Institut National des Langues et Civilisations Orientales

Mathieu Valette

Approche textuelle du lexique

Mémoire

présenté dans le cadre de
l'habilitation à diriger des recherches

Coordonné par François Rastier

Novembre 2009

Merci

À Annie Montaut, Sylviane Rémi-Giraud, Salah Mejri, André Salem, Pierre Zweigenbaum d'avoir accepté de participer à mon jury d'habilitation, de lire mes travaux, de les évaluer et d'évaluer ainsi mon habileté à diriger ceux des autres.

À mes collègues de l'ATILF, en particulier Evelyne Jacquey, Sandrine Ollinger, Coralie Reutenauer, Bertrand Gaiffe, Mick Grzesitchak, Etienne Petitjean avec qui j'ai grand plaisir à travailler. C'est peu dire que ce mémoire leur doit beaucoup. À Jean-Marie Pierrel, directeur de l'ATILF, pour son soutien constant à mes recherches et sa confiance.

À mes collègues de l'ERTIM, Evelyne Bourion, Marie Kraskovetz, Egle Ramdani, Rachid Belmouhoub, Jean-Michel Daube, François Stuck qui m'ont aidé à achever ce mémoire, à grand renfort d'encouragements, de café, de relectures, de thé, de photocopies de biscuits, et qui ont supporté mon (exceptionnelle) mauvaise humeur.

À Monique Slodzian pour l'apport décisif à ma formation intellectuelle de sa manière de penser le monde, pour m'avoir appris les enjeux du multilinguisme et la fonction sociale de la linguistique et, par son biais, la mienne, peut-être.

À François Rastier pour la richesse et la portée de sa réflexion scientifique et de son œuvre, et pour son accompagnement généreux et indéfectible tout au long de mon parcours de recherche. Encore merci.

Résumé

La linguistique doit prendre part et position face aux nouveaux enjeux théoriques et méthodologiques naissant autour du document numérique et de l'élaboration des connaissances, et ne pas laisser à d'autres disciplines (sciences de l'information et de la communication, informatique) le soin de décrire, seules, ces nouveaux objets sémiotiques. Leur diversité et leur complexité sont en outre à problématiser tant dans la perspective de la variété des pratiques sociales que dans celle du multilinguisme. L'élaboration conjointe de modélisations linguistiques et d'outils informatiques destinés à leur validation et leur mise en œuvre s'avère une condition nécessaire à leur description. Dans ce cadre général, notre objectif est de présenter un ensemble de propositions visant à situer l'étude du lexique dans le paradigme textuel. Plus précisément, notre projet est d'étudier les déterminations textuelles de la conceptualisation et de la lexicalisation des concepts.

Dans le premier chapitre, nous effectuons une revue critique des principaux modes de structuration et de représentation du lexique, en particulier dans la perspective d'un traitement automatique du sens. Nous exposons ensuite certaines propositions de la sémantique interprétative et textuelle de F. Rastier en la matière. Après une présentation de la notion de classes sémantiques, nous nous focalisons sur l'une d'entre elles, le taxème, et nous discutons plus particulièrement de son rôle dans la représentation de la praxis. Dans le deuxième chapitre, nous traitons de la représentation du lexique du point de vue du texte, c'est-à-dire du point de vue de l'agencement syntagmatique. Nous abordons les différentes objectivations sémantiques proposées par la théorie susmentionnée (isotopies, molécules sémiques) de façon à mettre en évidence le rôle de l'articulation lexique/texte dans la cohésion textuelle. Dans le troisième chapitre, nous présentons un ensemble de travaux réalisés dans la perspective d'une instrumentation de l'analyse sémantique des textes et du lexique faisant la synthèse des recherches relatées dans les deux précédents chapitres. Enfin, dans un quatrième chapitre, nous abordons la question de la conceptualisation et de la lexicalisation des concepts. Nous proposons une méthode de description fondée sur les propositions théoriques et les outils informatiques décrits précédemment. Nous présentons, enfin, un ensemble de prospectives et un programme de recherche relatif à l'approfondissement de notre approche dans la perspective des nouvelles applications de la linguistique, en particulier dans un contexte variationniste et multilingue.

Table des matières

Introduction
La linguistique aujourd'hui

L'ère numérique modifie sensiblement les pratiques liées au texte. De nombreuses applications, telles que la navigation sur Internet ou la Gestion Electronique de Documents (GED), sont demandeuses de nouvelles méthodologies et de nouvelles façons d'appréhender les textes dans toute leur diversité et leur complexité. Comme elles ont le texte, le corpus et l'archive comme matériau privilégié, les sciences humaines et sociales sont tenues de renouveler le discours scientifique sur le texte dans cette perspective numérique[1].

Jadis *science-pilote* des sciences humaines d'un point de vue épistémologique, la linguistique qui a, parmi ces objets d'étude, le texte, pourrait dans ce contexte prétendre aujourd'hui au statut de *science-pivot* des sciences de l'homme et de la société d'un point de vue méthodologique. Si elle bénéficie en effet d'une expérience du texte, tant d'un point de vue théorique (linguistique textuelle, analyse du discours, philologie) que pratique (linguistique de corpus), la linguistique demeure très en retrait dans les applications (fouille de textes, recherche d'information, ingénierie des connaissances, etc.), Pourtant, elle est à même de proposer de nouvelles méthodologies d'analyses du sens et d'outiller les sciences humaines pour que celles-ci s'approprient à leur tour la problématique du document numérique. Plus encore, il apparaît vital, pour la linguistique, de prendre part et position face aux nouveaux enjeux théoriques et méthodologiques naissants, et de ne pas laisser à d'autres disciplines (sciences de l'information et de la communication, informatique) le soin de décrire, seules, ces nouveaux objets sémiotiques.

[1] Des initiatives interdisciplinaires très dynamiques telles que le RTP-DOC du CNRS (http://rtp-doc.enssib.fr) témoignent de l'importance de ce débat.

1. De la linguistique de corpus à une science des textes instrumentée

Il n'est pas impossible que la « linguistique de corpus », en tant que discipline candidate des sciences du langage, ait vécu. Aujourd'hui, de plus en plus de linguistes, quels que soient leur discipline ou leurs objets d'étude, sont amenés à constituer des corpus numériques et à les étudier au moyen d'outils logiciels chaque année plus nombreux, sophistiqués et conviviaux. La banalisation de l'outil désenclave ainsi des pratiques réservées jusque-là à une petite minorité que l'informatique ne rebutait pas. Cette évolution technologique, si elle a une incidence méthodologique évidente (par exemple et minimalement, en substituant aux exemples construits des exemples attestés), n'a pas pour autant un impact fort sur la définition des objets de la linguistique. À la morphologie les corpus de mots, à la syntaxe les corpus de phrases, aux théories énonciatives les corpus d'énoncés. Et bien que tous ces objets d'étude proviennent de textes, ceux-ci ne sont que rarement considérés comme objet de science et sont réduits, par défaut, au statut préscientifique de ressource – un matériau brut dont la qualité est déterminée par la seule présence, après raffinage, de l'objet étudié[2]. On collecte ainsi de l'indénombrable : du texte, du corpus – jamais du mot ou de la phrase mais des mots et des phrases[3].

Or, le passage d'une problématique du texte numérisé à celle du *document numérique**[4] (lequel est, pour beaucoup, le vecteur d'une révolution aussi importante que jadis le passage du *volumen* au *codex*) constitue un des enjeux de la linguistique de corpus. La linguistique des textes s'est en effet jusque-là consacrée à l'analyse des textes littéraires ou politiques, aux genres globalement bien décrits. Elle se trouve désormais confrontée à une grande variété de discours et de genres nouveaux, indéterminés, polymorphes, souvent multilingues, et en permanente évolution[5]. S'il s'agit souvent d'une modernisation de pratiques anciennes, ces genres sont aussi la trace de nouvelles pratiques sociales.

---

[2] Lire Valette 2008 pour une critique du *ressourcisme** en linguistique.

[3] Cette observation fait écho à l'opposition proposée par (Tognini-Bonelli 2001), rapportée par (Mayaffre 2005), entre les études *corpus-based* et les études *corpus-driven*. Mayaffre conclut que « la frontière semble aujourd'hui déplacée mais sans être abolie. Elle ne sépare plus ceux qui utilisent les corpus et ceux qui ne les utiliseraient pas, mais les linguistes qui se servent des corpus pour valider leur hypothèse et ceux qui les servent pour construire leur savoir ».

[4] Les termes suivis lors de leur première occurrence d'un astérisque sont définis dans le glossaire du mémoire.

[5] Aujourd'hui, par exemple, les "pages personnelles" n'existent pour ainsi dire plus sur Internet, on leur a substitué les blogs.

Parmi les linguistiques du texte, les propositions théoriques de F. Rastier (sémantique interprétative, sémantique textuelle) participent à ce débat[6]. Ayant pour objet empirique le texte et non le mot, la phrase ou l'énoncé, traditionnellement privilégiés, cette linguistique-science des textes renoue avec une tradition rhétorique et herméneutique oubliée du XXème siècle et se focalise sur l'étude de la textualité, des genres textuels, des discours et de leurs corollaires (cohésion textuelle, intertextualité, etc.). Son appareil théorique est depuis le début des années 90 éprouvé par la linguistique de corpus et le TAL. L'instrumentation logicielle s'associe ainsi aux outils théoriques et conceptuels. Constitutive de la linguistique de corpus, elle donne lieu à ce qu'on pourrait appeler son « cercle vertueux » : d'un côté, les grandes masses de données textuelles ou documentaires nécessitent, pour être analysées et décrites, des instruments ad hoc, de l'autre, cette instrumentation permet de construire de nouveaux observables qui seraient demeurés invisibles autrement.

2. Outil théorique pour l'interprétation du sens

L'interprétation des textes assistée par ordinateur fait certes l'objet d'une riche littérature[7] mais elle est le plus souvent réduite à quelques aspects récurrents : désambiguïsation, identification des domaines et des thèmes. D'une manière générale, la question du sens est dominée par une approche lexicale : le sens est dans les mots et il se calcule compositionnellement à partir de ceux-ci. Or, ni la problématique du texte ni celle du document numérique ne peuvent s'en satisfaire exclusivement. Elles impliquent de prendre en compte les ensembles de textes, les corpus, l'archive, l'intertexte. Le mot demeure un objet d'importance, notamment dans les langues qui le privilégient, mais il doit être rapporté à sa juste proportion, un fragment du texte.

Les propositions théoriques de la sémantique textuelle (Rastier, 2001) permettent d'étudier la structuration sémantique d'un texte, par le biais d'*objectivations sémantiques**, c'est-à-dire des réseaux de traits sémantiques, ou *sèmes**, qui en assure la cohésion. La sémantique textuelle en détaille deux catégories : (i) l'isotopie, qui relève du *fond sémantique** – il s'agit de la récurrence d'un même sème sur un empan de longueur variable, de la phrase au corpus ; (ii), la molécule sémique, qui correspond à la *forme sémantique** – il s'agit de patrons stabilisés et récurrents de sèmes hétérogènes. À l'interface entre le lexique et le texte, les réseaux sémiques permettent d'étudier de

[6] Lire par exemple Rastier 2008.

[7] *Cf.* pour ne citer que les publications les plus récentes : (Corblin & Gardent, éd. 2005), (Enjalbert éd., 2005), (Condamines, éd. 2005).

manière approfondie à la fois le lexique, le texte et leur relation cohésive. Leur intérêt descriptif et applicatif a été montré dans différents contextes. Dans le cadre d'un projet d'identification des pages web racistes sur Internet, (Valette 2004) met en évidence la variété des formes sémantiques associées à une même unité lexicale dans deux corpus de textes contrastés racistes *vs* antiracistes. La forme *« étranger »*, dans les textes antiracistes, est par exemple en cooccurrence avec *« irrégularité »* et *« régularisation »*, tandis que dans les textes racistes, elle cooccurre de façon privilégiée avec *« illégalité »* et *« naturalisation »*. Le discours sur l'Étranger relève d'un fond sémantique partagé à la fois par les textes antiracistes et les textes racistes, mais ce qu'il en est dit, de l'Étranger, varie en fonction des sous-corpus.

Un programme de recherche pour une science des textes instrumentée pourrait, par conséquent, s'articuler autour de trois objectifs conjoints : approfondir les connaissances actuelles sur les objectivations sémantiques connues et notamment référencées par la sémantique textuelle, en identifier de nouvelles que la théorie n'a pas su jusque là reconnaître faute d'une instrumentation adéquate et, enfin, créer de nouveaux observables sémantiques, textuels et lexicaux.

C'est dans ce contexte théorique et ingénierique que nous souhaitons situer notre propos, en nous focalisant sur un aspect particulier de l'interprétation dont les incidences en termes d'application (recherche d'information, industrie de la connaissance) apparaissent identifiables aisément : l'émergence textuelle des concepts et leur lexicalisation.

3. Une approche textuelle du lexique

Le mot est un concept linguistique fragile. A la fois imprécis de par ses frontières théoriques et matérielles et ethnocentrique parce que les langues sans mot sapent tout espoir d'en faire un concept universel, il demeure néanmoins un mode d'aperception du langage parmi les plus intuitifs et, selon toute vraisemblance, le plus étudié en linguistique, en dépit du succès de la phrase et de l'énoncé depuis plus d'un demi-siècle. Une approche textuelle du mot pourrait, comme le suggère (Rastier 2001, 182-183) à propos d'une reconception possible du signe, s'inspirer d'un texte où Saussure dit :

> « […] vous n'avez plus le droit de diviser, et d'admettre d'un côté le mot, de l'autre sa signification. Cela fait tout un. Vous pouvez seulement constater le kénôme ∩ et le sème associatif ⊃⊂ » (Saussure 2001, 93)

Rastier – qui, à notre connaissance, est le seul linguiste à avoir discuté de cet extrait – normalise le kénôme suivant le symbole mathématique de l'intersection « ∩ » et l'interprète à l'aune de la représentation saussurienne du signe (Saussure [1916] 1972, 158) comme un « signifié ouvert vers des signifiants indéterminés ». Quand au sème associatif, il le restitue avec deux symboles d'inclusion accolés « ⊃⊂ », neutralisant de la sorte la petite intersection visible dans le texte saussurien. Le sème associatif est, selon Rastier, « le signe linguistique contextuellement défini » et lui permet de développer son concept de *passage**. Le signe (ou le mot) est vu comme un passage ontologiquement vide entre deux contextes, gauche et droite.

