le sens en contexte. La variation sémantique a été décrite comme un phénomène graduel, plus ou moins marqué. Nous avons choisi de nous focaliser sur les variations marquées, pour lesquelles les ressources lexicographiques ne fournissent pas les éléments d'interprétation nécessaires et suffisants. Ces variations sont associées à un sentiment de rupture et elles s'accompagnent de traces discursives, ce qui les rend accessibles à des traitements automatiques. Parmi les variations marquées, nous nous sommes limitée à celles qui présentent une certaine diffusion et qui, de ce fait, participent à la stabilisation d'un nouveau sens. Les variations sémantiques marquées en cours de diffusion rejoignent le champ de la néologie et elles sont assimilées à de la néologie sémantique.

Dans le chapitre I.2, nous avons positionné la néologie sémantique dans le champ de la néologie. Derrière la tripartition classique entre néologie de forme, néologie de sens et emprunt se cache une réalité plus complexe, où les frontières entre les différents types de néologie sont floues. Dans notre cadre, nous considérons qu'il y a néologie sémantique si, pour un signifiant donné, il existe un sens codé dans une ressource de référence. Les autres types de néologie ont été agencés par rapport à la néologie sémantique selon les procédés qui les génèrent. Le processus d'allocation de signifié décrit par la suite est élaboré pour la néologie sémantique, il est destiné à faire ressortir l'apport du contexte au sens. Nous avons décrit comment l'étendre aux autres formes de néologie, moyennant des traitements complémentaires. Ceux-ci sont destinés à générer un contenu sémantique analogue au sens codé à l'aide de ressources externes ou d'outils tels que des analyseurs morphologiques.

Au chapitre I.3, nous avons défini des éléments de modélisation de la néosémie, adaptés à une approche applicative et intégrés à un système théorique cohérent. Dans le cycle d'évolution du sens, nous ciblons un stade de diffusion relativement précoce sans pour autant relever de l'événement ponctuel. La diffusion se traduit par des événements répétés et elle oriente vers des indices quantitatifs. Certains indices quantitatifs contribuent essentiellement à

la détection de la néosémie, d'autres participent aussi à la qualification du nouveau sens. La priorité est donnée aux indices qualifiants, à savoir les cooccurrents d'ordre 1 de l'unité lexicale ciblée, le paradigme des concurrents issus du foisonnement néologique et les empreintes de fréquences spécifiques à des thèmes ou des domaines. L'émergence d'un nouveau sens peut donc être appréhendée à travers plusieurs axes d'observation, témoins d'un phénomène multiniveau. Pour intégrer ce dernier dans un système cohérent, nous adoptons un cadre théorique issu de la sémantique textuelle. Le sens codé y est représenté sous forme d'un ensemble structuré de traits sémantiques, les sèmes. En contexte, le sens émerge à travers des phénomènes de récurrences simples ou combinées de sèmes (isotopies ou faisceaux d'isotopies). À travers ces phénomènes, le contexte peut reconfigurer le sens codé par activation ou inhibition de sèmes, ou l'enrichir en nouveaux sèmes. La dynamique du sens est abordée à travers des descripteurs situés à plusieurs niveaux, dépendants les uns des autres : au niveau supra-lexical, avec des descripteurs du sens relativement généraux (thèmes, domaines) , au niveau lexical (cooccurrents d'ordre 1 et 2), et au niveau infra-lexical, avec des descripteurs du contenu sémantique d'unités lexicales (les sèmes ou traits sémantiques) . À travers ces différents niveaux de description, nous jouons à la fois sur l'information sémantique accessible directement (signifiants du niveau lexical) et sur l'information diffuse (niveaux infra-lexical, supra-lexical).

En deuxième partie, nous avons étudié des ressources et outils adaptés à une procédure d'allocation de signifié.

Au chapitre II.1, nous avons présenté et analysé les ressources retenues pour la partie applicative. Des ressources de deux types sont nécessaires : des ressources lexicographiques pour le sens codé et des ressources textuelles pour le sens en discours. La ressource choisie comme vivier de sens codés est le *Trésor de la Langue Française informatisé* (TLFi), qui valide un certain nombre de caractéristiques jugées importantes en terme d'accessibilité, de langue et de couverture du lexique. Le formalisme des entrées a été transformé pour pouvoir être combiné aux données textuelles et pour être conforme au modèle théorique. La plateforme Semy convertit les entrées en sacs de traits sémantiques, c'est-à-dire des ensembles de traits issus des constituants sémantiquement pleins des définitions et des domaines de rattachement. La ressource et le format de représentation choisis restent imparfaits, du fait des dates de rédaction du TLFi, de l'éparpillement du vocabulaire et de la faible structuration des sens codés après transformation des entrées. Bien que pénalisantes, ces limites n'ont pas été jugées suffisantes pour compromettre la procédure. La connaissance de ces limites a été intégrée à l'analyse pour expliquer les résultats et reconsidérer la validité de certaines pistes. Les ressources pour appréhender le sens en discours sont des corpus, caractérisés par une structure temporelle, spatiale (en fonction du voisinage de la cible) et thématique (présence de domaines). Trois corpus de presse ont été utilisés en pratique, pour aborder les phénomènes sous différents angles : le corpus 'Factiva', étendu dans le temps et présentant une diversité de domaines, pour étudier des phénomènes relevant du niveau supra-lexical (domaines) ; le corpus thématique 'Crise financière' pour observer les phénomènes sémantiques à un niveau de granularité plus fin (unités lexicales, sèmes), à travers des contrastes dans le temps ou entre voisinages local et global ; le corpus 'Outreau' pour observer la progression dans le temps de la diffusion d'un nouveau sens.

Au chapitre II.2, nous avons proposé des outils mathématiques pour construire le nouveau sens. Ces outils ont une triple finalité : faire apparaître des informations de nature différente en fonction de la structure interne ou conjointe des ressources ; extraire des unités saillantes sémantiquement pertinentes ; structurer ces unités. Pour ce faire, les ressources lexicographiques et textuelles sont considérées comme des espaces mathématiques, dont la

configuration varie en fonction de divers paramètres (temps, domaines, palier de textualité et ordre du voisinage de cooccurrence en corpus ; structure interne aux entrées et externe aux entrées dans le dictionnaire). Les diverses configurations sont converties en schémas d'urne, à la base du calcul d'indices statistiques. Ces indices servent à extraire des unités saillantes au voisinage de l'unité lexicale ciblée et à les structurer. Après comparaison d'une série d'indices répandus en textométrie, notre choix s'est porté sur un indice basé sur la loi hypergéométrique, les spécificités de Lafon. Nous avons donné un aperçu des types de structures que permettent d'obtenir les pondérations affectées aux unités via les spécificités, l'interaction des deux espaces, lexicographique et textuel, leurs découpages multiples et la diversité des observables (domaines, unités lexicales et sèmes).

En troisième partie, nous avons mis en place les grandes lignes d'un modèle applicatif, en accord avec le cadre théorique et fondé sur les ressources et outils choisis.

Le chapitre III.1 esquisse le déroulement d'une procédure étayée par des expériences illustratives, construites autour de cibles et corpus adaptés à chaque étape. L'acquisition du nouveau sens s'effectue par strates. La granularité sémantique est la clé d'organisation principale de la procédure : l'information sémantique est recherchée par affinements successifs de ce qu'on observe (domaines, unités lexicales, puis sèmes), avec dépendance d'un niveau d'observation au précédent. À chaque étape, les ressources ont été analysées à travers différents axes : des contrastes entre local et global (voisinage de la cible lexicale et reste du corpus), des contrastes entre domaines et des contrastes dans le temps. L'analyse en domaines a permis de qualifier l'émergence d'un nouveau sens de façon précise et structurée, aussi bien au niveau de la structuration des domaines que dans leur articulation avec le sens codé. L'analyse des unités lexicales a permis de préciser l'information apportée par les domaines. Les informations lexicales saillantes ont été reliées au sens codé manuellement et elles sont apparues comme des vecteurs d'un contenu sémantique diffus. L'analyse en traits sémantiques a permis d'expliciter ce contenu diffus et le lien avec le sens codé, en termes de reconfiguration et d'enrichissement. Les résultats au niveau des traits sémantiques étaient parfois discutables et assez fortement bruités. Des filtres en fonction des domaines et des résultats sur le plan lexical ont toutefois amélioré la qualité de l'information. De façon générale, cette démarche a permis de récupérer des informations nuancées et structurées sur le nouveau sens.

Au chapitre III.2, nous avons défini des pistes pour consolider les faiblesses de la procédure et l'enrichir. Des propositions ont été faites pour améliorer la représentation des sens codés, afin de réduire le problème de dispersion des données et certaines imprécisions du format de représentation choisi. Nous avons également dégagé des voies d'exploration sur d'autres façons d'exploiter le dictionnaire, sur l'extension de la procédure à d'autres objets d'étude et sur un affinement des traitements afin de parvenir à une représentation du sens encore plus nuancée.

La présente thèse n'avait pas pour ambition de résoudre complètement la question de l'allocation automatique d'un nouveau sens, objectif peu réaliste, mais elle a posé les jalons d'un protocole cohérent, structuré, avec un ancrage théorique solide, une proposition de mise en œuvre réaliste et vraisemblablement automatisable d'après les tests expérimentaux, et qui est enfin ouvert à des extensions multiples.

Conclusion

Bibliographie

AGIRRE E., EDMONDS P. (2006). *Word Sense Disambiguation: Algorithms and Applications* (Text, Speech and Language Technology). Springer, Secaucus, NJ, USA.

AHMAD, K., SCHIERZ A. & AL-THUBAITY A. (2002). « Discovery and Terminology », Actes de la conférence internationale *Terminology and Knowledge Engineering (TKE 2002)*, Nancy, France, 28-30 août 2002, pp.1-6.

ALAOUI Khalid (2008). « La néologie chez Larousse. Traitement et analyse d'un corpus de néologismes », *Néologie et terminologie dans les dictionnaires*, Jean-François Sablayrolles (dir.), Honoré Champion, Paris, pp.61-84.

ALLAN J., CARBONELL J., DODDINGTON G., YAMRON J., and YANG Y. (1998). « Topic detection and tracking pilot study final report. », *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pp.194-218.