La radicalité de cette conception du signe nous agrée et nous l'étendons, par méthode et provisoirement, à celle du mot. La lexicologie a recours en maints lieux théoriques et pratiques au contexte pour définir le contenu sémantique du mot (par la collocation notamment)[8], mais ce contexte est perçu d'un point de vue syntaxique ou micro-syntaxique. Il importe selon nous d'aborder le contexte du mot comme un objet et donc, dans la perspective de son objectivation. Dans ce cadre général, nous présentons un ensemble de propositions visant à situer l'étude du lexique dans le paradigme textuel. Plus précisément, notre projet est d'étudier les déterminations textuelles de la conceptualisation et de la lexicalisation des concepts.

4. Plan du mémoire

Dans le premier chapitre, nous effectuons une revue critique des principaux modes de structuration et de représentation du lexique, en particulier dans la perspective d'un traitement automatique du sens. Nous exposons ensuite certaines propositions de la sémantique interprétative et textuelle de F. Rastier en la matière. Après une présentation de la notion de classes sémantiques, nous nous focalisons sur l'une d'entre elles, le taxème, et nous discutons plus particulièrement de son rôle dans la représentation de l'expérience praxique. Dans le deuxième chapitre, nous traitons de la représentation du lexique du point de vue du texte, c'est-à-dire du point de vue de l'agencement syntagmatique. Nous abordons les différentes objectivations sémantiques proposées par la théorie susmentionnée (isotopies, molécules sémiques) de façon à mettre en évidence le rôle de l'articulation lexique/texte dans la cohésion textuelle. Dans le troisième chapitre, nous présentons un ensemble de travaux réalisés dans la perspective d'une instrumentation de l'analyse sémantique des textes et du lexique faisant la synthèse des recherches relatées dans les deux précédents chapitres. Enfin, dans un quatrième chapitre, nous abordons la question de

[8] Lire (Blumenthal & Hausman, éd., 2006).

la conceptualisation et de la lexicalisation des concepts. Nous proposons une méthode de description fondée sur les propositions théoriques et les outils informatiques auparavant décrits. Nous présentons, enfin, un ensemble de prospectives et un programme de recherche relatif à l'approfondissement de notre approche dans la perspective des nouvelles applications de la linguistique, en particulier dans un contexte variationniste et multilingue.

Chapitre 1
Structures du lexique pour l'analyse sémantique automatisée

1. Outils d'analyse et d'observation et observables

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

La linguistique de corpus propose de collecter des données attestées (c'est-à-dire non produites par le linguiste pour les besoins de l'analyse) et repose sur deux conditions d'observation conjointes : d'une part, l'usage d'outils d'analyse, d'autre part, la nécessité d'un grand volume de données. Ainsi, d'un côté le corpus impose l'usage d'instruments d'observation, de l'autre, l'instrumentation permet d'obtenir des mesures objectives. C'est ainsi que la linguistique de corpus s'est développée en même temps que les statistiques lexicales (Guiraud, 1960, Muller 1964, Lebart & Salem 1994). À l'outillage voué à la lecture (indexation, concordancier), on a associé des outils dédiés au comptage et à la pondération (calcul de l'écart réduit, loi hypergéométrique, etc.).

Que la science soit instrumentée ou non, et à moins d'adopter une position empiriste radicale, les données observables sont toujours des constructions. Doter la linguistique

d'une instrumentation implique aussi la production des données. Produire des données peut vouloir dire constituer des corpus de textes mais de plus en plus souvent construire des données métalinguistiques, c'est-à-dire créer des observables qui ne sont plus des données empiriques. On parle d'annotation en métadonnées ou d'enrichissement.

Ainsi, la communauté scientifique a su doter le français des instruments permettant l'annotation des parties du discours dans un texte. Autrement dit, les étiqueteurs morphosyntaxiques disponibles pour cette langue produisent des données métalinguistiques (lemmes, catégories grammaticales, etc.) réputées fiables. (Véronis 2000), s'appuyant sur des critères de robustesse et de diffusion, qualifie d'« opérationnel » ce stade d'avancement de l'instrumentation ; et (Habert 2005), privilégiant l'adhésion sociale (i.e. académique et scientifique) parle d'« homologation ».

Mais avant d'atteindre ce stade, la construction des données connaît de multiples phases de prospection. Le Traitement Automatique du Langage (TAL) vise dans une certaine mesure à doter la linguistique d'une instrumentation et est donc voué à ces prospections. Véronis (*op.cit.*) distingue deux stades en amont de l'opérationnalité :

- la recherche (« existence de travaux de nature prospective, mais ne donnant pas encore lieu à des implémentations utilisables »)
- le prototypage (« existence de logiciels dans les laboratoires, mais qui ne sont pas encore suffisamment robustes et suffisamment testés pour faire l'objet d'une diffusion importante »).

En 2000, il observait que l'annotation sémantique de texte, c'est-à-dire la création de données dédiées à l'analyse sémantique, n'en était qu'au stade de la recherche ou du prototypage. Pour son analyse, il s'appuyait sur les existants à l'époque, à savoir les ontologies de WordNet et EuroWordNet. Depuis, d'autres ressources ont été développées en premier lieu FrameNet (Fillmore *et al.* 2002).

2. Les ressources lexicales pour le TAL

(Habert 2005), s'inspirant de Simondon propose de distinguer, dans le champ de l'instrumentation du TAL, les outils, les instruments, les ressources et les dispositifs expérimentaux :

- les outils sont des « logiciels multi-usages, polyvalents, non spécifiquement orientés vers le traitement de données langagières mais qui peuvent être mis à contribution en linguistique » ;
- un instrument est un « logiciel qui prend en entrée une donnée langagière (du texte, de l'oral, un lexique…) et qui permet d'obtenir en sortie une représentation transformée (annotée) » ;
- les ressources sont des « données (corpus écrits, dictionnaires électroniques, bases de données) […]. Les données sont de plus en plus assorties d'instruments » ;
- et enfin, un dispositif expérimental est un « montage d'instruments, d'outils et de ressources servant à produire des « faits » dont la reproductibilité et le statut (l'interprétation) font l'objet de controverses (ex. Bruno Latour) ».

Dans le domaine des ressources, les lexiques pour le TAL dédiés à l'annotation de corpus sont en plein essor et évoluent rapidement. Si l'on s'en tient aux ressources existantes ou en cours de constitution, on constate que les options théoriques se répartissent en deux approches correspondant à l'opposition classique entre linguistique de la langue (approche paradigmatique) et linguistique du discours (approche syntagmatique) :

(i) Approches paradigmatiques : D'inspiration philosophique et terminologique, et principalement représentées par les thésaurus et les ontologies telles que WordNet (Fellbaum 1998), les approches paradigmatiques proposent une représentation close du monde ou du domaine, où la signification des unités lexicales dépend de relations hiérarchiques (hyperonymie, hyponymie, etc.) construites en fonction des référents qu'ils désignent. Centrées sur la référence et non sur les usages en discours, les ontologies relèvent peut-être autant, sinon davantage, de la philosophie que de la linguistique. Elles sont parfaitement légitimes dans la perspective terminologique des ontologies de domaine spécialisé (M.-Cl. Lhomme 2004, P. Zweigenbaum *et al.* 2007), mais elles sont insuffisantes pour rendre compte du sens des unités lexicales dès lors qu'elles sont actualisées dans des textes où l'ambiguïté est tolérée (discours politique ou journalistique), voire encouragée (discours littéraire), à l'inverse des textes techniques, notamment, où un terme désigne un concept ou un objet en principe sans équivocité. Les relations entre items retenues par tradition s'avèrent par ailleurs lacunaires. L'hyperonymie et l'hyponymie en particulier témoignent d'une représentation des connaissances ensembliste, abstraite, et non déterminée par les usages ou par quelques usages particuliers. WordNet et les ressources généralistes visant à l'universalité ne proposent finalement qu'une certaine représentation du lexique[9], indubitablement convaincante d'un point de vue

[9] Pour une critique plus développée, *cf.* Slodzian 1999, Rastier 2004, Hanks & Pustejovsky 2005.

encyclopédique en l'occurrence, mais nullement universelle d'un point de vue linguistique et sociolinguistique. On illustrera ces limites dans le paragraphe *Les structures du lexique défilent dans la rue*.

(ii) Approches syntagmatiques : Inspirées par la logique et la grammaire, les approches syntagmatiques s'intéresse au cotexte du lexique et donnent une large place à la syntaxe. Le Lexique Génératif (Hanks et Pustejovsky 2005), VerbNet (Kipper, Dang & Palmer, 2000 ; Kipper 2003, inspiré de Levin 1993) et FrameNet (Fillmore, Baker & Sato 2002, inspiré de Fillmore 1968) en sont d'excellents représentants. Dans FrameNet, par exemple, les unités lexicales sont livrées accompagnées d'une notice d'actualisation décrivant la combinatoire syntaxique (arguments) et sémantique (actants) de leurs différentes acceptions. Les approches syntagmatiques apparaissent plus pertinentes, linguistiquement parlant, que les approches paradigmatiques parce qu'elles tiennent compte de l'énonciation. En d'autres termes, elles ont pour objet le sens de l'unité lexicale dans l'énoncé plutôt que sa référence. Elles reposent toutefois sur une vision grammaticale du sens, où le syntagme et la phrase constituent les deux unités prises en compte, et adaptent les acteurs traditionnels de la sémantique de l'énoncé (thème/rhème, agent/patient, etc.). Mais si la signification d'une unité lexicale peut sans encombre être rapportée au lexique, son sens dépend dans une large mesure du texte dans son unité et du corpus d'accueil, autrement dit, des usages socialement codifiés et linguistiquement organisés en discours [10].

3. Le contexte entre le mot et le texte

Polguère fait l'hypothèse que les fenêtres cotextuelles larges, ont tendance à ramener des cooccurrents liés à des relations paradigmatiques, tandis que les fenêtres cotextuelles étroites, des cooccurrents liés par des relations syntagmatiques[11].

[10] Complétons cette typologie d'une observation transversale fondée sur le mode d'acquisition de la ressource. Distinguons : (i) l'acquisition introspective où les ressources sont produites à partir de l'expertise du linguiste et éventuellement validées *a posteriori* par des collections d'attestations issues de corpus (WordNet, Le Dico). (ii) l'acquisition empirique, où les ressources proviennent d'extraction sur corpus (FrameNet, Generative Lexicon, terminologie textuelle). L'acquisition en corpus est reconnue depuis quelques années déjà, pour sa nature objectivante et sa robustesse (Fuchs et Habert 2004), (Condamines 2005). Des efforts visant à renforcer encore cette méthode méritent d'être soulignés. Ainsi, Hanks et Pustejovsky ont proposé récemment de repérer des usages « normaux » de la valence des verbes à partir de l'analyse de patrons basée sur des corpus (Corpus Pattern Analysis, CPA) [Merci à Evelyne Jacquey de nous avoir signalé ces travaux]. Cette acquisition est semi-automatique.

[11] Communication personnelle. Le lexique francophone DiCo (Jousse & Polguère 2005, inspiré de Mel'čuk 1995, 2005) opte d'ailleurs explicitement pour une approche mixte, à la fois paradigmatique (dérivation sémantique) et syntagmatique (consignation des principales collocations). Mais son objet,

La question du contexte, passée la phrase (c'est-à-dire du paragraphe à l'intertexte), est cependant loin d'être toujours problématisée dans les ressources lexicales, même lorsqu'elles sont syntagmatiques. Pourtant, les problèmes conjoints de la dimension, de l'hétérogénéité et de la nature du cotexte, c'est-à-dire du contexte linguistique ont été très étudiés, comme l'atteste (Ide & Véronis 1998) et plus récemment (Crestan *et al.* 2003). Mais si déterminer une limite inférieure à la dimension d'une fenêtre cotextuelle semble possible, les limites supérieures sont plus difficiles à établir. Tout se passe comme si, à mesure que le cotexte s'élargit, il se rapproche du contexte, c'est-à-dire du contexte extralinguistique, lequel peut être, dans certains cas, assimilé à l'usage socialement déterminé. Or, c'est ce cotexte élargi qui nous intéresse particulièrement ici.

Prenons l'exemple d'un des succès reconnus de l'approche syntagmatique en matière de traitement automatique du sens : la désambiguïsation. (Ide & Véronis 1998 : 21-22), après avoir établi diverses qualités de cotexte, du plus étroit au plus large (le domaine thématique), font état de la théorie des scripts de (Schank & Abelson 1977) qui consiste à identifier dans un texte des scénarii préétablis pour choisir les emplois correspondant à la pratique ainsi formalisée (par exemple, au restaurant : entrer, commander, manger, sortir). La théorie des scripts repose certes sur une approche pratique des faits linguistiques, mais ces scripts sont des modèles conceptuels d'inspiration cognitiviste et ne sont pas hérités de l'analyse de corpus. En conséquence, ils ne relèvent pas du cotexte à proprement parler mais d'un préjugé sur le contexte projeté sur le cotexte. Substituer au préjugé une objectivation du contexte constitue l'enjeu d'une linguistique textuelle.

En effet, le contexte du mot ne se limite pas à une fenêtre de 5 ou 20 mots, mais à un intertexte qu'il importe de savoir simuler. On songe ici à la notion de « corpus réflexif » élaborée par (Mayaffre 2002), dans la continuité des observations de (Rastier 1998 : 17) selon qui « Le corpus est la seule objectivation possible (philologique) de l'intertexte, qui sinon demeure une notion des plus vagues ». Mayaffre choisit quant à lui de confondre le corpus (de travail) et l'archive dans la constitution de « macro-corpus réflexifs » dans lesquels s'organisent des réseaux sémantiques « cohérents et auto-suffisants » :

> L'objet d'étude […] et les sources (les archives […]) seront rassemblés et reliés dans le corpus, traités d'un même mouvement, par une même méthode. L'historien là encore n'aura pas ou moins à sortir de son corpus et de son traitement scientifique pour l'éclairer par un travail spécifique d'archive. Et le travail même d'archive sera partie intégrante du travail de saisie et de constitution du corpus. Plus largement encore, le développement de gros corpus

---

explicitement lexicologique, est circonscrit aux mêmes unités moyennes que la plupart des lexiques d'inspiration logico-grammaticale : mot, syntagme, phrase.

> réflexifs et enrichis permet d'envisager, à terme, l'intégration de la bibliographie dans le corpus lui-même. Un discours sera alors relié non seulement aux textes qui constituent son co-texte, mais aux écrits scientifiques permettant de mieux le comprendre (Mayaffre 2002).

Cette proposition est d'un intérêt théorique extrême. On peut l'illustrer à l'aide d'une petite expérience très simple. Prenons le premier quatrain de *Bohémiens en voyage*, de Ch. Baudelaire :

> La tribu prophétique aux prunelles ardentes
> Hier s'est mise en route, emportant ses petits
> Sur son dos, ou livrant à leurs fiers appétits
> Le trésor toujours prêt des mamelles pendantes.

L'ambiguïté de *« prunelles »*, à la fois « fruit du prunellier » et, par analogie, « pupille de l'œil » ou métonymiquement « yeux, regard » (d'après le *Trésor de la langue française*), est un pur effet de mitoyenneté lexicographique. Mais la simulation des pratiques littéraires, par l'étude des voisinages du mot *« ardent »* dans la base textuelle FRANTEXT, nous apprend que ce qui est ardent, dans la poésie du 19ème siècle, est, par ordre d'importance, l'œil, le soleil, le feu et le cœur, et non les petites prunes (*cf.* Figure 1).

Fréq. abs.	Formes
233	ardent
19	œil
19	soleil
17	feu
14	ton
9	cœur
9	jour
9	souffle
8	peuple
8	vie

Figure 1 : Etude des voisinages de la forme « ardent » dans la base textuelle FRANTEXT, textes non catégorisés. Fonction Voisinage de mot +/10 dans un regroupement de textes, genre « poésie » 1800-1900. Nombre d'occurrences du pivot : 231. Taille des voisinages explorés : 4620 occurrences. Nombre de graphies trouvées dans ces voisinages : 1278.

De fait, ce n'est pas tant le cotexte que le texte lui-même, en tant qu'il relève d'un discours (littéraire), d'un champ générique (la poésie) voire d'un genre (le sonnet italien) et d'une période (le 19ème siècle) qui permet de désambiguïser le mot. À l'inverse, si nous

nous interrogeons sur l'environnement cooccurrentiel de prunelles dans les traités du 20ème siècle, on y rencontrera *« alises »*, *« calvilles »*, *« châtaignes »*, *« cormes »*, *« figues »*, *« nèfles »*, *« noix »*, *« noisettes »*, *« pêches »*, *« poires »*, *« pommes »*, *« raisins »*, lesquels constituent quelques individus d'une classe sémantique des //fruits comestibles// qui n'est pas convoquée par prunelles dans la poésie du 19ème siècle[12].

Ainsi, à la question finalement centrale de l'articulation du sens lexical avec l'extra-linguistique, les notions de *discours** et de *genre** apportent une réponse linguistique semble-t-il crédible.

4. Les structures du lexique défilent dans la rue

Comme on l'a vu ci-dessus, le mot et la phrase constituent les seules unités prises en compte dans le développement de ressources lexico-sémantiques. Il faut y voir l'héritage de la grammaire et de la linguistique traditionnelle. La sémantique lexicale (sur laquelle s'adosse la problématique des ontologies) vise à décrire une sémantique des mots (Cruse 1986, 1995), (Pustejovsky 1995), (Kleiber 1999). La sémantique formelle, d'inspiration logicienne, s'occupe de décrire le sens des phrases (G. Fillmore 1968), (J.-P. Desclés 2004). Or, ces niveaux d'observation sont insuffisants. De même qu'un texte est davantage qu'une suite de mots ou de phrases, son contenu sémantique n'est pas réductible à l'addition des significations des mots ou des phrases qui le composent. Nous distinguons donc deux niveaux d'observation supplémentaires :

- au-delà de la phrase, le texte constitue une unité linguistique indiscutable, car l'interprétation d'un texte est dans une très large mesure conditionnée par le corpus dont il est issu. Le corpus est, en quelque sorte, le contexte du texte (voir Mayaffre, cité *supra*). Il correspond à une pratique discursive et fournit les clés interprétatives nécessaires à la compréhension du texte.

- en deçà du mot, la cohésion sémantique d'un texte est assurée par des réseaux de sèmes qui correspondent à ce que nous appellerons des *objectivations sémantiques**. Lorsque ces réseaux de sèmes sont hérités du corpus, ils fournissent un cadre interprétatif ; lorsqu'ils sont spécifiques au texte, ils le singularisent également.

[12] Étude des voisinages de la forme *prunelles*, dans la base textuelle FRANTEXT, textes non catégorisés. Voisinage de mot +/10 dans un regroupement de textes, genre « traité » 1918-2000. Nombre d'occurrences du pivot : 5 Taille des voisinages explorés : 241 occurrences. Nombre de graphies trouvées dans ces voisinages : 86.

Sans nier complètement l'intérêt des approches syntagmatiques et paradigmatiques présentées ci-dessus, la définition et l'exploitation des réseaux sémiques permettent de décrire aussi bien le contenu sémantique d'un emploi lexical que celui d'un texte et, par conséquent, de problématiser les interactions entre textes et usages lexicaux. En bref, les réseaux sémiques formalisent ce que beaucoup de théories sémantiques invoquent par l'expression généralement peu théorisée de contexte. Nous verrons, dans le chapitre suivant, comment rendre compte de ce contexte, en en proposant une objectivation. Auparavant, étudions quelques unes des propositions de la sémantique textuelle en manière de structuration du lexique.

La question des classes

La sémantique différentielle présente une approche typologique pour la structuration du lexique en termes de classes sémantiques liées à des situations d'usages. Ces classes organisent le lexique de façon différente des approches classiques par prototype (Kleiber 1990) ou hiérarchies (synonymie, hyperonymie, hyponymie, etc.). On, c'est-à-dire (Rastier 1987, 2001), distingue trois types de classes : les dimensions, les domaines et les taxèmes.

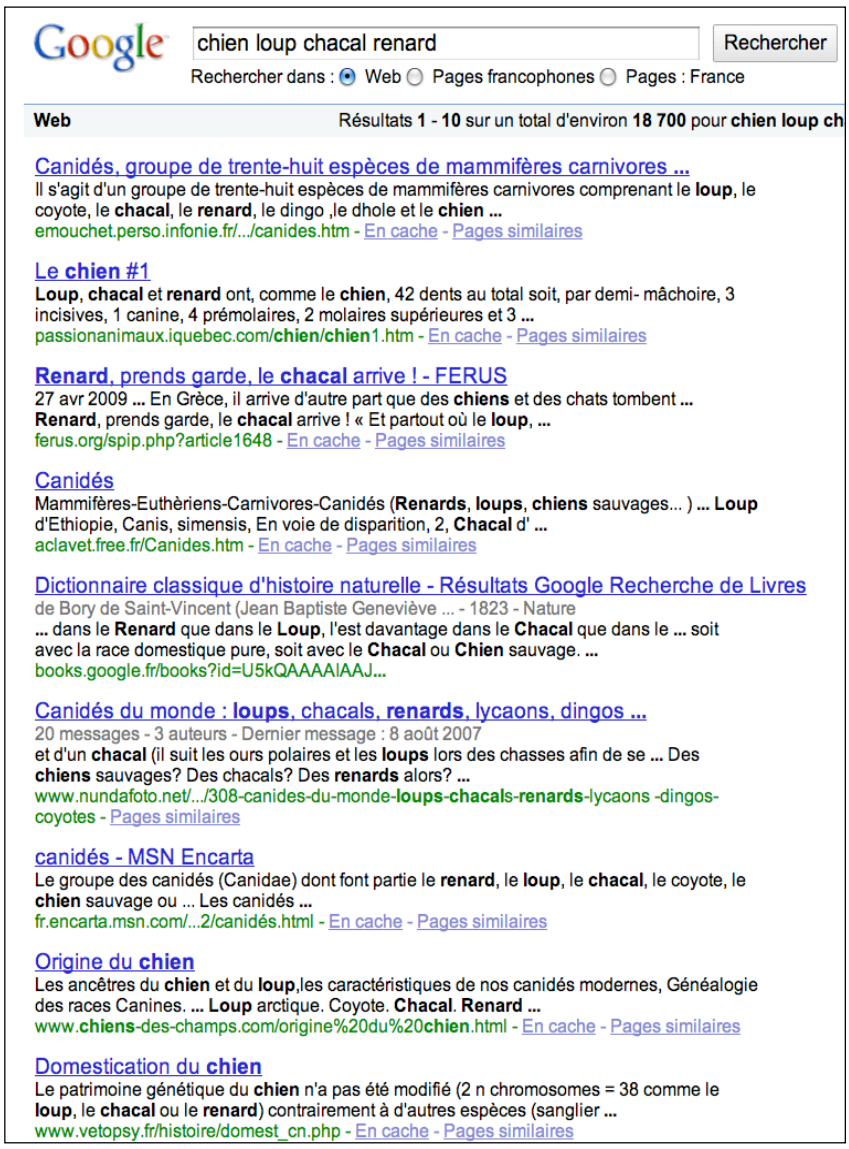
Les *dimensions* sont des classes très générales qui opposent des valeurs telles que /humain/ *vs* /animal/ ou /positif/ *vs* /négatif/, ou encore /animé/ *vs* /inanimé/. Par exemple, *« apporter*$^{/\text{inanimé}/}$ *»* et *« emmener*$^{/\text{animé}/}$ *»* :

> J'ai emmené$^{/\text{animé}/}$ ma souris$^{/\text{animé}/}$ chez le vétérinaire
> J'ai apporté$^{/\text{inanimé}/}$ ma souris$^{/\text{inanimé}/}$ au service après-vente
> J'ai emmené $^{/\text{animé}/}$ mon chien$^{/\text{animé}/}$ chez le vétérinaire.
> J'ai apporté$^{/\text{animé}/}$ mon chien$^{/?/}$ chez le [taxidermiste ?].

Les *taxèmes* constituent la classe la plus originale de la sémantique interprétative. Il s'agit de classes construites en fonction de situations d'usage. Par exemple, le sème /domestique/ intègre *« chien »* dans le taxème des //animaux domestiques//, ou plus précisément, dans celle des //animaux de compagnie//). Il établit ainsi une relation sémantique avec *« chat »*, *« poisson rouge »* ou *« hamster »*.

Pour comprendre l'intérêt de cette représentation taxémique, revenons le temps de quelques lignes à l'ontologie Wordnet. *« Dog »* y est en effet l'hyponyme de *« canis »* et ses *sister terms* sont *« wolf »*, *« jackal »*, *« fox »*, *« hyena »*, etc. Autrement dit, ils dessinent une classe sémantique cohérente, mais restreinte au seul domaine de la zoologie.

Or, la pratique du *dog*, si l'on peut dire, ne se résume pas à la zoologie. On identifie sans peine des classes sémantiques différentes liées à d'autres usages. Le chien est notoirement, comme nous venons de le signaler, un élément de la classe des //animaux de compagnie//, laquelle, outre qu'elle est d'une grande hétérogénéité zoologique, relève d'une autre pratique sociale. Le discours correspondant pourrait être qualifié de philozootique. Une petite expérience aidée d'un moteur de recherche est, à cet égard, très éclairante. Si l'on construit une requête « chien loup chacal renard », on rapatrie essentiellement des pages Web relevant du discours encyclopédique ou zoologique, avec des genres textuels tels que la définition, la monographie, etc. A l'inverse, une requête « chien chat canari hamster, poisson rouge » recrutera des pages Web, qui elles, ressortissent de genres tels que les annonces, des informations d'animalerie, des conseils vétérinaires, etc. (voir figure 2).



Google

chien loup chacal renard Rechercher

Rechercher dans : Web Pages francophones Pages : France

Web Résultats **1** - **10** sur un total d'environ **18 700** pour **chien loup ch**

Canidés, groupe de trente-huit espèces de mammifères carnivores ...
Il s'agit d'un groupe de trente-huit espèces de mammifères carnivores comprenant le **loup**, le coyote, le **chacal**, le **renard**, le dingo ,le dhole et le **chien** ...
emouchet.perso.infonie.fr/.../canides.htm - En cache - Pages similaires

Le **chien** #1
Loup, **chacal** et **renard** ont, comme le **chien**, 42 dents au total soit, par demi- mâchoire, 3 incisives, 1 canine, 4 prémolaires, 2 molaires supérieures et 3 ...
passionanimaux.iquebec.com/**chien**/**chien**1.htm - En cache - Pages similaires

Renard, prends garde, le **chacal** arrive ! - FERUS
27 avr 2009 ... En Grèce, il arrive d'autre part que des **chiens** et des chats tombent ... **Renard**, prends garde, le **chacal** arrive ! « Et partout où le **loup**, ...
ferus.org/spip.php?article1648 - En cache - Pages similaires

Canidés
Mammifères-Euthériens-Carnivores-Canidés (**Renards**, **loups**, **chiens** sauvages...) ... **Loup** d'Ethiopie, Canis, simensis, En voie de disparition, 2, **Chacal** d' ...
aclavet.free.fr/Canides.htm - En cache - Pages similaires

Dictionnaire classique d'histoire naturelle - Résultats Google Recherche de Livres
de Bory de Saint-Vincent (Jean Baptiste Geneviève ... - 1823 - Nature
... dans le **Renard** que dans le **Loup**, l'est davantage dans le **Chacal** que dans le ... soit avec la race domestique pure, soit avec le **Chacal** ou **Chien** sauvage. ...
books.google.fr/books?id=U5kQAAAAIAAJ...