ANSCOMBRE Jean-Claude (1998). « Regards sur la sémantique française contemporaine », *Langages*, 32e année, n°129, 1998. Diversité de la (des) science(s) du langage aujourd'hui.

ANTHONY L. (2005). « AntConc: A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit », *Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning.*

ASTRID Guillaume (2010) « Diachronie et synchronie : passerelles (étymo)logiques », *Texto !* [En ligne], URL : http://www.revue-texto.net/index.php?id=2557.

AUGER Pierre (2010). « Néologicité et extraction automatique de néologismes », in *Actes del i Congrés internacional de neologia de les llengües romàniques (CINEO 2008),* Barcelone, Espagne, 7-10 mai 2008, pp.117-121.

BAIDER Fabienne (2007). « Féminisation des noms de métiers. Une grande victoire ou une petite concession ? », *Proceedings of the 31st Congress in Functional Linguistics* (SILF2007), Lugo, Spain.

BALLABRIGA Michel (2005). *Sémantique textuelle 2.* Revue *Texto !,* mars 2005 [en ligne]. Disponible sur : <http://www.revue-texto.net/Reperes/Cours/Ballabriga2/index.html>. (Consultée le 12 avril 2011).

BARQUE Lucie, NASR Alexis, POLGUERE Alain (2010). « From the Definitions of the Trésor de la Langue Française to a Semantic Database of the French Language », *Proceedings of the 14th EURALEX International Congress*, Leeuwarden.

BARQUE Lucie, POLGUERE Alain (2009). « Structuration et balisage sémantique des définitions du Trésor de la Langue Française informatisé (TLFi) », *Fourth International Conference on Meaning-Text Theory*. Montréal.

BASTUJI J. (1974). « Aspects de la néologie sémantique », *Langages*, n°36, 1974.

BAUER Laurie, RENOUF Antoinette (2000). « Contextual clues to Word-Meaning », *International Journal of corpus Linguistics*, ISSN 1384-6655, Vol. 5, Nº 2, 2000, pp.231-259.

BECUE-BERTAUT M. (2003). « Comparaison des structures induites sur un ensemble de réponses ouvertes par le choix de l'unité statistique. », *Corpus n°2* (La distance intertextuelle).

BEJOINT Henri (2009). « Lexicographie et linguistique : le domaine anglais », *Lexique, 19, Changer les dictionnaires*, Presses Universitaires du Septentrion, pp.117-158.

BEN HARIZ OUENNICHE Soundous (2009). « Diminuer les fluctuations du sentiment néologique », *Neologica*, 3, pp.37-51.

BERNARD Pascale, MONTEMONT Véronique (2010). « Voyage au cœur du langage : le *Trésor de la Langue française* et *Frantext* », *Culture et recherche*, n°124, Hiver 2010-2011, 64p.

BERNET Charles, PIERREL Jean-Marie (2005). « Histoire de Frantext : constitution d'une base textuelle (1964-2002) et perspectives », in *L'édition électronique en littérature et dictionnairique : évaluation et bilan*, J.C. Arnould (eds), Presses Universitaires de Rouen, Éditions Champion, 2005.

BEUST Pierre (2002). « Un outil de coloriage de corpus pour la représentation de thèmes. », Actes des JADT2002, 13-15 mars 2002, Saint-Malo, pp.161-172.

BEUST Pierre (2007). « L'approche interactionniste en traitement automatique des langues : une discontinuité épistémologique ? », *Journées de Rochebrune (Rencontres interdisciplinaires sur les systèmes complexes naturels et artificiels) "Catastrophes, discontinuités, ruptures, limites, frontières"*, Megève, France. ISSN 1242-5125 ENST S (Paris), p. 69-80.

BOUQUET Simon (1998) : « Linguistique textuelle, jeux de langage et sémantique du genre », *Langages* n°129, *Diversité de la (des) science(s) du langage aujourd'hui*, S.Bouquet (ed), pp.112-124.

BOUSSIDAN Armelle, LUPONE Sylvain, PLOUX Sabine (2009). « La malbouffe : un cas de néologie et de glissement sémantique fulgurants. », *Atelier "Du thème au terme, émergence et lexicalisation des connaissances", 8e conférence internationale Terminologie et Intelligence Artificielle*, Toulouse, 20 novembre 2009.

BOUSSIDAN Armelle, PLOUX Sabine (2011). « Using topic Salience and Connotational Drifts to Detect Candidates to Semantic Change », *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011),* Oxford, IK. http://www.aclweb.org/anthology/W/W11/W11-0134.pdf.

BOUSSIDAN Armelle, RENON Anne-Lyse, FRANCO Charlotte, LUPONE Sylvain, PLOUX Sabine (2010). « Vers une méthode de visualisation graphique de la diachronie des néologies. », *colloque Néologie sémantique et Corpus,* Tübingen, Germany. À paraître.

BUZAN Tony, BUZAN Barry (2000). *The mind map book.* London:BBC Worldwide.

BRUNET Étienne (1984). « Le viol de l'urne », *La recherche française par ordinateur en langue et littérature,* Slatkine-Champion, Genève-Paris, 1984, pp.253-264.

BRUNET Étienne (2000). « Qui lemmatise dilemme attise », *Scolia, 11ème rencontres linguistiques en pays rhénan*, 13, pp.7-32.

BRUNET Étienne (2006). « Le corpus conçu comme une boule », Rastier, François, Ballabriga, Michel (dir.). *Corpus en Lettres et Sciences sociales : des documents numériques à l'interprétation, Actes du colloque international d'Albi, juillet 2006*. Publiés par Carine Duteil et Baptiste Foulquié. Paris : Texto, 2006. ISSN 1773-0120.

BRUNET Étienne (2007). « Fréquences et séquences : mise en œuvre dans Hyperbase », *Lexicometrica* 7, *Topographie et topologie textuelles*.

BRUNET Étienne (2011). *Hyperbase, Logiciel hypertexte pour le traitement documentaire et statistique de corpus textuels, Manuel de référence*, version 8.0 et 9.0, janvier 2011.

BRUNET Étienne (1996). « Les liens hypertexuels ou Abondance des liens ne nuit pas », *Ce qui compte, écrits choisis, tome 2, Méthodes statistiques*, Céline Poudat (éd), édité en 2001 pp.143-163.

Repris des Actes du colloque *Lexicographie et informatique,* Didier érudition, Paris, 1996, pp.299-318.

CABRE M. T., DOMENECH M., ESTOPA R., FREIXA J., SOLE E. (2003). « L´Observatoire de néologie: conception, méthodologie, résultats et nouveaux travaux. », *L´innovation lexicale*, Paris: Honoré Champion, pp.125-147.

CARAUX Gilles, PINLOCHE Sylvie (2005). « Permutmatrix : A Graphical Environment to Arrange Gene Expression Profiles in Optimal Linear Order, », *Bioinformatics*, 21, pp.1280-1281. Logiciel accessible en ligne à l'adresse http://www.lirmm.fr/~caraux/PermutMatrix/.

CARTIER Emmanuel, SABLAYROLLES Jean-François (2008). « Néologismes, dictionnaire et informatique », *Cahiers de lexicologie : Revue internationale de lexicologie et lexicographie*, ISSN 0007-9871, Nº 93, 2008, pp.175-192.

CHANGEUX Jean-Pierre (1972). « Le cerveau et l'événement », *Communications*, n°18, pp.37-47.

CHOMSKY Noam (1965). *Aspects of the theory of syntax*, MIT Press (trad.. fr. Millner, J.C.), Paris, Le Seuil, 1971.

CHURCH Kenneth, GALE William (2000). « Empirical estimates of adaptation: The chance of two Noriegas is closer to p/2 than p2 », *Proceedings of COLING 2000*, Saarbrücken, Germany, pp.173–179.

CHURCH Kenneth, HANKS Patrick (1989). « Word Association Norms, Mutual Information, and Lexicography », *Association for Computational Linguistics (ACL Proceedings*), Vancouver, Canada.

COHEN Jacob, COHEN Patricia, WEST Stephen G., AIKEN LEONA S. (2003). *Applied multiple régression/correlation analysis for the behavioral sciences.* 3e édition, édition revue de 1975.

CONDAMINES Anne, REBEYROLLE Josette, SOUBEILLE Anny (2004). « Variation de la Terminologie dans le Temps : une Méthode Linguistique pour Mesurer l'Évolution de la Connaissance en Corpus », Actes *Euralex International congress*, Université de Lorient, France, 6-10 juillet 2004, pp.547-557.

CORDIER Françoise (1996). « De l'intérêt d'une technique d'amorçage sémantique dans l'étude des relations lexicales et conceptuelles », *Questions de méthode et de délimitation en sémantique lexicale*, Hiltraud DUPUY-ENGELHARDT (pub), Presses Universitaires de Reims, pp.13-23.

COSERIU Eugenio (1973). *Synchronie, Diachronie et Histoire* (*Sincronía, diacronía e historia)*, seconde édition revue et élargie par l'auteur, Madrid : Gredos, 1973. Traduit en français en 2006, « Synchronie, diachronie et histoire. Chapitre I, L'apparente aporie du changement linguistique. Langue abstraite et projection synchronique. », *Texto!* [en ligne], mars 2006, vol. XI, n°1, disponible sur : <http://www.revue-texto.net/Saussure/Sur_Saussure/Coseriu_Diachronie1.html>. (Consultée le 17 novembre 2010).

Traduction de la seconde édition de 1973 de l'ouvrage d'Eugenio Coseriu, *Synchronie, Diachronie et Histoire* (*Sincronía, diacronía e historia*)

CUSIN-BERCHE Fabienne (2003). *Les mots et leurs contextes*. Presses Sorbonne Nouvelle.

DAILLE Béatrice (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. Thèse de doctorat, Université Paris 7.

DAILLE Béatrice (1994b). « Study and implementation of combined techniques for automatic extraction of terminology. The Balancing Act : Combining Symbolic and Statistical Approaches to Language », Proceedings of the *"Workshop of the 32nd Annual Meeting of the ACL"*, Las Cruces, New Mexico, USA.

DAOUST François (2011). *SATO 4, Manuel de référence*, Centre ATO, UQAM, Montréal. http://www.ling.uqam.ca/sato/satoman-fr.html.