Canidés du monde : **loups**, chacals, **renards**, lycaons, dingos ...
20 messages - 3 auteurs - Dernier message : 8 août 2007
et d'un **chacal** (il suit les ours polaires et les **loups** lors des chasses afin de se ... Des **chiens** sauvages? Des chacals? Des **renards** alors? ...
www.nundafoto.net/.../308-canides-du-monde-**loups**-**chacals**-**renards**-lycaons -dingos-coyotes - Pages similaires

canidés - MSN Encarta
Le groupe des canidés (Canidae) dont font partie le **renard**, le **loup**, le **chacal**, le coyote, le **chien** sauvage ou ... Les canidés ...
fr.encarta.msn.com/...2/canidés.html - En cache - Pages similaires

Origine du **chien**
Les ancêtres du **chien** et du **loup**,les caractéristiques de nos canidés modernes, Généalogie des races Canines. ... **Loup** arctique. Coyote. **Chacal**. **Renard** ...
www.**chiens**-des-champs.com/origine%20du%20**chien**.html - En cache - Pages similaires

Domestication du **chien**
Le patrimoine génétique du **chien** n'a pas été modifié (2 n chromosomes = 38 comme le **loup**, le **chacal** ou le **renard**) contrairement à d'autres espèces (sanglier ...
www.vetopsy.fr/histoire/domest_cn.php - En cache - Pages similaires

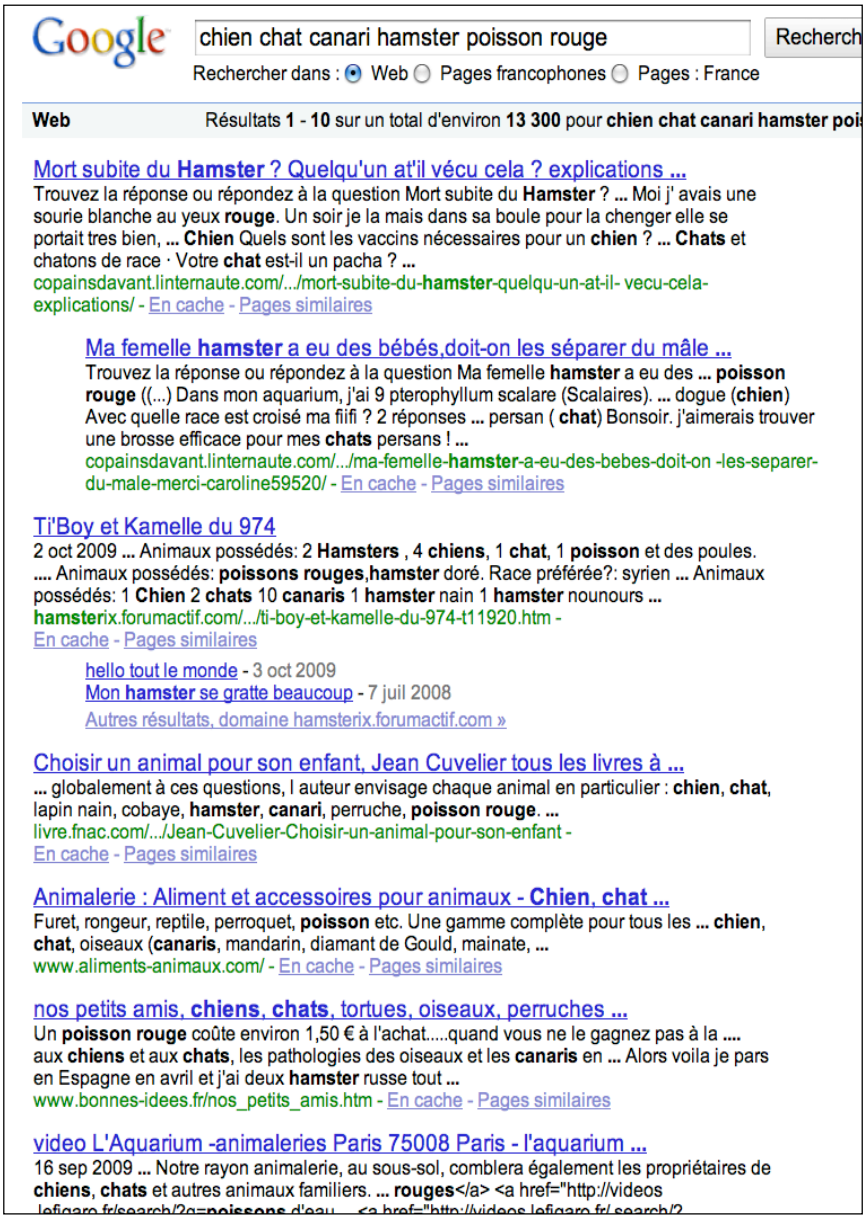


Figure 2 : résultat des requêtes *canidés* et *animaux de compagnie* sur Google (14 10 2009)

Les sèmes qui assurent la cohésion du taxème sont les *sèmes génériques**. Ceux qui au contraire, permettent de distinguer au sein d'une classe sémantique deux individus, sont appelés des *sèmes spécifiques**. Ainsi, dans *« chien »*, /domestique/ est un sème générique dans la classe des //animaux de compagnie//, et un sème spécifique dans la classe des //canidés//. Dans la classe des //canidés//, c'est l'inverse. Le caractère provisoire et idoine des taxèmes en fait tout l'intérêt par rapport aux nomenclatures figées, mais aussi toute la complexité en termes de traitement automatique[13]. Dans les textes, les taxèmes s'articulent ensemble et forment des champs sémantiques.

[13] On lira dans le chapitre 3 des exemples de production automatique de taxèmes.

Quant aux *domaines*, la dernière classe proposée par la sémantique textuelle, il s'agit de regroupements de taxèmes liés à une même pratique sociale. Dans l'exemple ci-dessus, on a vu que le taxème //canidés// relevait du domaine //zoologie//. Le nombre et la richesse des taxèmes dans un domaine donné témoignent du dynamisme de la pratique correspondante. Ainsi, le site philozootique http://vianimo.com, consacré aux chiens et aux chats, aligne en guise de sommaire, une vingtaine de rubriques qui correspondent à des classes sémantiques dont la plupart sont des taxèmes (figure 3).

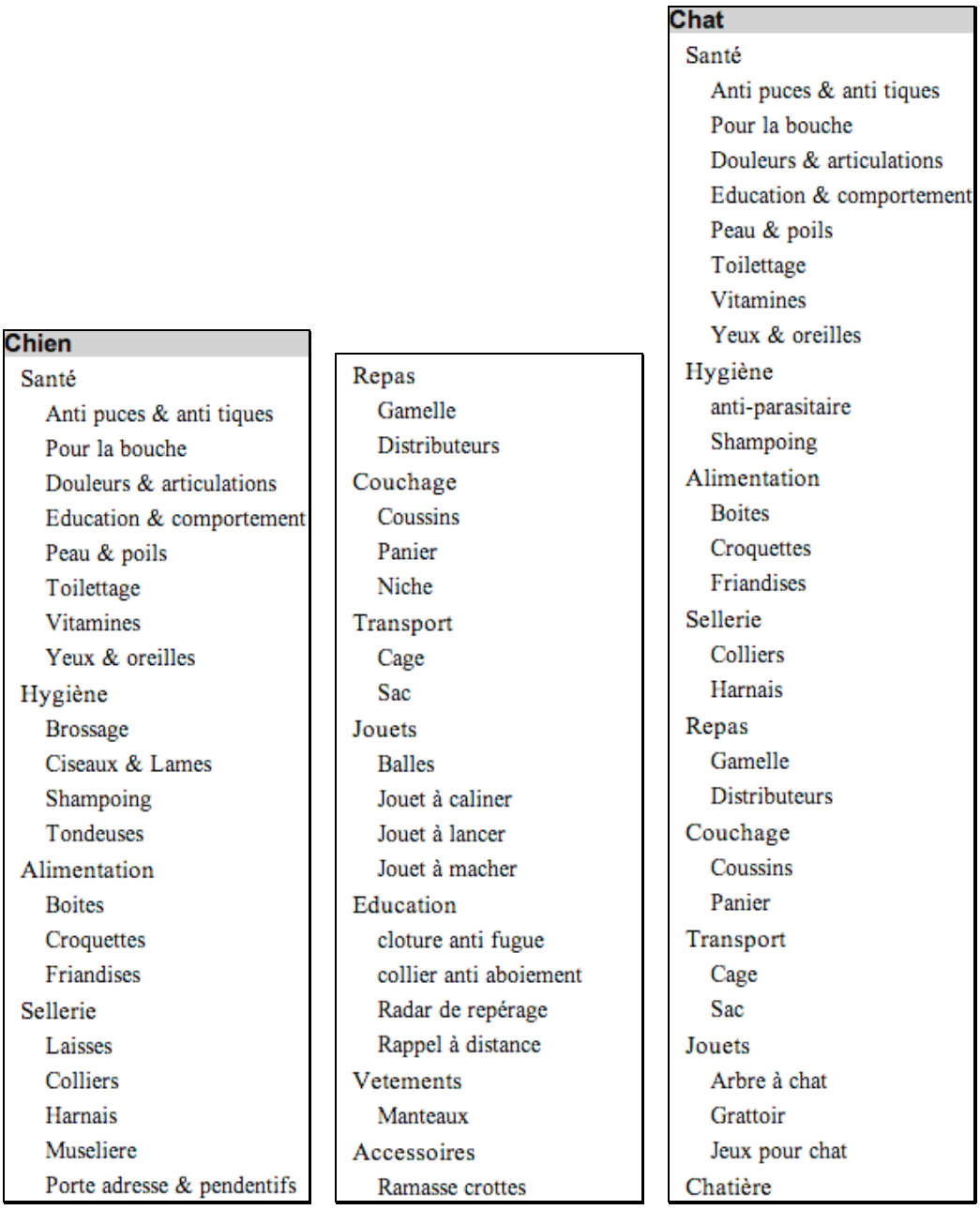


Figure 3 : Menu du site http://vianimo.com

Les taxèmes ont une valeur cognitive indiscutable mais ils sont avant tout pragmatiques et leur idonéité doit sans cesse être questionnée. L'activité langagière consiste dans une large mesure à déployer et partager des classes sémantiques, à les enrichir et à les élaguer, suivant des modalités que nous évoquerons ultérieurement.

Si, dans les dictionnaires, les taxèmes ne sont pas formellement identifiés (ce qui pourrait représenter un enjeu de taille pour la lexicographie), les domaines, quant à eux, le sont davantage, ils sont en général documentés, mais de façon assez hétérogène.

Chapitre 2
Les mots dans les textes

Adoptons un empirisme de méthode : les mots apparaissent dans deux types d'objet matériel ; le texte, objet construit de façon syntagmatique, où le mot est actualisé, dans un état qu'on pourrait qualifier de dynamique et le dictionnaire, objet construit de façon paradigmatique, où le mot est dans un état passif, en attente d'une actualisation. Qu'est-ce qu'un mot, par-delà ces types d'objet ? Il est un signe constitué d'une forme et d'un contenu. La forme est acoustique ou graphique, on l'appelle le signifiant, le contenu est, dans le paradigme structuraliste une collection de propriétés sémantiques plus ou moins articulées entre elles, qui constituent le signifié. Certaines traditions distinguent toutefois le signifié à proprement parler de son contenu, suivant l'opposition langue/parole. Ainsi, sur le modèle du lexème et de la lexie, on distingue le *sémème* en langue et la *sémie* en discours. Cette distinction, en première approximation, outrepasse notre ambition descriptive, même si l'on pourrait assimiler la relation du dictionnaire au texte à l'opposition langue/discours. Pour des raisons de clarté, mais aussi parce que la pratique du TAL rend peu pertinente ladite opposition, nous retiendrons le seul signifié[14]. Les propriétés sémantiques qu'il contient sont des sèmes. Elles sont d'ordre métalinguistique et résultent d'une analyse ou d'une validation humaine effectuées, par exemple, par un linguiste. Cohérente, les collections de sèmes relèvent de la catégorie générale des objectivations sémantiques. On peut en distinguer deux types, selon que l'on se trouve dans une problématique du texte ou dans une problématique du dictionnaire.

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

1. Les signifiés

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

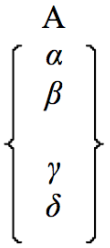


Figure 4 : un signifié composé des sèmes α, β, γ, δ associé à un signifiant A

A désigne le signifiant (c'est-à-dire, en pratique, le mot graphique) et α, β, γ, δ entre accolades désignent le signifié proprement dit, c'est-à-dire les sèmes regroupés. Tous ces traits n'ont pas nécessairement le même poids ni la même valeur sémantique, comme nous l'avons vu dans le chapitre précédent, ce que figure le saut de ligne entre β et γ. Certains sèmes, par exemple, sont génériques d'une classe donnée. Ainsi, le signifié du mot *« chien »* peut être décrit comme la collection de sèmes énumérés dans la partie gauche de la figure 5. Mais dans la mesure où ces sèmes sont, répétons-le, purement métalinguistiques, leur liste n'est ni exhaustive, ni fermée, et ne relève pas d'une quelconque valeur de vérité. Dès lors, le contenu sémantique de *« chien »* construit par un enfant de quatre ans pourrait fort bien correspondre au signifié de droite, toujours sur la figure 6[15].

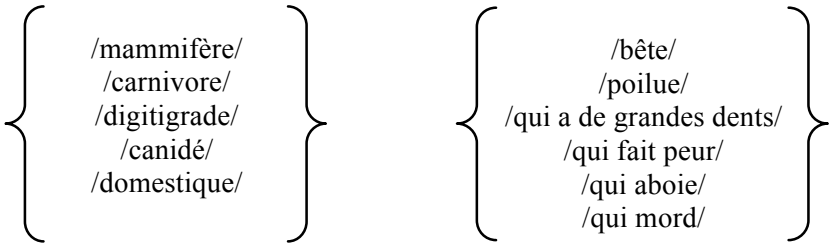


Figure 5 : deux signifiés pour *« chien »*.

Aucun de ces signifiés, qui sous-tendent deux définitions lexicographiques parfaitement distinctes, n'est meilleur qu'un autre. Les propriétés sémantiques diffèrent par leur seul contexte de construction. Le premier est savant, le second se fonde sur une ou plusieurs expériences sensibles ou fantasmées.

[15] Dans le taxème des animaux terrifiants, il est possible que « loup » côtoie le chien à une courte encablure sémique : {/bête/, /poilue/, /qui a de grandes dents/, /qui hurle/, /qui dévore/}.

Un texte est, formellement parlant, un alignement de signes (de mots) suivant des règles de construction syntaxiques. On le représentera ainsi :

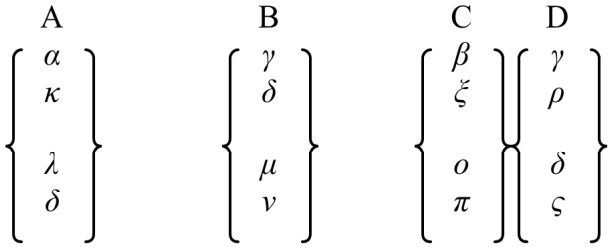


Figure 6 : Un texte (alignement de mots constitués chacun d'un signifiant (A, B, C, D) et d'un signifié)

2. Les réseaux sémiques

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

2.1. Les fonds sémantiques (isotopies)

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

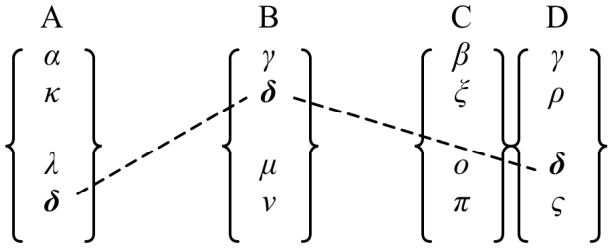


Figure 7 : Une isotopie (fond sémantique) (récurrence du trait sémantique δ dans le texte)

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

> Violence urbaine/ville/ : l'expression est née récemment, il y a plus ou moins dix ans. Une expression qui s'applique plus exactement à certaines villes/ville/ de banlieue/ville/, celles qui rassemblent un grand nombre d'habitants d'origine étrangère, au point d'être devenues de véritables ghettos/ville/16.

Minimalement, cela signifie que ce texte traite de ville. Mais les isotopies peuvent relever d'une description plus fine. On peut distinguer par exemple, dans un premier temps, des isotopies domaniales, par exemple, dans le texte suivant, on observe une isotopie domaniale /botanique/ passablement dense.