DASSI Etienne (2003). « Question de sémantique. De la néologie autour de la téléphonie au Cameroun », *Sudlangues, n°2, juin 2003*. Accessible en ligne : http://www.sudlangues.sn/spip.php?rubrique23.

DELANOË Alexandre (2010). « Statistique textuelle et séries chronologiques sur un corpus de presse écrite. Le cas de la mise en application du principe de précaution », in Actes des *10e Journées d'Analyse statistique des Données Textuelles (JADT 2010)*, 9-11 juin 2010, Rome, pp.561-572.

DENDIEN Jacques, PIERREL Jean-Marie (2003) « Le trésor de la langue française informatisé. Un exemple d'informatisation d'un dictionnaire de langue de référence », *TA*L, 44-2, 11-37.

DETRIE Catherine (2011). « La niche et le bouclier : rôle du processus métaphorique dans la construction d'une désignation sociale et événementielle », colloque Langage, discours, événements, 31 mars, 1-2 avril 2011, Villa Finaly, Florence, accessible à http://syled.univ-paris3.fr/colloques/langage-discours-evenements-2011/pages/39.html.

Dictionnaire historique de la langue française (1992). éd. du Robert, sous la direction d'Alain Rey, Paris.

DROUIN Patrick (2003). « Acquisition de termes simples fondée sur les pivots lexicaux spécialisés », *Conférence TIA-2003,* Strasbourg, 31 mars et 1er avril 2003.

DROUIN Patrick, PAQUIN Annie, MENARD Nathan (2006). « Extraction semi-automatique des néologismes dans la terminologie du terrorisme », in Actes des *8e Journées d'Analyse statistique des Données Textuelles (JADT 2006)*, 19-21 avril 2006, Besançon.

DUBOIS Jean (1969). « Énoncé et énonciation », *Langages*, 4e année, n°13, 1969, L'analyse du discours, pp.100-110.

DUBOIS Jean, GIACOMO Mathée, GUESPIN Louis, MARCELLESI Christiane, MARCELLESI Jean-Baptiste, MEVEL Jean-Pierre (1994). *Dictionnaire de linguistique et des sciences du langage*. Larousse, Paris, France.

DURY Pascaline (2008). « Les noms du pétrole : une approche diachronique de la métonymie onomastique », Lexis, E-Journal in English Lexicology, http://screcherche.univ-lyon3.fr/lexis/.

DUNNING Ted (1993). « Accurate Methods for the Statistics of Surprise and Coincidence », *Computational Linguistics*, 19(1), pp.61-74.

DUTEIL-MOUGEL Carine (2004). « Introduction à la sémantique interprétative. », *Texto !* décembre 2004 [en ligne]. Disponible sur : <http://www.revue-texto.net/Reperes/Themes/Duteil/Duteil_Intro.html>. (Consultée le …/10/2010).

DUVIGNAU K., GAUME B., NESPOULOUS J.-L. (2004). « Proximité sémantique et stratégies palliatives chez le jeune enfant et l'aphasique. » Handicap langagier et recherches cognitives: apports mutuels [numéro spécial]. *Revue Parole*, *31-32*, in J.-L. Nespoulous & J. Virbel (Eds.), pp.219-255.

ENJALBERT Patrice (2005). *Sémantique et traitement automatique du langage naturel*. Traité IC2, Cognition et traitement de l'information, Hermès – Lavoisier, 410 p.

EQUIPE TXM (2011). « Présentation – FAQ », *site du projet Textométrie*, http://textometrie.ens-lyon.fr, consulté le 15 septembre 2011.

EVERT Stefan (2005). *The Statistics of Word Cooccurrences : Word Pairs and Collocations*. Thèse de doctorat, Université de Stuttgart.

FERRARESI A., BERNARDINI S., PICCI G., BARONI M. (2010). « Web Corpora for Bilingual Lexicography: A Pilot Study of English/French Collocation Extraction and Translation ». In Xiao, R. (ed.) *Using Corpora in Contrastive and Translation Studies*. Newcastle: Cambridge Scholars Publishing.

FERRARI Stéphane (2006). « Rhétorique et compréhension », *Compréhension des langues et interaction,* Gérard Sabah (dir.), Hermès Lavoisier, ISBN 2-7462-1256-0, pp.195-224.

FERRARI Stéphane (1997). *Méthode et outils informatiques pour le traitement des métaphores dans les documents écrits*. Thèse de doctorat en Informatique de l'Université de Paris XI.

FIRTH John Rupert (1957). *Papers in Linguistics 1934-1951*. London: Oxford University Press.

FLECHON Geneviève (1998). « Expérience de rédaction : la mise au point de quelques rubriques synchroniques dans le *Trésor de la Langue Française*, 1ère Partie », *International Journal of Lexicography*, Volume 11, n°2, juin 1998, pp.87-110.

FORESTIER Mathilde, VELCIN Julien, GANASCIA Jean-Gabriel (2009). « Un cadre formel pour la veille numérique sur la presse en ligne. », *Atelier Veille Numérique (EGC-VN 09)*, Strasbourg, Janvier 2009.

FUCHS Catherine (2008). « L'incertitude interprétative dans l'activité de langage », *Actes de Savoirs*, 5, revue de l'IUF, Paris:PUF, pp.41-57.

GADET Françoise (2003). *La variation sociale en français*. Paris : Ophrys.

GARDIN B., MARCELLESI Ch., MORTUREUX M.-F., LEFEVRE G. (1974). « À propos du "sentiment néologique" », *Langages* n° 36, pp.45-52.

GASIGLIA Nathalie (2009). « Évolutions informatiques en lexicographie : ce qui a changé et ce qui pourrait émerger », *Lexique, 19, Changer les dictionnaires*, Presses Universitaires du Septentrion, pp.235-298.

GAUME Bruno (2004). « Balades Aléatoires dans les Petits Mondes Lexicaux », *I3: Information Interaction Intelligence.*

GAUME Bruno (2006). « Rapport de fin de projet – DiLan : Du Trésor de la Langue Française Informatisé à une plateforme de ressources linguistiques pour le web sémantique et l'école », rapport final au 15/10/06 du projet 2004 *DiLan*. Responsable scientifique : Brno Gaume (IRIT).

GERARD Christophe (2010). « L'individu et son langage : idiolecte, idiosémie, style », *PhiN (Philologie im Netz*) 51/2010, p. 1-40. Consultable sur http://www.phin.de.

GEVAUDAN Paul (2002). « Fondements sémiologiques du modèle de la filiation lexicale », *Philologie im Netz*, volume 22/2002:1, pp.1-26.

GHEORGHITA Inga (2011). « Ressources lexicales au service de recherche et d'indexation des images », RECITAL 2011, Montpellier, 27 juin – 1er juillet 2011.

GRZESITCHAK Mick (2008). « Annotation sémantique : profilage textuel et lexical », *Lexicographie et informatique : bilan et perspective* à l'occasion, colloque international du 50e anniversaire du lancement du projet du *Trésor de la Langue Française*, 23-25 janvier 2008, Nancy.

GRZESITCHAK Mick, JACQUEY Evelyne, VALETTE Mathieu (2007). « Systèmes complexes et analyse textuelle : Traits sémantiques et recherche d'isotopies », *ARCo'07 – Cognition, Complexité, Collectif, Acta-Cognitica*, pp.227-235.

GUERIN Emmanuelle (2008). « Le français standard : une variété située ? », *Congrès Mondial de Linguistique Française – CMLF'08*, Durand J., Habert B., Laks B. (éds), ISBN 978-2-7598-058-3, Paris, 2008, Institut de Linguistique Française.

GUILBERT Louis (1965). *La formation du vocabulaire de l'aviation*, Paris, Librairie Larousse, 1965.

GUILBERT Louis (1971). « La néologie scientifique et technique », *La banque des mots*, n°1, pp.45-54.

GUILBERT Louis (1975). *La créativité lexicale*, Paris, Larousse.

GUILBERT Louis (1977). « Néologie et néologismes », *Beiträge zur Romanischen Philologie,* XVI, n°1, pp.113-118.

HABERT Benoît (2000). « Création de dictionnaires et typologie des textes » in Tyvaert J.-E. (éd.) *L'Imparfait, Philologie électronique et assistance à l'interprétation de textes,* Actes des Journées scientifiques 1999 du CIRLEP, Reims : Presses Universitaires de Reims, pp.171-188.

HANKS Patrick (2000). « Do word meanings exist ? » *Computers and the Humanities*, 34(1-2):205-215. Special Issue on SENSEVAL.

HARDIE A. (2009). « CQPweb – combining power, flexibility and usability in a corpus analysis tool », colloque *ICAME 30*, Lancaster, 27-31 May 2009.

HARRIS Zellig S. (1968). *Mathematical Structures of Languages*. New York : John Wiley & Sons.

HARRIS Z.S., BALAGNA J. (1970). « La structure distributionnelle », *Langages*, n°20, Analyse distributionnelle et structurale, pp.14-34.

HEIDEN Serge (2000). *Manuel de référence des Expressions CQP (2000)*, Documentation du logiciel Weblex, http://weblex.ens-lsh.fr/doc/weblex/refregexpcqp.html.

HEIDEN Serge, TOURNIER Maurice (2001). « Lexicométrie textuelle, sens et stratégie discursive », in *Simposio internacional de análisis del discurso*, Madrid : Spain (2001)

HEIDEN Serge (2004). « Interface hypertextuelle à un espace de cooccurrences: implémentation dans Weblex », *JADT 2004: 7èmes Journées d'Analyse des Données Textuelles*, pp.577-588.

HENRY Françoise (1990). « *Informatisation du Trésor de la langue française : problèmes et perspectives* », *Cahiers de lexicologie, 56-57,* Actes du colloque franco-danois de Lexicographie, Copenhague, 19-20 septembre 1988, pp.201-212.

HENRY Françoise (1996). « Les paramètres de l'analyse dans la pratique lexicographique », *Sémiotiques*, 11, pp.13-32.

IDE Nancy, VERONIS Jean (1998). « Word Sense Disambiguation: The State of The Art », Special Issue of *Computational Linguistics*, 24 (1), pp.1-40.

ILLOUZ G., HABERT B., FLEURY S., FOLCH H., HEIDEN S., LAFON P. (1999). « Maîtriser les déluges de données hétérogènes », Atelier Thématique *TALN 1999*, Cargèse, 12-17 juillet 1999.