> En effet, comme il avait cultivé les unes près des autres des espèces différentes, les sucrins s´étaient confondus avec les maraîchers, le gros Portugal avec le grand Mogol – et le voisinage des pommes d´amour/bot/ complétant l´anarchie, il en était résulté d´abominables mulets qui avaient le goût de citrouilles.
> Alors Pécuchet se tourna vers les fleurs/bot/. Il écrivit à Dumouchel pour avoir des arbustes/bot/ avec des graines/bot/, acheta une provision de terre de bruyère et se mit à l´œuvre résolument. Mais il planta des passiflores/bot/ à l´ombre, des pensées/bot/ au soleil, couvrit de fumier les jacinthes/bot/, arrosa les lys/bot/ après leur floraison, détruisit les rhododendrons/bot/ par des excès d´abattage, stimula les fuchsias/bot/ avec de la colle forte, et rôtit un grenadier, en l´exposant au feu dans la cuisine. Aux approches du froid, il abrita les églantiers/bot/ sous des dômes de papier fort enduits de chandelle ; cela faisait comme des pains de sucre, tenus en l´air par des bâtons. Les tuteurs des dahlias/bot/ étaient gigantesques ; – et on apercevait, entre ces lignes droites les rameaux tortueux d´un sophora/bot/-japonica qui demeurait immuable, sans dépérir, ni sans pousser. Cependant, puisque les arbres/bot/ les plus rares prospèrent dans les jardins de la capitale, ils devaient réussir à Chavignolles ? et Pécuchet se procura le lilas des Indes, /bot/ la rose de Chine/bot/ et l´Eucalyptus/bot/, alors dans la primeur de sa réputation. Toutes les expériences ratèrent. (G. Flaubert, *Bouvard et Pécuchet)*

Dans le fragment ci-dessous, s'actualise une isotopie taxémique (isotopie *microgénérique* dans la terminologie rastiérienne), à partir de la classe sémantique maintenant bien connue du lecteur des //animaux de compagnie//

> Des sociétés ou associations mettent à votre disposition une équipe d'assistants chargés de venir faire une ou plusieurs visites régulières à votre domicile, pour nourrir votre chien/dom/, chat/dom/, canari/dom/, ou hamster/dom/, lui prodiguer des soins adaptés (sorties, jeux, câlins) et lui apporter une présence rassurante.[17]

[16] http://www.lacathode.org/cqfs/viol.htm
[17] http://www.linternaute.com/acheter/depart-vacances/animaux/animaux.shtml

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

En bref, l'isotopie constitue le socle du parcours interprétatif. Elle permet d'une part, de zoner les textes et d'autre part, de les inscrire dans l'intertexte.

Zonage isotopique

Le *zonage isotopique** est la manifestation spatiale d'un faisceau d'isotopies. Le zonage participe à l'identification de passages. Si l'on reprend l'exemple de *Bouvard et Pécuchet*, on est en mesure de distinguer trois zones globalement consécutives qui correspondent à trois périodes dans la narration des errances jardinières de Pécuchet :

> En effet, comme il avait cultivé les unes près des autres des espèces différentes, les sucrins s´étaient confondus avec les maraîchers, le gros Portugal avec le grand Mogol – et le voisinage des pommes d´amour/com/ complétant l´anarchie, il en était résulté d´abominables mulets qui avaient le goût de citrouilles.
> Alors Pécuchet se tourna vers les fleurs/orn/. Il écrivit à Dumouchel pour avoir des arbustes avec des graines, acheta une provision de terre de bruyère et se mit à l´oeuvre résolument.
> Mais il planta des passiflores/com/ à l´ombre, des pensées au soleil, couvrit de fumier les jacinthes, arrosa les lys/orn/ après leur floraison, détruisit les rhododendrons/orn/ par des excès d´abattage, stimula les fuchsias/orn/ avec de la colle forte, et rôtit un grenadier, en l´exposant au feu dans la cuisine.
> Aux approches du froid, il abrita les églantiers sous des dômes de papier fort enduits de chandelle ; cela faisait comme des pains de sucre, tenus en l´air par des bâtons. Les tuteurs des dahlias/orn/ étaient gigantesques ; – et on apercevait, entre ces lignes droites les rameaux tortueux d´un sophora/orn/-japonica qui demeurait immuable, sans dépérir, ni sans pousser.
> Cependant, puisque les arbres les plus rares prospèrent dans les jardins de la capitale, ils devaient réussir à Chavignolles ? et Pécuchet se procura le lilas des Indes/odo/, la rose de Chine/odo/ et l´Eucalyptus/odo/, alors dans la primeur de sa réputation. Toutes les expériences ratèrent.

La première période correspond à la brève isotopie /comestible/ (*« pomme d'amour »*, *« passiflores »* auxquels on pourrait ajouter *« citrouilles »*)[18], la seconde période à

[18] Le lecteur peut s'étonner de quelques partis-pris en matière d'annotation. Par exemple, pourquoi *« passiflore »* est-il étiqueté comme plante comestible quand elle semble avant tout esthétique et que le

l'isotopie /ornemental/, qui couvre avec une grande régularité les deux alinéas suivants. Enfin, une troisième zone isotopique est identifiable par le sème /odorant/

Certes les isotopies sont insuffisantes à segmenter le texte, mais comme souvent dans l'approche hétérarchique que propose la sémantique textuelle, il s'agit d'indices corrélés à d'autres clés interprétatives. Par exemple, les changements de zones sont accompagnés de petites isotopies de la destruction ou de la ruine.

Insertion intertextuelle

Minimalement, les isotopies intertextuelles organisent la cohésion d'un corpus (par exemple, l'ensemble des textes où il est question de botanique, ou d'animaux domestiques). Mais elles peuvent également avoir une fonction d'évocation en référence à l'intertexte. Par exemple, le roman d'Annie Ernaux, *Les années* (Gallimard, 2008) comprend un prologue constitué d'une suite de courts textes, fragmentaires et elliptiques, comme des réminiscences chaotiques des années passées qui feront l'objet de la narration proprement dite. L'un des fragments est une réécriture/contraction d'un passage de la *Recherche du temps perdu*, comme « un cité de mémoire ». Nous reproduisons ci-dessous le cité de mémoire et le texte de Proust correspondant :

> Notre mémoire est hors de nous, dans un souffle pluvieux du temps (*Les années*, p. 17)

> La meilleure part de notre mémoire est hors de nous, dans un souffle pluvieux, dans l'odeur de renfermé d'une chambre ou dans l'odeur d'une première flambée, partout où nous retrouvons de nous-mêmes ce que notre intelligence, n'en ayant pas d'emploi, avait dédaigné, la dernière réserve du passé, la meilleure, celle qui, quand toutes nos larmes semblent taries, sait nous faire pleurer encore. (*A l'ombre des jeunes filles en fleur*, p. 643[19])

En ajoutant « du temps » à la suite du souffle pluvieux, Ernaux restitue le fond sémantique du passage original. Ainsi, le sème /temps/ est actualisé deux fois ; une première dans « notre mémoire » (/temps/$_1$), une seconde dans « temps » (/temps/$_2$). Le passage est ainsi enclos dans une isotopie qui assure à la fois la cohésion du passage ernalien et son intertextualité. Cependant, elle fait subir au sens du syntagme « souffle pluvieux » de *La recherche* une altération remarquable, car de concret (« souffle pluvieux$^{/\text{atmosphérique}/}$, odeur

zonage ornemental est enclenché. C'est parce que notre parti-pris est d'annoter ce texte au moyen des seuls sèmes explicitement identifiables dans les définitions du *TLF*i, simulant ainsi le procédé TAL qui sera décrit dans le chapitre 3.

[19] Cité ici d'après l'édition Gallimard 1962.

de renfermé »), il devient métaphorique (« souffle pluvieux du temps/duratif/ »), vraisemblablement parce que les deux signifiés, le temps duratif et le temps atmosphérique, partagent le même signifiant. Cette transformation, outre qu'elle restitue le fond sémantique (l'isotopie) est une façon d'en abstraire le sens : c'est toute *La recherche* qui est ainsi concentrée dans ce « temps /duratif/ ».

2.2. Les formes sémantiques

Plusieurs sèmes distincts peuvent être instanciés ensemble dans des textes différents avec une certaine régularité. Ces groupements s'appellent des thèmes, ou des *formes sémantiques**. On peut les représenter de la façon suivante :

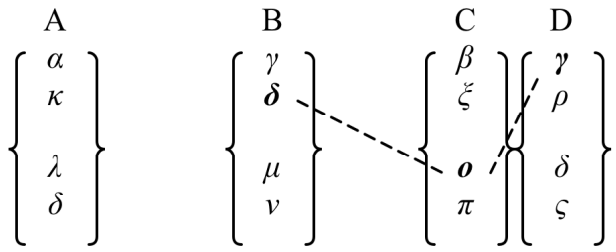


Figure 8 : une forme sémantique
(groupement stabilisé des sèmes δ, o et γ dans différents textes)

Par exemple dans les textes suivants :

> Nicolas Sarkozy et les émeutiers/émeute/ de banlieues/ville/ sont le produit d'un même terreau[20]

> Prenez un patron de CRS [...] qui gère une échauffourée/émeute/ dans un quartier/banlieue/, il sait qu'il doit laisser les jeunes gredins s'escrimer sur les vitrines[21]

Les sèmes /émeute/ et /ville/ sont en cooccurrence rapprochée. Ils constituent une petite forme sémantique, c'est-à-dire une unité sémantique dont le signifiant est discontinu et surtout variable et qu'on pourrait assimiler, par exemple, à l'unité lexicale « émeute urbaine » :

> Les émeutes urbaines{/émeute//ville/} sont récurrentes en France depuis le début des années 80. Les premières émeutes/émeute/ ont lieu en 1979, à Vaulx-en-Velin/ville/[22]

[20] http://www.voltairenet.org/article130994.html
[21] http://www.forumdesforums.com/modules/news/article.php?storyid=14613
[22] http://www.babnet.net/cadredetail-3296.asp

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

> Le citoyen est libre$^{/liberté/}$ de fumer$^{/fumer/}$ ou de ne pas fumer, de manger de la salade si ça lui chante et des rillettes s'il en a envie[23]
>
> Ils ont la liberté de fumer$^{\{/liberté//fumer/\}}$, soit, je serai le dernier à la leur retirer, sauf dans les endroits publics[24].

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

> Notre métier ne consiste pas à inciter des gens à fumer$^{/fumer/}$. Il consiste à offrir des marques de qualité à des adultes qui ont déjà pris la décision$^{/liberté/}$ de fumer$^{/fumer/}$ […]. C'est pourquoi nous sommes convaincus que fumer$^{/fumer/}$ devrait être le seul fait d'adultes conscients des risques de fumer$^{/fumer/}$. (JTI)
>
> Fumer$^{/fumer/}$ repose sur une décision$^{/liberté/}$ individuelle qui ne peut être que le fait d'adultes informés des risques liés au tabagisme (ALTADIS)
>
> Nous sommes convaincus que le choix$^{/liberté/}$ de fumer$^{/fumer/}$ doit être le choix d'adultes avertis et conscients des risques, un choix qui exclut de fait les jeunes, non adultes. (BAT)
>
> Nous nous engageons ainsi à communiquer de manière responsable avec les adultes qui ont délibérément choisi$^{/liberté/}$ de fumer$^{/fumer/}$. (ALTADIS)
>
> JTI s'engage à fabriquer des cigarettes$^{/fumer/}$ de qualité pour les adultes qui choisissent$^{/liberté/}$ de fumer$^{/fumer/}$ par plaisir. (JTI)[25]

De la sorte, les industriels du tabac construisent une nouvelle forme sémantique {/liberté//fumer/adulte/}. Toutefois, *« adultes »* subit un certain nombre de modalités valorisantes, le plus souvent sous la forme d'adjectifs qualitatifs *(« conscients »*,

[23] http://www.le-tigre.net/

[24] http://www.philo5.com/

[25] Ces exemples sont issus des recherches de l'équipe linguistique (ERTIM) du projet C-MANTIC.

« informés », *« avertis »*, etc.) de sorte que se construit un parcours interprétatif liberté de fumer + adulte/valorisant/. Ce discours de valorisation de l'adulte libre de fumer, autrement dit, de l'adulte fumeur, induit une lecture spéculaire telle que le jeune, non adulte est dévalorisé. C'est explicite dans le troisième exemple (« un choix qui exclut de fait les jeunes, non adultes »). On peut donc construire les deux formes sémantiques ci-dessous :

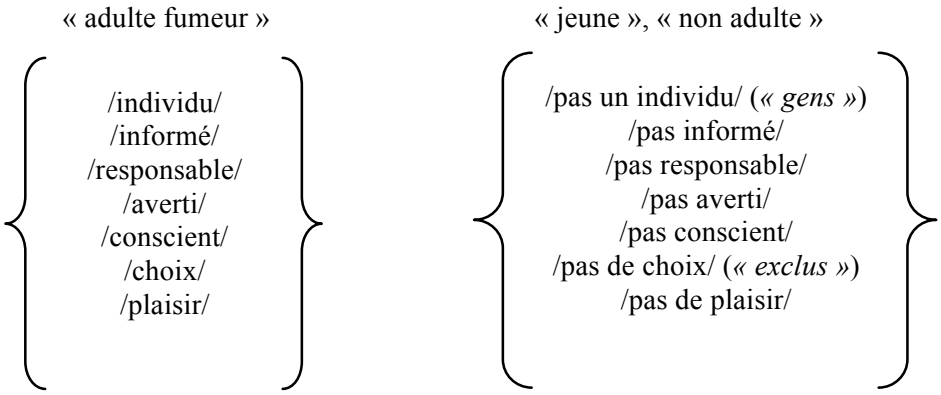


Figure 9 : formes sémantique d'*« adulte »* et de *« non adulte »* dans les textes de l'industrie du tabac

3. Interpréter, une négociation entre signifiés et formes sémantiques

On l'a vu, le sens d'une unité lexicale dépend autant du contexte dans lequel il apparaît que de sa définition première. Ainsi, on distingue parfois la signification (du dictionnaire) et le sens (dans le texte). En complément de l'opposition sèmes génériques *vs* sèmes spécifiques (qui permet d'introduire la notion de classe sémantique), la sémantique interprétative opère une distinction entre des sèmes inhérents et des sèmes afférents.

Les sèmes inhérents peuvent être considérés comme plus « définitoires » (par exemple, pour *« chien »*, on aura : /qui aboie/, /canidé/, /quadrupède/, etc.) que les sèmes afférents qui relèvent de l'usage qui est fait du mot dans les textes. Les sèmes afférents sont donc issus des contextes d'énonciation. C'est le cas dans l'exemple de la figure 9 où le signifié d'*adulte* n'est représenté ici qu'avec des sèmes afférents. On retiendra qu'il s'agit de sèmes hérités d'un ensemble de contextes et susceptibles d'être recontextualisés à l'identique.