IMBS Paul (1971). *Trésor de la Langue Française*, t.1, Préface, p. IX-XLVII.

ISSAC F., OUENNICHE S. (2010). « Pour une veille néologique à partir du web : l'outil telanaute », in *Actes del i Congrés internacional de neologia de les llengües romàniques (CINEO 2008),* Barcelone, Espagne, 7-10 mai 2008, pp.1165-1173.

JACQUEY Evelyne, KISTER Laurence, GRZESITCHAK Mick, GAIFFE Bertrand, REUTENAUER Coralie, OLLINGER Sandrine, VALETTE Mathieu (2010). « Thésaurus et corpus de spécialité sciences du langage : approches lexicométriques appliquées à l'analyse de termes en corpus », *TALN 2010*, Montréal, 19-23 juillet 2010.
www.iro.umontreal.ca/~felipe/TALN2010/Xml/Papers/all/taln2010_submission_166.pdf.

JANSSEN Maarten (2009). « Detección de Neologismos: una perspectiva computacional », *Debate Terminológico,* Vol. 5, pp.68-75.

KATSBERG-SJÖBLOM Margareta (2007). « Attirance thématique : fréquences ou séquences ? Extraction des isotopies sémantiques d'un corpus textuel », Actes des *JLC 5 (5e journées de la linguistique de corpus)*, 13-15 septembre 2007, Lorient.

KATZ SLAVA M. (1996). « Distribution of content words and phrases in text and language modelling. », *Natural Language Engineering*, 2(2), 15–59.

KELLER Tanja**,** TERGAN Sigmar-Olaf (2005). « Visualizing Knowledge and Information : An Introduction, *Knowledge and Information Visualization* », Springer, pp.1-23.

KILGARRIFF Adam (1997). « The hard parts of lexicography. », *International Journal of lexicography* 11 (1): 51-54. http://www.kilgarriff.co.uk/Publications/1997-K-IJL-HardParts.pdf.

KILGARRIFF Adam (2001). « Comparing corpora », *International Journal of Corpus Linguistics*, 6:1 (2001), 97–133.

KILGARRIFF Adam (2005). « Language is never, ever, ever random », *Corpus Linguistics and Linguistic Theory*, 1(2), pp.263-276.

KILGARRIFF A., HUSAK M., MCADAM K., RUNDELL M. and RYCHLÝ P. (2008). « GDEX: Automatically finding good dictionary examples in a corpus. », Actes d'*Euralex*, Barcelone, Espagne.

KILGARRIFF Adam, RYCHLY Pavel, SMRZ Pavel and TUGWELL David (2004). « The Sketch Engine », Actes d'*Euralex*. Lorient, France, juillet 2004, pp.105-116.

KLEIBER G. (1999). *Problèmes de sémantique. La polysémie en question.* Lille : Presses Universitaires du Septentrion.

KOHONEN T. (1989). *Self-Organization and Associative Memory,* Springer-Verlag, Berlin.

KUPIETZ Marc, BELICA Cyril, KEIBEL Holger, WITT Andreas (2010). « The German Reference Corpus DEREKO: A primordial sample for linguistic research. » In: Calzolari, Nicoletta et al. (eds.): *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC 2010)*. Valletta, Malta: European Language Resources Association (ELRA), 1848-1854. http://www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf.

L'HOMME Marie-Claude, BODSON Claudine, RENATA Stela Valente (1999). « Recherche terminographique semi-automatisée en veille terminologique : expérimentation dans le domaine médical », Depecker Loïc et Rousseau Jean-Louis (dir.) : *Nouveaux outils pour la néologie,* dans *Terminologies nouvelles,* n°20, décembre 1999, Bruxelles, Agence de la francophonie et Communauté française de Belgique, ISSN : 1015-5716, pp.25-36.

LABBE Cyril, LABBE Dominique (2001). « Que mesure la spécificité du vocabulaire ? », in *Lexicometrica n°3,* http://lexicometrica.univ-paris3.fr/article/numero3/specificite2001.PDF.

LAFON P. (1980) - « Sur la variabilité de la fréquence des formes dans un corpus », *Mots n°1*, pp.127-165.

LAFON P. (1981) - « Statistiques des localisations des formes d'un texte », *Mots*, N° 2, mars 1981, pp.157-188.

LAFON P. (1984). *Dépouillements et statistiques en lexicométrie.* Genève-Paris, éd. Slatkine – Champion, pp.54-77.

LANDHEER R. (2001). « La métaphore, une question de vie ou de mort ? », *Semen*, 5, pp.25-37.

LANDHEER R. (2005). « L'hyperbole : Figure de l'exagération illusionniste et foyer d'une polysémisation féconde », Publication inédite de la Conférence faite à l'Université de Franche-Comté à Besançon, le 26-05-2005, accessible en ligne sur le site La métaphore en question à l'adresse : http://www.info-metaphore.com/articles/landheer-hyperbole-figure-exageration-illusionniste-foyer-polysemisation-feconde-conference-universite-franche-comte.html.

LARSSON Staffan (2007). « A general Framework for semantic plasticity and negotiation ». In H. C. Bunt, editor, *Proceedings of the Seventh International Workshop on Computational Semantics (IWCS-7)*.

LE NY Jean-François (1989). « Accès au lexique et compréhension du langage : la ligne de démarcation sémantique », *Lexique*, 8, 65-85.

LEBART Ludovic (2004). « Validité des visualisations de données textuelles », Actes des *7e Journées internationales d'Analyse statistique des Données Textuelles (JADT 2004)*, Louvain, 10-12 mars 2004, pp.708-715.

LEBART Ludovic (2008). « L'analyse des données des origines à 1980 : quelques éléments », *Journal électronique des Probabilités et de la Statistique,* vol. 4, n°2, décembre 2008.

LEBART Ludovic, MORINEAU Alain, PIRON Marie (1995). *Statistique exploratoire multidimensionnelle*. Dunod, Paris, 1995. ISBN 2 10 002886 3.

LEBART L., PIRON Marie (2011). *Data and Text Mining. Visualization, inférence, classification. Logiciel d'analyse exploratoire multidimensionnelle des données numériques et textuelles.* Manuel d'utilisation de DTM-Vic. http://www.dtmvic.com/06_ManualF.html.

LEBART Ludovic, PIRON Marie, STEINER Jean-François (2003). *La sémiométrie – Essai de statistique structurale*. Dunod, Paris. ISBN 2-10-008105-5.

LEBART Ludovic, SALEM André (1988). *Analyse statistique des données textuelles*. Dunod. ISBN 2-04-018779-0.

LEBART L., SALEM A. (1994). *Statistique Textuelle*, Dunod, 344 p.

LECOLLE Michelle (2007). Support de cours de sémantique lexicale, licence Sciences du Langage, Université Paul Verlaine Metz, non publié.

LECOLLE Michelle (2007b). « Polysignifiance du toponyme, historicité du sens et interprétation en corpus », *Corpus*, n°6, pp.101-125.

LECOLLE Michelle (2009). « Changement de sens du toponyme en discours : de *Outreau* « ville » à *Outreau* « fiasco judiciaire » », *Les Carnets du Cediscor,* 11, 2009, PSN, pp.91-106.

LECOLLE Michelle (2004). « Mot d'introduction sur la métonymie », Accessible en ligne sur le site La métaphore en question à l'adresse : http://www.info-metaphore.com/articles/mot-d-intro-sur-la-metonymie.html.

LEMAIRE B. (2008). « Limites de la lemmatisation pour l'extraction de significations. » JADT 2008: 9es Journées internationales d'Analyse statistique des Données Textuelles (2008).

LEMNITZER Lothar, ULE Tylman (2011). *Die Wortwarte - auf der Suche nach den Neuwörtern von morgen.* In: URL: http://www.wortwarte.de/ (http://www.wortwarte.de/Projekt/index.html). Consulté le 11 juillet 2011.

LOISEAU Sylvain (2007). « CorpusReader : un dispositif de codage pour articuler une pluralité d'interprétations », *Corpus* [En ligne], n°6 | décembre 2007 , mis en ligne le 02 juillet 2008, Consulté le 04 avril 2011. URL : http://corpus.revues.org/index1282.html.

LOISEAU Sylvain, GREA Philippe, MAGUE Jean-Philippe (2010). « Dictionnaires, théorie des graphes et structures lexicales », *Revue de Sémantique et Pragmatique*. Juin 2010. Numéro 27. pp.51-78.

LONGREE Dominique, LUONG Xuan, MELLET Sylvie (2008). « Les motifs : un outil pour la caractérisation topologique des textes », *Actes des 9e JADT*, PUL, vol. 2, pp.733-744, Heiden S. et Pincemin B. (eds).

LONGREE Dominique, MELLET Sylvie, POUDAT Céline (2010). « Les taggers, auxiliaires heuristiques en ADT ? », In *Proceedings of 10th International Conference Journées d'Analyse Statistique des Données Textuelles*, Rome, Italie, 9-11 juin 2010, pp.1195-1206, ISBN 978-88-7916-450-9, http://lexicometrica.univ-paris3.fr/jadt/jadt2010/tocJADT2010.htm.

LUX-POGODALLA Veronika et POLGUERE Alain (2011) : « Construction of a French Lexical Network : Methodological Issues », *Proceedings of the International Workshop on Lexical Resources (WoLeR 2011)*, Ljubljana. À paraître.

MACQUEEN J. (1967). «Some methods for classification and analysis of multivariate observations », Proceedings of the fifth Berkeley Symposium on Mathematics, Statistics and Probabilities, Vol.1, pp.281-291.

MANNING Christopher D., SCHÜTZE Hinrich (2003). *Foundations of statistical natural language processing*. Cambridge, London, MIT Press (ed). ISBN : 0-262-13360-1.

MARTIN Robert (1992). *Pour une logique du sens*. 2e édition revue et augmentée. PUF.

MARTIN Robert (2001). *Sémantique et automate.* PUF. ISBN 2 13 0519296.

MARTINEZ Camille (2009). *L'évolution de l'orthographe dans les Petit Larousse et les Petit Robert 1997-2008 : une approche généalogique du texte lexicographique.* Thèse de doctorat, Université de Cergy-Pontoise.

MARTINEZ Camille (2011). « Mots nouveaux des dictionnaires », site du Club d'orthographe de Grenoble, accessible à l'adresse : http://orthogrenoble.net/page-de-camille-club-orthographe-grenoble.html. Consulté le [23 septembre 2011].

MARTINEZ William (2003). *Contribution à une méthodologie de l'analyse des cooccurrences lexicales multiples dans les corpus textuels*, Thèse pour le doctorat en Sciences du Langage, Université de la Sorbonne nouvelle - Paris 3.

MAYAFFRE Damon (2006). « Faut-il prendre en compte la composition grammaticale des textes dans le calcul des spécificités lexicales ? Tests logométriques appliqués au discours présidentiel sous la Vème République », in Actes des *8e Journées d'Analyse statistique des Données Textuelles (JADT 2008)*, 12-14 mars 2008, Lyon, pp.673-681.

MAYAFFRE D. (2007). « L'analyse de données textuelles aujourd'hui : du corpus comme une urne, au corpus comme un plan. Bilan sur les travaux actuels de topographie / topologie textuelle », *Lexicometrica 7*, http://www.cavi.univ-paris3.fr/lexicometrica/numspeciaux/ special9/mayaffre.pdf.

MAYAFFRE Damon (2008). « L'entrelacement lexical des textes, co-occurrences et lexicométrie », *Texte et Corpus,* 3, pp.91-102.

MAYAFFRE Damon (2008b). « De l'occurrence à l'isotopie. Les co-occurrences en lexicométrie », *Sémantique et Syntaxe 9*, pp.53-72.

MAYAFFRE Damon (2008c). « *Quand " travail ", " famille ", " patrie " co-occurrent dans le discours de Nicolas Sarkozy. Étude de cas et réflexion théorique sur la co-occurrence* », in Actes des *8e Journées d'Analyse statistique des Données Textuelles (JADT 2008)*, 12-14 mars 2008, Lyon, pp.811-822.

MEJRI Salah (1995). *La néologie lexicale*, Publications de la Faculté des lettres, Manouba, Tunisie, 1995, 380p.

MEJRI Salah (2010). « Néologie et traitement automatique », in *Actes del i Congrés internacional de neologia de les llengües romàniques (CINEO 2008),* Barcelone, Espagne, 7-10 mai 2008, pp.99-110.

MEL'CUK Igor (1997). *Vers une linguistique Sens-Texte,* leçon inaugurale, Paris, Collège de France, pp.41-57.

MEL'CUK Igor (2008). « Phraséologie dans la langue et dans le dictionnaire », *Repères et Applications (VI), XXIV Journées Pédagogiques sur l'Enseignement du Français en Espagne, Barcelone, 3-5 septembre 2007.*

MEL'CUK Igor, CLAS André, POLGUERE Alain (1995). *Introduction à la lexicologie explicative et combinatoire.* Duculot, Paris/Louvain-la-Neuve.

MEL'CUK Igor, POLGUERE Alain (2006). « Dérivations sémantiques et collocations dans le DiCo/LAF. *Langue française*, 150, pp.66-83.

MELLET Sylvie (2003). « Lemmatisation et encodage grammatical : un luxe inutile ? », in *Lexicometrica n°3, http://lexicometrica.univ-paris3.fr/article/numero3/sm2001.pdf.*

MELLET Sylvie, Barthélemy Jean-Pierre (2009). « La topologie textuelle : légitimation d'une notion émergente », *Lexicometrica* 7, *Topographie et topologie textuelles* (*cf.* http://www.cavi.univ-paris3.fr/lexicometrica/numspeciaux/special9/mellet.pdf).

MISRA Hemant, YVON François (2010). « Modèles thématiques pour la segmentation de documents », in *Proceedings of 10th International Conference Journées d'Analyse Statistique des Données Textuelles*, Rome, Italie, 9-11 juin 2010.

MISSIRE Régis. (2005) « Rythmes sémantiques et temporalité des parcours interprétatifs ». *Texto !* 2005 [en ligne]. Disponible sur : http:// www.revue-texto.net/Inedits/Missire.html (Consultée le 24 juin 2010).

MISSIRE Régis (2006). *Glossaire de sémantique.* Disponible sur http://www.revue-texto.net/Inedits/Missire/Missire_th_glossaire.pdf.

MOESCHLER J., REBOUL A. (1994). *Dictionnaire encyclopédique de pragmatique*. Paris: Seuil.

MULLER Charles (1964). *Essai de statistique lexicale. L'illusion comique de Pierre Corneille*, Paris, Klincksieck.

MULLER Charles (1968). *Initiation à la statistique linguistique*, Larousse.

MULLER Philippe, HATHOUT Nabil, GAUME Bruno (2006). « Synonym Extraction Using a Semantic Distance on a Dictionary », *Proceedings of the HLT/NAACL workshop Textgraphs*, Radev, Dragomir et Rada Mihalcea (eds), New York, NY. Association for Computational Linguistics, pp.65–72.

NAMER Fiammetta (2009). *Morphologie, lexique et Traitement Automatique des Langues - Le système DériF: TIC et Sciences cognitives*. London: Hermès Sciences Publishing, 448p.

NAMER Fiammetta, BOUILLON Pierret, JACQUEY Evelyne (2007). « Un lexique génératif pour le français », *TALN 2007*, Toulouse, 5-8 juin 2007, pp.233-242.

NAZAR Rogelio, VIDAL Vanesa (2010). « Aproximación cuantitativa a la neología », *Actes de CINEO*, pp.867-880.

NYCKEES V. (2000). « Changement de sens et déterminisme socio-culturel », *Théories contemporaines du changement sémantique,* J. François (Éd.), Leuven: Peeters (Société de Linguistique de Paris), pp.31-58.

OLLINGER Sandrine, VALETTE Mathieu (2010). « La créativité lexicale : des pratiques sociales aux textes », in *Actes del i Congrés internacional de neologia de les llengües romàniques (CINEO 2008),* Barcelone, Espagne, 7-10 mai 2008, pp.881-891.

PECINA Pavel, SCHLESINGER Pavel (2006). « Combining Association Measures for Collocation Extraction. », *Proceedings of the 21th International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006),* Sydney, Australia, 2006.

PEDERSEN T. (1996). « Fishing for exactness. », in *Proceedings of the South-Central SAS Users Group Conference*, Austin, TX.

PERLERIN Vincent (2004). *Sémantique légère pour le document. Assistance personnalisée pour l'accès au document et l'exploration de son contenu*. Thèse de doctorat en informatique, université de Caen, Basse-Normandie.

PETRALLI Alessio (1999). « Néologismes, internationalismes et mondialisation », *Terminologies nouvelles*, n°20, déc. 1999, pp.60-71.

PICTON Aurélie (2009). *Diachronie en langue de spécialité. Définition d'une méthode linguistique outillée pour repérer l'évolution des connaissances en corpus. Un exemple appliqué au domaine spatial*. Thèse de Doctorat en Sciences du Langage. Université Toulouse 2.

PINCEMIN Bénédicte (1999). *Diffusion ciblée automatique d'informations : conception et mise en œuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents*, Thèse de Doctorat, Linguistique, Université Paris IV (Sorbonne), 6 avril 1999, n°99PA040027, 806 pages.

PINCEMIN Bénédicte (2002) - « Sémantique interprétative et analyses automatiques de textes : que deviennent les sèmes ? », Benoît Habert (dir.), Dépasser les sens iniques dans l'accès automatisé aux textes, *Sémiotiques*, 17, décembre 1999, pp.71-120.

PINCEMIN Bénédicte (2008). « Modélisation textométrique des textes », *JADT 2008 : 9 Journées internationales d'Analyse statistique des Données Textuelles*, pp.949-960.

PINCEMIN Bénédicte (2009). « Analyse stylistique différentielle à base de marqueurs : compatibilité, potentiel et perspectives d'une approche logicielle textométrique », *Journées d'étude Le style et sa modélisation*, Tours, 10-11 décembre 2009, pp.54-61.

PINCEMIN Bénédicte, ISSAC Fabrice, CHANOVE Marc, MATHIEU-COLAS Michel (2006). « Concordanciers : thème et variations », *Actes des 8es Journées internationales d'Analyse statistique des Données Textuelles (JADT 2006)*, Jean-Marie Viprey et al. (éds), Besançon :Presses Universitaires

de Franche-Comté, ISBN 2.84867130.0, vol. II, pp.773-784. [En ligne sur le site *Lexicometrica* : http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2006/PDF/II-069.pdf].

PLOUX Sabine, BOUSSIDAN Armelle, FRANCO Charlotte, RENON Anne-Lyse (2011). « Dynamic diachronic modelling », présentation en ligne du *Semantic Atlas,* http://dico.isc.cnrs.fr/en/diachro.html. Consulté le 24 octobre 2011.

PLOUX Sabine, VICTORRI Bernard (1998). « Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes », *TAL*, Vol 39/1, pp.161-182.

POLGUERE Alain (2002). « Le sens linguistique peut-il être visualisé ? », in D. Lagorgette et P. Larrivée (dir.) : *Représentations du sens linguistique*, coll. "Lincom Studies in Theoretical Linguistics", 25, Munich : Lincom Europa, pp.89-103, http://olst.ling.umontreal.ca/pdf/Polguere2002.pdf.

POLGUERE Alain (2003) Étiquetage sémantique des lexies dans la base de données DiCo, TAL, Vol. 44 No 2.

POLGUERE Alain (2008). *Lexicologie et sémantique lexicale*. Presses de l'Université de Montréal, 2008, 2e édition.

PONS Pascal, LATAPY Mathieu (2006). « Computing communities in large networks using random walks » (version longue), *Journal of Graph Algorithms and Applications* (JGAA), Vol. 10, no. 2, pp. 191-218.

PRUVOST Jean (2002). *Les dictionnaires de langue française*. PUF, Coll. Que sais-je ?, Paris, 2002.

PRUVOST Jean, SABLAYROLLES Jean-François (2003). *Les néologismes*, PUF, Coll. Que sais-je?, Paris, 2003.

RADERMACHER Ruth (2004). *Le Trésor de la Langue Française. Une étude historique et lexicographique*, Thèse de doctorat, Université Marc Bloch, Strasbourg, 2004.