Par exemple, le sème /pauvre/ est fréquemment actualisé dans *« population »* lorsqu'il est cooccurrent du lexème *« banlieue »*. Mais il ne s'agit pas d'un sème inhérent : rien dans la banlieue ne la prédispose à accueillir de façon privilégiée une population pauvre. Par

exemple, l'expression *« jeunes des banlieues »*, dans le discours journalistique (qui est aujourd'hui largement prescriptif) ou dans le discours politique (qui lui ressemble) ne signifie pas tous les jeunes résidant en banlieue, mais certains jeunes, défavorisés, résidant en banlieue. C'est un sème afférent, « subjectif » ou « socialement normé », c'est-à-dire circonscrit d'un point de vue historique, géographique et socioculturel. Aux Etats-Unis, l'opposition est inverse : les pauvres vivent en centre-ville. En revanche, lorsque « banlieue » s'intègre dans certaines lexies composées telles que « *banlieue de l'ouest parisien »*, le sème /pauvreté/ est inhibé par le sème /bourgeois/, afférent à *« ouest parisien »*, quand bien même la banlieue ouest de Paris est tout aussi hétérogène que la banlieue dans son ensemble.

Retenons l'idée que les sèmes contenus dans un signifié ne sont pas tous égaux. Leur nature et leur poids varient en synchronie (tous les sèmes n'ont pas la même valeur dans le signifié) comme en diachronie (un même sème peut évoluer dans le temps, changer de poids, disparaître, etc.). En bref, les sèmes oscillent entre stabilité et instabilité. Le signifié est constitué de sèmes résistant à la variation, sinon permanents et de sèmes instables. Cette variabilité résulte de l'enrichissement (ou de l'appauvrissement) du signifié à mesure que le mot est actualisé dans les textes, et notamment à mesure qu'il participe à des formes sémantiques. 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. En somme, on pourrait dire, pastichant une formule de (Rastier 2001, 92), que tout mot placé dans un texte en reçoit des déterminations sémantiques, et modifie potentiellement le signifié de chacun des mots qui le composent.

On posera que le sens résulte de négociations entre des signifiés et des formes sémantiques. Nous faisons l'hypothèse que, de la même façon que le cortex visuel traite moins d'informations issues du nerf optique que d'information stockée en mémoire, l'interprétation résulte autant – sinon davantage – d'une « effervescence sémique » (reconfiguration du signifié, négociation, convocation des afférences possibles en mémoire, etc.) que du texte lui-même.

Chapitre 3
Un dictionnaire sémique pour l'analyse textuelle du lexique

Les lexiques sémantiques généralistes décrits précédemment (WordNet, FrameNet, etc.) reposent sur l'hypothèse qu'il existe une « langue générale » opposée aux langues de spécialité (objets de la terminologie). À l'opposé, l'approche textuelle pose qu'il n'existe pas de langue générale, mais des pratiques sociales variées (la médecine, le droit, le journalisme, etc.) auxquelles correspondent des discours (discours scientifique, discours juridique, discours médiatique, etc.), des genres textuels (l'article, le compte-rendu, etc.) et des domaines (pédiatrie, droit constitutionnel, économie, etc.) (Adam 1999, Rastier 2001). La notion de ressource sémantique généraliste s'avère donc insuffisante dans cette approche de la langue dans ses usages.

Toutefois, le TAL a nécessairement recours à des outils et à des ressources (Véronis 2004, Habert 2005). La mise en œuvre d'un dispositif expérimental de sémantique textuelle fait l'objet de ce chapitre. Nous y évoquons la réalisation d'une ressource sémantique pour l'annotation de corpus et la mise en place des méthodologies l'accompagnant.

1. Lexiques sémantiques généralistes ou particuliers ?

La légitimité des lexiques sémantiques est logiquement soumise à leurs usages, et ceux-ci ne sont pas toujours faciles à appréhender. Faisant figure d'exception de par sa notoriété, et en dépit de ses innombrables défauts, WordNet est relativement à l'abri des critiques concernant sa finalité. Sa transposition EuroWordNet, quoique plus sophistiquée, est cependant contestée (Slodzian 1999). Dès lors, les projets de développement ou de

transposition des ressources (anglo-saxonnes) en d'autres langues comme le français, tels que FR.FrameNet (Fillmore *et al.* 2002)[26], posent des problèmes non triviaux d'objectifs. Il n'est pas impossible que la réalisation d'un lexique à large couverture inspirée d'une théorie vise en premier lieu à équiper ladite théorie d'une ressource idoine, c'est-à-dire à se donner les moyens de la tester et de la valider. C'est le cas par exemple de FrameNet et de la sémantique des cadres de (Fillmore 1976), du DiCo et de la théorie Sens-Texte de (Mel'čuk, Clas & Polguère 1995) (Polguère 2000), des classes d'objets et du Lexique-Grammaire construits à la suite des travaux de (Gross 1975, 2005). De fait, la réalisation d'un lexique à large couverture est une entreprise ambitieuse qui n'est pas forcément initiée par les « usagers » des ressources – sinon les lexicographes, mais plutôt par les promoteurs des théories. Les « usagers », en général, ont des besoins ponctuels ; soit ils empruntent les ressources existantes et s'en satisfont peu ou prou, soit ils en produisent eux-mêmes des lacunaires et néanmoins suffisantes. Qu'une application particulière mobilise un lexique « général » apparaît difficilement soutenable : une tâche relève d'un domaine et d'un usage déterminé auquel correspond un discours spécifique. N'est alors requis que le seul lexique particulier dudit discours. Pragmatique, la terminologie textuelle soumet la constitution de ressources aux sources, c'est-à-dire aux corpus de textes. (Bourigault & Slodzian 2000 : 30) expliquent qu'« étant donné un domaine d'activité, il n'y a pas une terminologie, qui représenterait le savoir sur le domaine, mais autant de terminologies que d'applications dans lesquelles ces terminologies ont été utilisées ». D'une certaine façon, cette proposition oppose aux lexiques généralistes un principe de réalité textuelle qui dépasse largement les fenêtres contextuelles, quelle que soit leur largeur. Les propositions d'une « sémantique légère » (Perlerin 2004) vont dans ce sens. En cela, les projets dont l'ambition affichée est lexicographique comme FrameNet ou le DiCo, peuvent difficilement prétendre à des applications autres que lexicographiques.

Nous évoquerons ici la réalisation en cours d'un lexique sémantique alternatif aux approches décrites dans le chapitre 1, élaboré à partir des critiques que nous avons formulées à cette occasion. Du point de vue paradigmatique, nous proposons d'envisager les relations entre unités lexicales non pas en termes de construction hiérarchique mais en termes de classes sémantiques dont la cohésion est assurée, comme nous l'avons vu, par des sèmes. Du point de vue syntagmatique, l'instanciation des unités lexicales dans les textes ne se fait pas au niveau de la phrase ou de l'énoncé, mais au niveau des réseaux sémiques.

[26] http://libresource.inria.fr/projects/framenet/

La ressource lexico-sémantique en cours de développement s'apparente donc à un dictionnaire sémique[27]. A termes, il s'agira d'une collection de dictionnaires sémiques relevant au minimum de discours particuliers. La ressource (que nous avons parfois nommé DIXEM), est conçue en premier lieu à des fins analytiques et prospectives : notre objectif est d'étudier l'économie générale du contenu sémantique des unités lexicales dans des corpus de textes structurés, dans une perspective tant diachronique que synchronique. Les applications TAL et ingénieriques (Recherche d'Information, etc.) ne sont pas négligées mais n'entrent pas dans la problématique de ce mémoire.

2. Constitution du dictionnaire sémique et éléments de structuration

Nous avons transformé le Trésor de la Langue Française informatisé (Dendien & Pierrel 2003), désormais *TLF* pour la réalisation du dictionnaire sémique. Le *TLF* est doté de 100 000 mots et de 270 000 définitions. L'extraction fut réalisée suivant une hypothèse minimaliste et robuste : une définition est un signifié mis en texte[28]. Ainsi, des définitions, nous retenons les lemmes des substantifs, adjectifs, verbes et de certains adverbes. Il constitue les candidats-sèmes* qui constituent le signifié d'une unité lexicale en attente d'actualisation. Ce que nous appelons ici et ultérieurement candidat-sème est l'étiquette sémantique résultant d'un traitement automatique. En validant le candidat-sème, le linguiste peut lui octroyer un statut de sème (pour une définition plus précise, *cf.* le glossaire).

Par exemple, pour une définition telle que :

> LAURACÉES. Famille de plantes dicotylédones, comprenant des arbres et des arbustes, à feuilles simples, alternes et persistantes, qui croissent dans les régions chaudes et tempérées.

Nous extrayons le signifié suivant :

> LAURACÉES {/famille/, /plante/, /dicotylédone/, /comprendre/, /arbre/, /arbuste/, /feuille/, /simple/, /alterne/, /persistant/, /croître/, /région/, /chaud/, /tempéré/}

Hélas, trois fois hélas, « comprendre un énoncé, c´est oublier le plus vite possible une grande partie de l´information sémique » nous enseigne (Pottier 1987 [1992], 82). C'est à ce titre qu'une part non négligeable des efforts produits dans le cadre de ce projet consiste

[27] Lire Pincemin 1999 sur l'exploitation du concept de sème en TAL.
[28] Lire (Rastier 1987), 41 ; (Martin 2001).

à éliminer des candidats-sèmes, tant au niveau paradigmatique de la ressource – il s'agit alors d'éliminer du bruit) que syntagmatique de l'annotation de corpus (il s'agit de sélectionner les bons candidats pour leur allouer le statut de sèmes – et on souhaite cette sélection la plus automatique possible). Beaucoup des recherches rapportées ici visent à ouvrir des pistes pour l'organisation et l'élagage massif des signifiés dans le dictionnaire et dans les textes.

Un travail de normalisation a par exemple été effectué par (Ramdani 2007). Il s'est agi d'une part de regroupements des candidats-sèmes morphologiquement apparentés sous une étiquette unique et d'autre part, d'identifier certaines parties des structures formelles des définitions et du genre lexicographique de manière à repérer certains types de candidats-sèmes, pour les neutraliser ou pour leur allouer des qualités particulières (sèmes dimensionnels par exemple). Ce travail n'a toutefois pas encore été intégré dans les ressources opérationnelles.

2.1. Construction de classes sémantiques

Les premiers travaux (Valette, Estacio *et al.* 2006) ont porté sur la constitution de taxèmes au sein d'un domaine donné et sur la structuration des signifiés en fonction des classes obtenues. Ils ont montré que la réalisation locale de classes sémantiques à partir des définitions dictionnairiques était possible, même si elle ne pouvait se substituer à un apprentissage sur corpus.

Une étude sur le sème /arbre/ nous a en effet permis de distinguer, par Classification Ascendante Hiérarchique (CAH), des sous-classes pertinentes, tant d'un point de vue gnosique (Essences d'arbre, Parasites) que praxique (Plantation, Bûcheronnage, Arboriculture) sans que de telles classes n'aient été dessinées a priori, ni par nous, ni par les lexicographes du *TLF* (voir Figure 10).

Puis, nous avons procédé à une pondération interne des signifiés à l'aide d'un calcul d'écart réduit à l'intérieur d'une classe, celle des Essences d'arbres (la plus importante). Pour ce faire, nous avons choisi de pondérer tous les signifiés de façon à en dégager l'organisation interne relative à la classe considérée. L'objectif est de dégager les sèmes génériques de chaque classe et les sèmes spécifiques des éléments des différentes classes. Pour arriver à nos fins, nous utilisons un calcul classique en lexicométrie : l'écart réduit. Dans ce contexte l'écart réduit se calcule ainsi :

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

Où f_E est la fréquence du sème observée dans l'entrée, f_c la fréquence du sème observée dans la classe, p est la proportion de l'entrée dans la classe et q le complément de p[29].

Classe	Fréq.	Pourcent.	Contenu de la classe	Exemples
1	105	29,33	Essences d'arbres	lauracées, rutacées, ramboutan, etc.
2	94	26,26	Plantation (techniques et outils)	racinage, scarifier, ombrophile, etc.
3	17	4,75	Bûcheronnage	lardoire, solivage, lambourde, etc.
4	16	4,47	Parasites de l'arbre	miastor, rhynchite, processionnaire, etc.
5	15	4,19	Arboriculture	sarter, palmette, provin, palissage, etc.

Figure 10 : Caractérisation de la classification obtenue pour les cinq classes les plus importantes (extrait de Valette, Estacio-Moreno *et al.* 2006)

La pondération des signifiés de chaque entrée à l'aide de l'écart réduit de la sous-classe des Essences d'arbres donne des résultats globalement satisfaisants. Pour chaque signifié, les candidats-sèmes génériques (c'est-à-dire ceux dont l'écart réduit est le plus bas) permettent de distinguer des sous-classes potentielles. Par exemple, la lexie « sycomore » (figure 12) semble appartenir à la sous-classe des essences qui valent pour leur bois et la lexie *« monbin »* (figure 14) à celle valant pour leurs fruits.

Cette organisation du signifié a une incidence déterminante sur les candidats-sèmes spécifiques, dans la mesure où ceux-ci sont spécifiques par rapport aux éléments de la sous-classe esquissée par les candidats-sèmes génériques. Ainsi, /léger/ et /imputrescible/ caractérisent le bois du sycomore, tandis que /saveur/, agréable/, /citron/ renvoient au fruit du monbin. De la même façon, si /famille/ est le candidat-sème générique commun à *« sophora »* (figure 11) et *« manglier »* (figure 13), leurs sèmes spécifiques caractériseront les arbres en tant qu'ils appartiennent à des familles (identifiables dans le signifié : légumineux dans un cas, rhizophoracées dans l'autre). Autrement dit, leur propriété remarquable est d'être des arbres et non de donner des fruits ou de produire un bois

[29] Pour l'entrée i de la classe C_K, $p_{i,K} = \frac{\sum_{l=1}^{|V|} E_i^l}{N_K}$ et $q_{i,K} = 1 - p_{i,K}$. NB : dans cette étude, les travaux statistiques ont été réalisés par A. Estacio-Moreno.

particulier. De fait, le sophora est caractérisé par sa fonction ornementale (sèmes spécifiques /jardin/, /avenue/, /ornement/) et le manglier par les endroits où il croît (sèmes spécifiques /plage/, /lagune/, /maritime/).