RAMDANI Egle (2007). *Du dictionnaire de langue au lexique TAL – la construction d'une ressource pour l'annotation sémantique des textes*, Mémoire de Master.

RASTIER F. (1987). *Sémantique interprétative*, Paris, PUF.

RASTIER François (1989). *Sens et textualité*. Paris, PUF.

RASTIER François (1996). « La sémantique des thèmes - ou le voyage sentimental. », *Texto ! 1996* [en ligne]. Disponible sur : <http://www.revue-texto.net/Inedits/Rastier/Rastier_Themes.html>. (Consultée le 15 avril 2011).

RASTIER F. (2001). *Sémantique et recherches cognitives*, Paris, PUF, pp.110-111.

RASTIER François (2008). « Obscure référence », *Zeitschrift für französische Sprache und Literatur - Beiheft n°35 (Dénomination, phraséologie, reference)*. P. Frath, coord., Franz Steiner Verlag, Stuttgart.

RASTIER François (2006). « Formes sémantiques et textualité », *Langages*, 2006/3 n°163, (éd) Armand Colin, ISSN 0458-726X, pp.99-114, disponible à l'adresse : http://www.cairn.info/revue-langages-2006-3-page-99.htm.

RASTIER François (2008). « Rhétorique et interprétation des figures », in *Figures de la figure,* Presses Universitaires de Limoges, pp.81-101.

RASTIER François, Cavazza Marc, Abeillé Anne (1994). *Sémantique pour l'analyse*. Masson.

RASTIER François, VALETTE Mathieu (2009). « De la polysémie à la néosémie », *Le français moderne*, S. Mejri (éd.), *La problématique du mot*, 77, pp.97-116.

RAYSON P., ARCHER D., PIAO S. L., MCENERY T. (2004). « The UCREL semantic analysis system. », *Proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks*

in association with 4th International Conference on Language Resources and Evaluation (LREC 2004), 25th May 2004, Lisbon, Portugal, pp.7-12.

RECANATI François (2007). *Le sens littéral.* Éditions de l'éclat, Paris-Tel-Aviv, collection *Tiré à part*.

REINERT Max (1983). « Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte », *Les cahiers de l'analyse de données*, tome 8, n°2 (1983), pp.187-198. http://www.numdam.org/item?id=CAD_1983__8_2_187_0.

REINERT Max (2002). *Alceste, Manuel de référence*, Université de Saint-Quentin-en-Yvelines, CNRS.

RENOUF Antoinette (2010). « Identification automatique de la néologie lexicologique et sémantique : questions soulevées par notre méthode », in *Actes del i Congrés internacional de neologia de les llengües romàniques (CINEO 2008),* Barcelone, Espagne, 7-10 mai 2008, pp.129-141.

REUTENAUER Coralie (2009). « Analyse et modélisation sémantiques à partir de ressources lexico-sémantiques », [En ligne], Vol. XIV (2009) n°1 (coordonné par Évelyne Bourion), URL : http://www.revue-texto.net/index.php?id=2095. Rapport de stage de fin d'études.

REUTENAUER Coralie, LECOLLE Michelle, JACQUEY Evelyne, VALETTE Mathieu (2009). « Outreau en n sèmes, Outreau en 5 temps, Diachronie de la représentation sémique d'une unité lexicale ». In Actes de l'Atelier "Du thème au terme", Conférence internationale *Terminologie et Intelligence Artificielle* (TIA), Toulouse, France, 21 novembre 2009, http://www.irit.fr/TIA09/thekey/tdmatelier1.htm.

REUTENAUER Coralie, LECOLLE Michelle, JACQUEY Evelyne, VALETTE Mathieu (2010). « Sémème au macroscope : genèse et variation sémiques d'une unité lexicale ». In *Proceedings of 10th International Conference Journées d'Analyse Statistique des Données Textuelles*, Rome, Italie, 9-11 juin 2010, ISBN 978-88-7916-450-9, pp.467-478, http://lexicometrica.univ-paris3.fr/jadt/jadt2010/tocJADT2010.htm.

REUTENAUER Coralie, VALETTE Mathieu, JACQUEY Evelyne (2009). « Proposition pour l'enrichissement sémantique de corpus », Actes des *6e Journées de la Linguistique de Corpus*, Lorient, 10-12 septembre 2009, http://web.univ-ubs.fr/corpus/jlc6.html#publi2009.

REUTENAUER Coralie, VALETTE Mathieu, JACQUEY Evelyne (2009). « De l'annotation sémantique globale d'un texte à l'interprétation locale d'un mot ». In *Cognitica, ARCo'09, Actes de Colloque de l'Association pour la Recherche Cognitive, Interprétation et problématiques du sens*, Rouen, 9-11 décembre 2009, http://arco09.colloques.univ-rouen.fr/spip.php?article25.

REUTENAUER Coralie (2010). « Propositions pour la détection automatique de la néosémie », colloque *Néologie Sémantique et corpus : une rencontre de méthodes*, Tübingen, Allemagne, 29-30 avril 2010. À paraître.

REY Alain (1974). «Essai de definition du concept de néologisme », *Actes du colloque international de terminologie,* O.L.F., 1974, Québec.

REY Alain (2008). *De l'artisanat des dictionnaires à une science du mot. Images et modèles*, Paris, A. Colin.

RINCK Fanny, TUTIN Agnès (2007). « Annoter la polyphonie dans les textes : le cas des passages entre guillemets », *Corpus*, 6:79-100.

ROSENTHAL Robert (1994). « Parametric measures of effect size », *The handbook of research synthesis*, Haaris Cooper et Larry V. Hedges (éds), Volume 236, chapitre 16, pp.232-244.

ROSSIGNOL Mathias, SEBILLOT Pascale (2002). « Automatic generation of sets of keywords for theme characterization and detection.», Actes des JADT2002, 13-15 mars 2002, Saint-Malo.

ROUSSEAU Jean-Louis (2010). « La néologie : foisonnement et harmonisation », in *Actes del i Congrés internacional de neologia de les llengües romàniques (CINEO 2008),* Barcelone, Espagne, 7-10 mai 2008, pp.189-195.

ROUSSEAU Jean-Louis, DEPECKER Loïc (1999). « Nouveaux outils pour la néologie », *Terminologies nouvelles,* n°20, déc. 1999, pp.2-3.

ROY Thibault (2007). *Visualisations interactives pour l'aide personnalisée à l'interprétation d'ensembles documentaires.* Thèse d'informatique Université de Caen.

ROY Thibault, BEUST Pierre (2005). « La cartographie thématique de corpus : une solution aux problèmes de veille documentaire ? », Actes de ISKO-France 2005, ISBN : 2-86480-817-X.

ROY Thibault, FERRARI Stéphane, BEUST Pierre (2005). « Cartographie de corpus pour l'étude de métaphores conceptuelles », *Journées de Linguistique de Corpus*, Lorient, Presses Universitaires de Rennes.

SABLAYROLLES Jean-François (2000). *La néologie en français contemporain : examen du concept et analyse de productions néologiques récentes*. Champion, Paris.

SABLAYROLLES Jean-François (2002). « Fondements théoriques des difficultés pratiques du traitement des néologismes », *Revue française de linguistique appliquée*, VII-1, pp.97-111.

SABLAYROLLES Jean-François (2003). « Métaphore et évolution du sens des unités lexicales », *Cahier du CIEL* 2000-2003, pp.109-124.

SABLAYROLLES Jean-François (2006). « Lacunes, flottements et trop-pleins », *Journées scientifiques sur la terminologie linguistique, Caen, FRANCE (12/05/2005)*, *Syntaxe et sémantique,* 2006, n° 7 (189 p.), ISSN 1623-6742, pp.79-89.

SABLAYROLLES Jean-François (2010). « *Extraction automatique et types de néologismes : une nécessaire clarification », »*, colloque *Néologie Sémantique et corpus : une rencontre de méthodes*, Tübingen, Allemagne, 29-30 avril 2010. À paraître.

SALEM André (1987). *Pratique des segments répétés. Essai de statistique textuelle*. Paris, Klincksieck.

SALEM André (1988). « Approches du temps lexical – Statistique textuelle et séries chronologiques », *Mots,* n°17, pp.105-143.

SALEM A., LAMALLE C., MARTINEZ W., FLEURY S., FRACCHIOLLA B., KUNCOVA A., MAISONDIEU A. (2003). « *Lexico3 – Outils de statistique textuelle. Manuel d'utilisation.* », Syled-CLA2T, Université de la Sorbonne nouvelle – Paris 3 : http://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW.

SALTON G., WONG A., YANG C.S. (1975). « A vector space model for information retrieval », *Journal of the American Society for Information Science*, 18(11):613-620, Nov. 1975.

SAMMON J. (1969). « A non linear mapping for data structure analysis. », *IEEE Transactions on Computing*, (18), 401–409.

SAUSSURE (DE), Ferdinand (1960). *Cours de Linguistique Générale*. pubs : Charles Bally et Albert Sechehaye ; avec la collaboration de Albert Riedlinger. Paris, Payot.

SAUVANET P., (2000), *Le rythme et la raison*. Paris, Kimé.

SCHAEFFER S.E. (2007). « Graph clustering », *Computer Science Review*, 1(1):27–64, août 2007.

Shannon C.E. (1948). « A Mathematical Theory of Communication », *The Bell System Technical Journal*, Vol.27 (1948), pp.379-423, pp.623-656.

SCHMID H. (1995). *TreeTagger - a language independent part-of-speech tagger.* Institute for Computational Linguistics of the University of Stuttgart. [Disponible à l'adresse http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/]

SCOTT M., TRIBBLE C. (2006). *Textual Patterns: Key words and corpus analysis in language education.* Philadelphia: John Benjamins (2006).

SOUTET Olivier (1995). *Linguistique.* PUF.

SPERBER Dan, WILSON Deirdre (1989). *La pertinence.* Traduit de l'anglais par Abel Gerschenfeld et Dan Sperber. Paris, Les éditions de Minuit, imprimé en 2009, copie de 1989.

TABOSSI P. (1988). « Effects of context on the immediate interpretation of unambiguous nouns », *Journal of Experimental Psychology : Learning, Memory and Cognition*, 14, pp.153-162.

TANGUY Ludovic (1997). *Traitement automatique de la langue naturelle et interprétation : contribution à l'élaboration informatique d'un modèle de la sémantique interprétative*, Thèse de Doctorat, Informatique, Université de Rennes I, 7 mai 1997, 207 pages.

TOURNIER J. (1985). *Introduction descriptive à la lexicogénétique de l'anglais contemporain*, Paris-Genève, Champion-Slatkine.

TANNIER Xavier (2006). *Extraction et recherche d'information en langage naturel dans les documents semi-structurés*. Thèse de doctorat, École des Mines de Saint-Étienne, Septembre 2006.

VALETTE Mathieu (2004). « Sémantique interprétative appliquée à la détection automatique de documents racistes et xénophobes sur Internet », *Approches Sémantiques du Document Numérique, Actes du 7e Colloque International sur le Document électronique, 22-25 juin 2004*, P. Enjalbert et M. Gaio, eds, 2004, pp.215-230.

VALETTE Mathieu (2008). « À quoi servent les lexiques sémantiques ? Discussion et proposition », *Description linguistique pour le traitement automatique du français*, M. Constant, A. Dister, L. Emirkanian & S. Piron, éd., *Cahiers du CENTAL*, n°5 – décembre 2008, Presses Universitaires de Louvain, 43-58.

VALETTE Mathieu (2009). *Approche textuelle du lexique*, mémoire pour l'Habilitation à Diriger des Recherches, Institut National des Langues et Civilisations Orientales, Paris.

VALETTE Mathieu (2010) « Méthodes pour la veille lexicale », in : *Actes de la journée d'étude Le dictionnaire électronique. Quelles perspectives pour les sciences humaines et sociales ?*, Leila Messaoudi (éds.), Publication du laboratoire Langage et société, Université Ibn Tofail Kénitra (disponible sur http://hal.archives-ouvertes.fr/).

VALETTE Mathieu (2010b), « Propositions pour une lexicologie textuelle », *Les configurations du sens,* Peter Blumenthal & Salah Mejri, éd., *Zeitschrift für Französische Sprache und Literatur*, 37, Franz Steiner Verlag (éd.), pp.171-188.

VALETTE Mathieu, ESTACIO-MORENO Alexander, PETITJEAN Etienne, JACQUEY Evelyne (2006). « Éléments pour la génération de classes sémantiques à partir de définitions lexicographiques. Pour une approche sémique du sens », *Verbum ex machina, Actes de la 13ème conférence sur le traitement automatique des langues naturelles (TALN 06)*, P. Mertens, C. Fairon, A. Dister, P. Watrin (éds). *Cahiers du CENTAL,* 2.1, UCL Presses Universitaires de Louvain. Volume 1, pp.357-366. (en ligne sur *Texto !*).

VARELA Francisco (1996). *Invitation aux sciences cognitives*. Éditions du Seuil.

VENANT Fabienne (2010). «Continu en sémantique », *JSM10, Semaine Nancéienne de Sémantique formelle*, Nancy, 22-24 mars 2010, http://jsm.loria.fr/jsm10/documents/lectures/venant.pdf.

VERONIS Jean (2004). « L'étiquetage sémantique des corpus », *Le Français Moderne*. *2004/1,* 27-38.

VICTORRI Bernard (1994). « La construction dynamique du sens », *Passions des formes – à René Thom*, R. Porte (Éd.), pp.733-747.

VICTORRI Bernard (2002). « Espaces sémantiques et représentation du sens », *Textualité et nouvelles technologies*, éc/arts, 3.

VIPREY Jean-Marie, SCHEPENS Philippe (2010). « Dérivation lexicale et dérive du discours : «mutualiser, mutualisation» », *Proceedings of 10th International Conference Journées d'Analyse Statistique des Données Textuelles* (JADT 2010), Rome, Italie, 9-11 juin 2010, ISBN 978-88-7916-450-9, pp.489-498, http://lexicometrica.univ-paris3.fr/jadt/jadt2010/tocJADT2010.htm.

WALTER H. (1984). « L'innovation lexicale chez les jeunes parisiens », *La linguistique*, n°20, 1984/2, pp.69-84.

WIERZBICKA Anna (1972). *Semantic Primitives*, Frankfurt am Main: Athenäum-Verl.

YATSKO V.A., STARIKOV M.S., BUTAKOV A.V. (2010). « Automatic Genre Recognition and Adaptive Text Summarization », *Automatic Documentation and Mathematical Linguistics,* Vol.44, No. 3, © Allerton Press Inc., ISSN 0005 1055, pp.111-120.

Annexes

Annexe 1. Comparaison d'indices statistiques – un exemple

La présente étude confronte les différents indices statistiques présentés au (chapitre II.2, 2.1.3.d), à savoir :

- l'écart-réduit *z* ;
- le t-score ;
- le χ^2 ;
- le G^2 ;
- les spécificités provenant du test exact de Fisher (spéc.) ;
- l'information mutuelle (IM).

L'objectif est d'illustrer l'écart entre les résultats produits par les indices. La comparaison des indices s'appuie sur un exemple de petite taille. Il est construit sur un ensemble d'unités lexicales provenant du corpus 'Outreau'. Celles-ci sont analysées en fonction de leur répartition dans le sous-corpus des voisinages d'*Outreau*, qui compte 40 474 occurrences de formes lexicales, par rapport au reste du corpus, d'une taille de 395 330 occurrences de formes. Les unités observées présentent des profils variés en termes de fréquence dans le sous-corpus et de fréquence dans le corpus.

	fréquence dans le sous-corpus (O_{11})	fréquence dans le corpus ($O_{11}+O_{12}$)
appel	16	223
arrestations	4	14
audition	5	117
commune	11	27
découverte	5	17
détention	9	382
doute	14	224
enfant	96	570
fouille	8	11
inculpés	11	20
interpellations	8	12
judiciaire	21	474
libérée	6	26
maltraitances	5	10
pédophile	73	213
pédophilie	60	333
police	29	114
politiques	5	39
réseau	127	364
réseaux	8	32
révélations	16	40
suspicions	5	21
témoignages	15	107

Figure A1.1 : Fréquences des unités lexicales observées

Les unités ne jouent pas de rôle majeur dans la comparaison d'indices, qui aurait pu être réalisée à partir de valeurs simulées, mais elles sont là pour donner aux résultats un ancrage dans une réalité textuelle.

Des scores sont affectés à chaque unité en fonction des quatre paramètres que sont (1) le nombre d'occurrences de l'unité dans le sous-corpus ; (2) le nombre d'occurrences de l'unité dans le corpus ; (3) la taille du sous-corpus ; (4) la taille du corpus. Un score est établi pour chacun des 6 indices comparés.

Les valeurs se répartissent comme suit selon la table contingence présentée au (chapitre II.2, 2.1.3.b) :

	D=voisinages d'*Outreau*	D≠voisinages d'*Outreau*	
X=unité observée	**fréquence dans le sous-corpus** **O_{11}**	O_{12}	**fréquence dans le corpus** **$O_{11}+O_{12}$**
X≠unité observée	O_{21}	O_{22}	$O_{21}+O_{22}$
	taille du sous-corpus **$O_{11}+O_{21}$= 40 474**	$O_{12}+O_{22}$	**taille du corpus** **N = 395 330**

Figure A1.2 : Table de contingence adaptée à l'exemple étudié

Les scores affectés aux unités sont les suivants :

	écart-réduit *z*	t-score	χ^2	G^2	spéc.	IM
appel	-1,43	-1,71	-2,28	-3,64	-1,54	-0,52
arrestations	2,14	1,28	5,12	5,25	1,42	1,48
audition	-2,02	-3,13	-4,54	-8,19	-1,98	-1,27
commune	4,95	2,48	27,33	24,66	4,52	1,99
découverte	2,47	1,45	6,8	6,9	1,72	1,52
détention	-4,82	-10,04	-25,85	-52,47	-8,78	-2,12
doute	-1,87	-2,39	-3,88	-6,41	-1,93	-0,72
enfant	4,92	3,84	27,09	33,42	6,38	0,71
fouille	6,47	2,43	46,74	34,94	5,84	2,82
inculpés	6,25	2,69	43,6	35,43	6,08	2,42
interpellations	6,1	2,39	41,58	31,81	5,41	2,7
judiciaire	-3,96	-6,01	-17,42	-31,23	-5,76	-1,21
libérée	2,04	1,36	4,66	5,16	1,51	1,17
maltraitances	3,92	1,77	17,2	14,43	2,78	2,28
pédophile	10,96	5,99	133,95	128,7	22	1,74
pédophilie	4,43	3,34	21,95	26,47	5,22	0,81
police	5,07	3,21	28,67	30,66	5,68	1,31
politiques	0,5	0,45	0,28	0,38	0,78	0,32
réseau	14,69	7,96	240,93	230,04	37	1,76
réseaux	2,6	1,67	7,58	8,16	2,02	1,28
révélations	5,88	2,97	38,55	35,02	6,16	1,96
suspicions	1,94	1,27	4,2	4,6	1,39	1,21
témoignages	1,22	1,04	1,66	2,18	1,28	0,45

Figure A1.3 : Valeurs retournées par les indices

Pour comparer les résultats retournés par les différents indices, nous utilisons une comparaison de rangs plutôt que des valeurs exactes, comme préconisé au (chapitre II.2, 2.1.3.d). Les lignes du tableau des rangs ci-dessous sont ordonnées en fonction des rangs associés aux spécificités, indice que nous avons retenu comme référence (chapitre II.2, 2.1.3.e).

	écart-réduit *z*	t-score	χ^2	G^2	spéc.	IM
réseau	1	1	1	1	1	7
pédophile	2	2	2	2	2	8
enfant	9	3	9	6	3	16
révélations	6	6	6	4	4	6
inculpés	4	7	4	3	5	3
fouille	3	9	3	5	6	1
police	7	5	7	8	7	11
interpellations	5	10	5	7	8	2
pédophilie	10	4	10	9	9	15
commune	8	8	8	10	10	5
maltraitances	11	11	11	11	11	4
réseaux	12	12	12	12	12	12
découverte	13	13	13	13	13	9
libérée	15	14	15	15	14	14
arrestations	14	15	14	14	15	10
suspicions	16	16	16	16	16	13
témoignages	17	17	17	17	17	17
politiques	18	18	18	18	18	18
appel	19	19	19	19	19	19
doute	20	20	20	20	20	20
audition	21	21	21	21	21	22
judiciaire	22	22	22	22	22	21
détention	23	23	23	23	23	23

Figure A1.4 : Rangs obtenus à partir des valeurs classées par ordre décroissant

Pour compléter l'analyse des rangs, un coefficient de corrélation est calculé sur les rangs pour tout couple d'indices.