Au vu des résultats obtenus, la réalisation préliminaire de classes sémantiques à partir de définitions dictionnairiques semble possible. Les items retenus qui composent une définition peuvent être légitimement considérés comme des sèmes minimaux et ce, malgré une segmentation sommaire et la perte sensible d'information que l'éclatement syntaxique induit. La classification automatique permet de distinguer des sous-classes pertinentes, tant d'un point de vue gnosique (Essences d'arbres, Parasites) que praxique (Plantation, Bûcheronnage, Arboriculture) sans que de telles classes n'aient été dessinées a priori, ni par nous, ni par les lexicographes du *TLF*.

Certes, il faut reconnaître que le domaine choisi, la sylviculture, est exceptionnellement bien structuré et correspond à un savoir encyclopédique davantage que praxéologique. Mais la caractérisation des signifiés à l'intérieur d'une classe donne à voir une organisation sémantiquement pertinente, où les sèmes spécifiques sont susceptibles d'être opérationnels, notamment dans la perspective textuelle qui est la nôtre. En effet, les sèmes /jardin/, /avenue/ et /ornement/ semblent plus caractérisants du sophora que le fait qu'il s'agisse d'un arbre exotique, ou encore qu'il appartienne à la famille des légumineuses (figure 11). En d'autres termes, le sens apparaît valorisé au détriment de la référence.

L'étude présentée ici avait, rappelons-le, une visée exploratoire. Il s'agissait de déterminer la pertinence d'un recours à un dictionnaire de langue dans l'élaboration d'un dictionnaire de sèmes initial destinée à l'annotation de corpus. Les résultats se sont avérés positifs mais des apprentissages sur corpus ciblés, en faisant apparaître la régularité des réseaux sémiques, permettront de stabiliser les signifiés. Cette seconde phase aura notamment pour effet de valider ou mettre à jour les résultats de notre classification préalable[30].

[30] Travaux actuellement réalisés par M. Grzesitchak.

Candidat	Ecart réduit
jardin	20,84
avenue	20,84
ornement	10,35
légumineux	9,23
exotique	6,14
famille	1,84

Figure 11 : Signifié pondéré de *« sophora »*.Définition : Arbre exotique de la famille des Légumineuses, servant à l'ornement des jardins et des avenues (extrait de Valette, Estacio-Moreno *et al.* 2006)

Candidat	Ecart réduit
figuier	20,84
léger	11,98
imputrescible	11,98
érable	11,98
bois	3,09

Figure 12 : Signifié pondéré de *« sycomore »*. Définition : 1. Figuier au bois léger et imputrescible. 2. Érable. 3. [P. méton.] Bois de l'un de ces arbres (extrait de Valette, Estacio-Moreno *et al.* 2006).

Candidat	Ecart réduit
plage	18,04
lagune	18,04
maritime	18,04
intertropical	12,72
rhizophoracées	12,72
croître	4,03
région	2,88
famille	1,46

Figure 13 : Signifié pondéré de *« manglier »*. Définition : Arbre de la famille des Rhizophoracées, qui croît dans les lagunes et les plages maritimes des régions intertropicales (extrait de Valette, Estacio-Moreno *et al.* 2006).

Candidat	Ecart réduit
saveur	11,98
agréable	11,98
citron	11,98
pulpe	7,50
renfermer	6,82
taille	5,86
exotique	4,95
région	2,68
fruit	1,99

Figure 14 : Signifié pondéré de *« monbin »*. Définition : Arbre des régions exotiques, dont les fruits, de la taille d'un citron, renferment une pulpe à la saveur agréable (extrait de Valette, Estacio-Moreno *et al.* 2006).

3. Des objectivations sémantiques

Le dictionnaire sémique a servi et sert encore à un certain nombre d'expérimentations d'annotation. On abordera dans ce paragraphe différents travaux exploratoires visant à évaluer son utilisation dans l'identification des fonds et des formes sémantiques.

3.1. Le fond sémantique

3.1.1. Les isotopies intertextuelles

Grzesitchak réalise actuellement un algorithme destiné à identifier, dans un corpus de texte homogène les isotopies domaniales les plus à même de profiler ledit corpus. L'enjeu de ce travail est avant tout de pouvoir sélectionner précisément les signifiés des unités lexicales du corpus, autrement dit, d'effectuer une désambigüisation par le domaine. C'est selon nous une façon de simuler la présomption d'isotopie (l'interprétation implique d'inhiber les sèmes non pertinents). Si le protocole est encore inachevé, il donne toutefois

Forme	**Frq.Tot.**	**Fréquence**	**Coeff.**
/arts dramatique/	78	47	27
/beaux-arts/	1687	292	19
/anatomie%humaine/	70	32	14
/littérature/	1811	277	12
/photographie/	553	111	12
/typographie/	482	97	11
/textile/	229	56	10
/architecture/	1283	196	9
/théâtre/	1092	174	9
/esthétique%histoire/	36	17	9
/esthétique/	89	29	9
/hippisme/	269	58	8
/spectacles/	408	79	8
/manège/	201	44	7
/arts/	1592	223	7

Figure 15 : Isotopies domaniales candidates identifiées sur un corpus d'articles du *Monde diplomatique* que la rédaction avait indexés comme relevant des arts (spécificités supérieures ou égales à 7) (d'après un document de travail de M. Grzesitchak).

des résultats encourageants. Ainsi, il est capable de distraire de la masse hétérogène des candidats-sèmes domaniaux un faisceau isotopique pertinent pour un corpus donné. A titre

d'exemple, sur 14 articles du *Monde diplomatique* traitant, si l'on se réfère aux mots-clés suggérés par la rédaction, des arts, on obtient après analyse contrastive sur un corpus de référence constitué de 110 articles, la liste suivante (figure 15), organisée par coefficient de spécificité[31] :

En bref, ce travail de qualification domaniale (mésogénérique, dans la terminologie rastiérienne) apparaît, tant pour des raisons TAL qu'interprétative, une étape prioritaire dans le travail d'annotation sémique de corpus.

classement freq	Candidats	freq, texte	moyenne 1	moyenne 2	écart-type 1	écart-type 2
401	/prisonnier/	22	2,75	2,75	0,96825	0,96825
495	/travailler/	18	2,25	2,25	0,96825	0,96825
366	/prétendre/	24	3	3	1	1
361	/étranger/	24	3	3	1,22474	1,22474

Figure 16 : Candidats-sèmes présent dans 100% des paragraphes dans l'article du *Monde diplomatique*« Toulon, ville amirale du front national » (août 96) (d'après Grzesitchak *et al.* 2007).

3.1.2. Les isotopies intratextuelles

Les travaux exploratoires exposés par (Grzesitchak *et al.*, 2007) ont montré en outre qu'il était possible d'observer des récurrences de sèmes non domaniaux (c'est-à-dire des isotopies potentielles de sèmes spécifiques) dans un texte isolé. Par exemple, dans un article du *Monde diplomatique* qui traite de l'administration de la ville de Toulon par le Front National (« Toulon, ville amirale du front national » (août 96), le candidat-sème /harceler/ apparaît dans 90% des paragraphes alors que ni le mot *« harceler »*, ni ses dérivés morphologiques ne sont actualisés dans le texte. De la même façon, en classant les candidats-sèmes du texte « La politique française d'immigration mise à l'épreuve » (juillet 97) par ordre croissant selon l'écart-type calculé à partir de leurs fréquences dans la totalité des paragraphes puis selon le taux de présence dans les paragraphes, on fait émerger parmi les 16 premiers candidats-sèmes, /travailler/ et /étranger/ qui semblent relever directement du sujet du texte (à savoir, le travail clandestin).

L'article « Replis communautaires à Sarcelles » (janvier 96) présente un intérêt particulier. Dans ce texte où il est essentiellement question des jeunes de la ville de

[31] La chaîne de traitement utilisé interface une plateforme d'annotation (SEMY) et le logiciel Lexico3 (Salem *et al.* 2003). Une version récente de SEMY intègre désormais le calcul de spécificités suivant le même algorithme que celui de Lexico3 (Spécifié par C. Reutenauer).

Sarcelles, on a en 9ème position, le candidat /enfant/. Or, le mot « enfant » est toujours absent du texte. Mais si, précédemment, on pouvait interpréter les candidats comme des impressions (ex. harcèlement), il semble que l'auteur ici, pratique l'euphémisme en employant des mots tels que « jeune » ou « adolescent » pour qualifier les protagonistes. Le sème isotopique isolé donne accès plus crûment à la réalité des faits. En bref, l'article parle des jeunes de Sarcelles ; le fond sémantique présume qu'il s'agit d'enfants.

classement freq.	**Candidats**	**freq. texte**	**présence para**	**moyenne 1**	**moyenne 2**	**écart-type 1**	**écart-type 2**
482	changer	24	66.7 %	6	4	1	1,82574
481	complet	24	66.7 %	6	4	1,22474	1,91485
476	pose	24	66.7 %	6	4	1,58114	2,08167
478	jouer	24	66.7 %	6	4	1,58114	2,08167
486	aboutir	24	66.7 %	6	4	1,58114	2,08167
392	titre	29	66.7 %	7,25	4,83	0,82916	2,08868
451	convenable	26	66.7 %	6,5	4,33	1,5	2,15390
466	déplacement	25	66.7 %	6,25	4,17	1,63936	2,16239
418	**enfant**	**28**	**66.7 %**	**7**	**4,67**	**1,41421**	**2,22544**
483	artiste	24	66.7 %	6	4	1,87083	2,23607
461	fondamental	25	66.7 %	6,25	4,17	1,78536	2,23814
465	déplacer	25	66.7 %	6,25	4,17	1,78536	2,23814
429	surtout	27	66.7 %	6,75	4,5	1,63936	2,27303
438	indéfini	27	66.7 %	6,75	4,5	1,63936	2,27303
421	confiance	28	66.7 %	7	4,67	1,58114	2,29912

Figure 17 : Candidats-sèmes présent dans 100% des paragraphes dans l'article du *Monde diplomatique* que la rédaction avait indexés comme relevant des arts (spécificités supérieures à 6).

Encore expérimentale, cette recherche sur les régions peu explorées de l'infralexicalité devrait bénéficier d'une typologie des sèmes et des isotopies (objet de la thèse de M. Grzesitchak).

3.2. Les formes sémantiques

(Reutenauer *et al.* 2007) cherche a évaluer les déformations subies par une forme sémantique dont l'élément stable est constitué de l'unité lexicale *« économie réelle »* dans un corpus de presse constitué de 1587 articles, tirés du le *Figaro* et de l'*Humanité* entre septembre 2008 et février 2009. Le thème du corpus est la crise économique et financière. Le corpus se présente sous forme de deux versions parallèles : la version lexicale, d'un million d'occurrences de formes, et une version « sémique » de 23 millions d'occurrences de candidats-sèmes. A partir d'un calcul de spécificités (implémentation Lexico3, Salem *et*

al. 2003), on cherche à voir l'influence de la ligne éditoriale des différents titres sur l'environnement d'*« économie réelle »*, à la fois à partir de l'usage traditionnel des formes et de l'image sémique du corpus.

Candidats	**Spécificité**	**Candidats**	**Spécificité**	**Candidats**	**Spécificité**
Budget#subst	21	appréciable#adj	10	effondrement#subst	9
particulier#subst	16	capitaliste#adj	10	enthousiasme#subst	9
ressource#subst	16	collision#subst	10	financier#adj	9
régir#v	15	contagion#subst	10	galaxie#subst	9
argent#subst	14	décisif#adj	10	intense#adj	9
particulier#adv	14	dysfonctionnement#subst	10	noeud#subst	9
répercussion#subst	14	économie#subst	10	pathologique#adj	9
théâtre#subst	13	époux#subst	10	phénomène#subst	9
bien#subst	12	profond#subst	10	progressif#adj	9
chômage#subst	12	subit#adj	10	retentissement#subst	9
déterminant#adj	11	boursier#adj	9	rupture#subst	9
diminution#subst	11	craindre#v	9	sous-production#subst	9
néfaste#adj	11	D=dramaturgie	9	surproduction#subst	9
ralentissement#subst	11	développement#subst	9		
roi#subst	11	économique#adj	9		

Figure 18 : Candidats-sèmes les plus spécifiques des paragraphes contenant « économie réelle » *(extrait de Reutenauer et al. 2009)*

La figure 19 montre les candidats-sèmes les plus spécifiques du voisinage sur le corpus total On y observe un faisceau probable d'isotopies autour des candidats /budget/, /argent/, /capitaliste/, /boursier/ et /collision/, /répercussion/, /contagion/, /pathologique/, /effondrement/ d'une part.

L'étude mesure la sensibilité des informations au contexte éditorial des deux quotidiens du corpus, par exemple, l'influence du journal sur le signifié du mot-pôle.

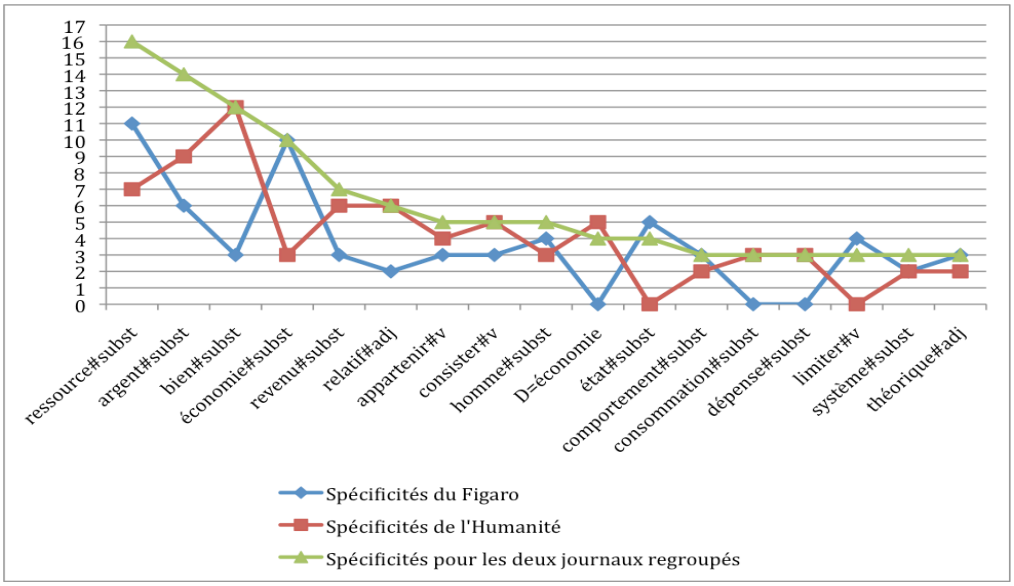


Figure 19 : Spécificités en fonction des candidats-sèmes d' « économie réelle », pour les candidats de spécificité supérieure à 3 sur l'ensemble des passages : étude de l'apport de chaque journal à la représentation globale (*(extrait de Reutenauer et al. 2009)*

Il semble ainsi que *l'Humanité* active par exemple les sèmes /bien/ (substantif), /revenu/, /consommation/ et /dépense/, tandis que /ressource/ et /économie/ sont actualisés par *le Figaro*. On en déduit une vision plus globalisante, plus macroéconomique dans *le Figaro* et une vision plus localisante, plus en lien avec les consommateurs dans *l'Humanité*. Cette analyse est corroborée par la lecture humaine. En dépit d'un bruit persistant, l'étude de Reutenauer *et al.* montre une convergence des résultats sur les plans sémique et lexical pour des axes sémantiques majeurs.