	écart-réduit z	t-score	χ^2	G^2	spéc.	IM
écart-réduit *z*	1	0,92	1	0,98	0,96	0,86
t-score	0,92	1	0,92	0,95	0,97	0,67
χ^2	1	0,92	1	0,98	0,96	0,86
G^2	0,98	0,95	0,98	1	0,99	0,81
spéc.	0,96	0,97	0,96	0,99	1	0,75
IM	0,86	0,67	0,86	0,81	0,75	1

Figure A1.5 : Coefficients de corrélation calculés sur les rangs des indices

L'analyse du tableau de rangs, complétée par le tableau de corrélation, permet d'observer que, globalement, les indices produisent le même classement. Les résultats convergent, sauf pour l'information mutuelle. Le classement qu'elle retourne n'est pas radicalement différent, mais il présente tout de même des écarts notables. Les unités lexicales dont la fréquence dans le corpus est faible sont nettement privilégiées, ce qui n'est pas le cas pour les autres indices. Parmi les 5 indices restants, le χ^2 et l'écart-réduit *z* donnent exactement les mêmes résultats. Le t-score est l'indice qui se dissocie le plus de l'ensemble des 5 indices considérés. L'indice qui retourne les résultats les plus proches des spécificités est le G^2, même si ses résultats sont également proches du χ^2 et de l'écart-réduit *z*.

Annexe 2. Estimation de l'approximation dans le calcul des spécificités

Pour le calcul des spécificités, (Lafon, 1984) propose d'utiliser l'approximation suivante, issue de la formule de Stirling :

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}$$

Il précise que l'approximation est de 10^{-8} pour n≥33. Nous souhaitons préciser l'approximation obtenue pour la formule provenant de la loi hypergéométrique, à savoir :

$$p(X \geq k) = \frac{\binom{f}{k}\binom{T-f}{t-k}}{\binom{T}{t}} = \frac{f!(T-f)!t!(T-t)!}{k!(f-k)!(t-k)!(T-f-t+k)!T!} = \frac{\prod_{n\in\{f,T-f,t,T-t\}} n!}{\prod_{d\in\{k,f-k,t-k,T-f-t+k-f,T\}} d!} \quad (1)$$

k est la fréquence de l'unité observée dans le sous-corpus, *f* la fréquence dans le corpus, *t* la taille du sous-corpus et *T* la taille du corpus. On exclut de cette étude les cas où *k=0* et où *k=f*, où $\binom{f}{k} = 1$.

On note $p'(X \geq k)$ l'approximation de $p(X \geq k)$ en remplaçant les factorielles par la formule de Stirling correspondante.

Le développement limité de *n!* est le suivant :

$$n! = 1 + \frac{1}{12n} + \frac{1}{288n^2} - \frac{139}{51840n^3} + o\left(\frac{1}{n^4}\right)$$

D'où :

$$\frac{n!}{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}} = e^{-\frac{1}{12n}}\left(1 + \frac{1}{12n} + \frac{1}{288n^2} - \frac{139}{51840n^3} + o\left(\frac{1}{n^4}\right)\right)$$

Le développement limité de $e^{-\frac{1}{12n}}$ est :

$$e^{-\frac{1}{12n}} = 1 - \frac{1}{12n} + \frac{1}{288n^2} - \frac{5}{51840n^3} + o\left(\frac{1}{n^4}\right)$$

Par substitution dans l'équation puis développement du produit des développements limités, on a :

$$\frac{n!}{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}} = \left(1-\frac{1}{12n}+\frac{1}{288n^2}-\frac{5}{51840n^3}+o\left(\frac{1}{n^4}\right)\right)\left(1+\frac{1}{12n}+\frac{1}{288n^2}-\frac{139}{51840n^3}+o\left(\frac{1}{n^4}\right)\right)$$

$$= 1-\frac{1}{360n^3}+o\left(\frac{1}{n^4}\right)$$

En réinjectant dans (1), on obtient :

$$\frac{p(X\geq k)}{p'(X\geq k)} = \frac{\prod_{n\in\{f,T-f,t,T-t\}} 1-\frac{1}{360n^3}+o\left(\frac{1}{n^4}\right)}{\prod_{d\in\{k,f-k,t-k,T-f-t+k-f,T\}} 1-\frac{1}{360d^3}+o\left(\frac{1}{d^4}\right)} = \frac{1-\sum_{n\in\{f,T-f,t,T-t\}}\left(\frac{1}{360n^3}+o\left(\frac{1}{n^4}\right)\right)}{1-\sum_{d\in\{k,f-k,t-k,T-f-t+k-f,T\}}\left(\frac{1}{360d^3}+o\left(\frac{1}{d^4}\right)\right)}$$

Soit, en utilisant e développement limité de $\frac{1}{1-x} = 1+x+o\left(x^2\right)$:

$$\frac{p(X\geq k)}{p'(X\geq k)} = \left(1-\sum_{n\in\{f,T-f,t,T-t\}}\left(\frac{1}{360n^3}+o\left(\frac{1}{n^4}\right)\right)\right)\left(1+\sum_{d\in\{k,f-k,t-k,T-f-t+k-f,T\}}\left(\frac{1}{360d^3}+o\left(\frac{1}{d^4}\right)\right)\right)$$

$$\frac{p(X\geq k)}{p'(X\geq k)} = 1-\frac{1}{360}\sum_{\substack{n\in\{f,T-f,t,T-t\}\\ d\in\{k,f-k,t-k,T-f-t+k-f,T\}}}\left(\frac{1}{n^3}-\frac{1}{d^3}+o\left(\frac{1}{n^4}\right)+o\left(\frac{1}{d^4}\right)\right)$$

Comme les tailles respectives du sous-corpus et du corpus *t* et *T* sont très grandes devant la fréquence de l'unité considérée dans le sous-corpus et le corpus, on peut considérer que les seuls termes non négligeables sont ceux qui ne comportent ni *t,* ni *T*. Le fait que *t* et *T* soient négligeables devant les fréquences se traduit par l'ajout d'un terme en $o\left(\frac{1}{f^3}\right)$, qui permet d'éliminer le terme en $o\left(\frac{1}{f^4}\right)$.

$$\frac{p(X\geq k)}{p'(X\geq k)} = 1-\frac{1}{360}\left(\frac{1}{f^3}-\frac{1}{k^3}-\frac{1}{(f-k)^3}+o\left(\frac{1}{f^3}\right)+o\left(\frac{1}{k^4}\right)+o\left(\frac{1}{(f-k)^4}\right)\right)$$

Donc l'écart relatif entre la valeur exacte et la valeur approchée est :

$$\frac{p(X\geq k)-p'(X\geq k)}{p'(X\geq k)} = \frac{1}{360}\left(\frac{1}{k^3}+\frac{1}{(f-k)^3}-\frac{1}{f^3}+o\left(\frac{1}{f^3}\right)+o\left(\frac{1}{k^4}\right)+o\left(\frac{1}{(f-k)^4}\right)\right)$$

L'élimination des termes en *t* et *T* revient à considérer que l'écart entre la valeur exacte et la valeur approchée provient principalement du terme $\binom{f}{k}$. Les termes en *o* deviennent négligeables lorsque leur argument tend vers 0, autrement dit, pour *k, f* et *f-k* suffisamment grands. Nous présentons l'écart relatif calculé avec et sans les termes en *o* pour les valeurs les

plus faibles de *k* et *f*, afin de vérifier que ce qui est présenté comme négligeable l'est effectivement, même pour les valeurs les plus faibles.

k	f	$\frac{1}{360}\left(\frac{1}{k^3}+\frac{1}{(f-k)^3}-\frac{1}{f^3}\right)$	Écart exact
1	2	0,0052	0,0043
1	3	0,0030	0,0026
2	3	0,0030	0,0026
1	4	0,0028	0,0024
2	4	0,00065	0,00066
3	4	0,0028	0,0024
1	5	0,0028	0,0023
2	5	0,00043	0,00043
3	5	0,00043	0,00043
4	5	0,0028	0,0023
1	6	0,0028	0,0023
2	6	0,00038	0,00038
3	6	0,00019	0,00019
4	6	0,00038	0,00038
5	6	0,0028	0,0023
1	7	0,0028	0,0023
2	7	0,00036	0,00036
3	7	0,00014	0,00013
4	7	0,00014	0,00013
5	7	0,00036	0,00036
6	7	0,0028	0,0023

Tableau A2.1 : Écart relatif entre les spécificités approchées et exactes pour les fréquences faibles

Les termes en *o* sont effectivement négligeables même pour les valeurs faibles. L'approximation donnée par les termes en x^{-3} peut être vue comme un bon indicateur pour estimer l'écart entre les spécificités exactes et approchées.

Par ailleurs, l'écart entre spécificité exacte et spécificité approchée est inférieur à 10^{-2}. La fonction $f \longrightarrow \frac{1}{360}\left(\frac{1}{k^3}+\frac{1}{(f-k)^3}-\frac{1}{f^3}\right)$ est décroissante en *f*, donc cet écart de 10^{-2} est amené à diminuer au fur et à mesure que *f* augmente. Autrement dit, il faut aller au-delà du troisième chiffre significatif pour constater des différences entre la spécificité exacte et la spécificité approchée, quelle que soit la fréquence dans le corpus et dans le sous-corpus. Plus la fréquence augmente, plus le seuil de chiffres significatifs est élevé. À titre d'illustration, Lexico3 calcule les spécificités à partir de la méthode proposée par (Lafon, 1984), donc à partir de la valeur approximative. Les valeurs retournées comportent 1 à 2 chiffres significatifs (valeurs entières allant en valeur absolue de 1 à 50, puis saturation à 50), qui correspondent donc à ce qui aurait été obtenu à partir d'un calcul exact quelle que soit la fréquence considérée.

Index

Index général

Index des exemples