3.3. L'humilité du sème

Les sèmes ne sont pas des universaux, ils sont de simples traits sémantiques, des valeurs modestes élaborées, construites par le linguiste ou par quelque artifice mécanique pour les besoins d'un texte ou d'un corpus. Face aux universaux, aux invariants et aux primitives, qui sont des géants, puissants, translingues et peu nombreux, les sèmes opposent leur indénombrabilité (l'inventaire en est inachevé, il s'en crée chaque fois que nécessaire de nouveau) et surtout leur dépendance déterminante à la langue, la pratique et la culture. En fait, s'ils n'ont guère, ou pas de valeur intrinsèque, leurs associations – syntagmatiques en particulier – sont leur puissance.

Chapitre 4
Texte et conceptualisation

Nous proposons, pour aborder la question de la conceptualisation et de la lexicalisation des concepts, d'adopter une méthode de description fondée sur les propositions théoriques et les outils informatiques décrits précédemment. Nous présenterons ensuite un ensemble de prospectives et un programme de recherche relatif à l'approfondissement de notre approche dans la perspective des nouvelles applications de la linguistique, en particulier dans un contexte variationniste et multilingue.

1. Signifiés et concepts, mots et termes

La problématique de la veille lexicale et conceptuelle se mesure à l'aune de sa parente, la veille terminologique[32] qui vise à constituer des lexiques spécialisés (langue de métier, langue de spécialité). Anciennement fondée sur une approche onomasiologique, la terminologie textuelle (Bourigault et Slodzian 1999) repose dans ses derniers développements, sur l'extraction de syntagmes (par patrons morphosyntaxiques et par des méthodes statistiques, Daille 1994) à partir de corpus de textes. Mais la comparaison fait long feu pour des raisons méthodologiques et théoriques. Si l'on peut sans faillite épistémologique parler des langues de métier ou de spécialité, il semble improbable de distinguer, par opposition, une langue générale, comme c'est pourtant parfois le cas. Accepter l'idée qu'il y a une langue générale conduit à se poser de nombreux problèmes artefactuels, en particulier en lexicologie, tel que celui de la polysémie. Or, tous les actes énonciatifs et interprétatifs s'inscrivent dans des pratiques sociales. Cela signifie, d'une part, que les textes appartiennent à des discours et à des genres déterminés qui contraignent

[32] Nous excluons d'emblée la « néologie d'aménagement » telle qu'elle est pratiquée par exemple par la Société française de terminologie. Elle consiste à créer de façon plus ou moins pertinente des mots nouveaux, le plus souvent pour les substituer à des lexies pourtant attestées mais d'origine anglo-saxonne.

tous les paliers de complexité du texte (lexique, syntaxe, sémantique), et d'autre part, qu'ils s'insèrent dans des domaines particuliers, dans lesquels on ne rencontre pas en général de polysémie : « antenne » s'actualise dans le domaine entomologique et dans celui de l'émission d'ondes radio, mais ces deux domaines ne s'interpénètrent qu'exceptionnellement.

La raison théorique ressortit à l'opposition entre le terme et la lexie. Alors que le terme a, en règle générale, une signification précise et exprime une idée définie de façon univoque[33], il n'en est pas de même en ce qui concerne la lexie. Le *TLF* observe que « la frontière entre "lexie" et "énoncé libre" n'est pas nettement tracée ; la phraséologie occupe un domaine intermédiaire, selon un continuum allant de la suite lexicalisée au syntagme et à l'énoncé simplement fréquent en discours et prévisible en langue ». On ajoute que la lexie ne répond pas au critère d'univocité. Ses frontières, tant formelles que sémantiques, sont beaucoup plus incertaines.

2. Forme sémantique et concept

2.1. Hypothèse générale

Rappelons les principaux présupposés nécessaires à notre propos, empruntés ou inspirés de la sémantique textuelle :

(i) Le texte est la trace de pratiques sociales. Un texte est produit et interprété dans des situations liées à des pratiques sociales, lesquelles sont identifiables en termes de discours et de genre. Le discours correspond à la pratique (par exemple, le discours journalistique, le discours scientifique, le discours médical, *etc*.) et le genre à des normes de production et d'interprétation des textes relatives au discours considéré. Ainsi, dans le discours journalistique, on trouve le fait divers, l'article, le reportage, l'éditorial, etc.

Nous faisons ainsi l'hypothèse générale que le néologisme, qu'il soit formel (c'est-à-dire qu'il relève du signifiant) ou sémantique (c'est-à-dire qu'il relève du signifié), subit les contraintes discursives et génériques exercées sur les textes dans lesquels il s'actualise. Plus précisément, si tout discours est a priori créatif à proportion de la vitalité de la pratique sociale correspondante, les genres, quant à eux, présentent un potentiel néologique variable. Ainsi, parmi les genres argumentatifs du discours littéraire, le pamphlet est réputé créatif, l'essai est plus conservateur.

[33] Sur l'équivocité en terminologie, *cf.* néanmoins Cabré 1999.

(ii) Le sens se décrit, comme nous l'avons vu, en termes de sèmes. Les unités lexicales s'organisent en classes sémantiques structurées en fonction de sèmes partagés qui unifient la classe (sèmes génériques) et de traits sémantiques particuliers qui différencient les éléments de la classe (sèmes spécifiques).

(iii) la cohésion des textes est assurée par des réseaux de sèmes (cohésion intratextuelle). Ces réseaux correspondent à des fonds sémantiques (sèmes récurrents, ou isotopie, organisés en faisceaux) et à des formes sémantiques (groupements stabilisés de sèmes). Fonds et formes sémantiques assurent également l'articulation du texte avec l'intertexte.

Dans ce contexte, nous entendons donner un rôle privilégié à la notion de forme sémantique. Nous proposons de la considérer comme le signifié d'un signe sans signifiant synthétique attitré. Inversement les signes sont des formes sémantiques lexicalisées. D'un point vue interprétatif, signifiés et formes sémantiques sont des groupements sémiques compacts et associés à un signifiant stable et synthétique dans un cas, discontinu et sans lexicalisation privilégiée dans l'autre cas. Soit, en résumé, l'hypothèse suivante et son corollaire :

Hypothèse : La forme sémantique est le signifié d'un signe sans signifiant stabilisé et synthétique attitré.
Corollaire : Une lexie est composée d'une forme sémantique associée à un signifiant stabilisé et synthétique.

On représentera cette équivalence de la façon suivante :

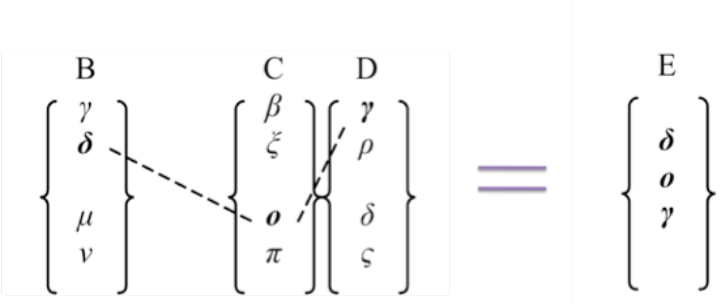


Figure 20 : équivalence entre la forme sémantique {γ, o, δ} et le signifié de E

La lexicalisation est donc un moment particulier de l'actualisation des formes sémantiques ou, si l'on préfère, la lexie est un cas particulier de forme sémantique. Le schéma de la figure 21 donne à voir l'évolution générale des formes sémantiques dans cette perspective. L'enjeu désormais est de les approfondir et de les valider. Les paragraphes ci-après présenteront différentes réalisations et outils mis en œuvre dans le but

d'éprouver notre modèle. Nous exposerons notamment quelques résultats d'une recherche menée sur la détection des néologismes de formes, ce qui nous permettra d'illustrer la question des contraintes intertextuelles exercées sur la conceptualisation. Puis, dans un autre paragraphe, nous exposerons quelques propositions pour la détection de l'innovation sémantique ; ce sera l'occasion d'illustrer la question des contraintes intratextuelles exercées sur la formation des néologismes.

2.2. Les phases de développement d'un concept

Distinguons trois phases dans le développement d'un concept (voir figure 21).

(i) Thématisation

Au cours de la première phase, appelée thématisation, la forme sémantique se stabilise. Tant du point de vue du signifié que du signifiant, ses éléments constitutifs tendent à se figer, à varier de moins en moins. Avant la lexicalisation, une représentation (un concept) peut en effet exister textuellement de façon plus ou moins ténue, à l'état de forme(s) sémantiques(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. De même, une forme sémantique peut se scinder en plusieurs sous-thèmes, lesquels peuvent coexister dans un même contexte ou se spécialiser en fonction de différentes problématiques. Plusieurs formes sémantiques génétiquement distinctes peuvent se rencontrer, s'enchevêtrer, se regrouper, pour finalement se séparer ; elles peuvent également cohabiter durablement sans se confondre, mais, par exemple, en échangeant ponctuellement quelques sèmes. Par exemple, certains mouvements écologistes partisans de la désindustrialisation ont développé dans les années 70 une critique de la société de consommation accompagnée d'une phraséologie variée (autour des motsde *« post-productivisme »*, *« convivialité »*, etc.).

(ii) Lexicalisation

La deuxième phase, la lexicalisation, correspond au figement lexical de la forme sémantique stabilisée. Le groupement de sèmes devient un signifié et la lexie se voit pourvue d'un signifiant fixe. C'est à partir de la lexicalisation que l'on peut d'ailleurs parler de néologisme ou de nouveau concept. On considère que cette lexicalisation concerne d'abord un domaine restreint et demeure relativement circonscrite à des discours donnés. Ainsi, *« décroissance »* en France, a concentré ces dernières années bon nombre des thèmes des années 70 évoqués à l'instant dans le discours politique.

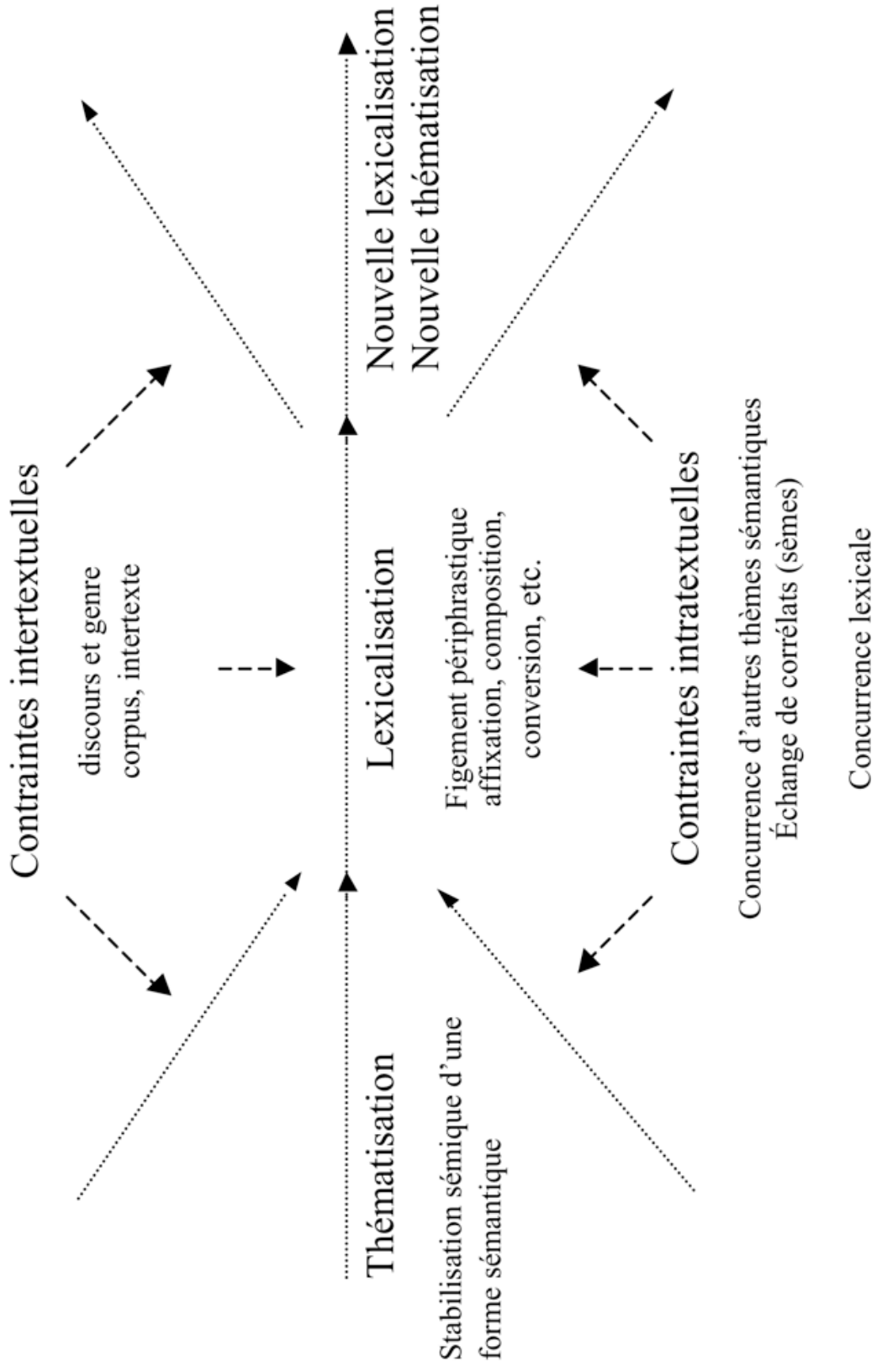


Figure 21 : Forme sémantique et lexicalisation du concept

(iii) Altération

La troisième phase consiste en l'altération du néologisme. C'est typiquement à ce moment-là que l'on peut parler de néosémie (*cf. supra*). Le néologisme peut par exemple participer à l'émergence de nouveaux domaines ou sous-domaines, faire l'objet de changement de domaine, d'inflexions thématiques, de spécialisation, etc. *« Décroissance »* connaît actuellement un usage plus varié, politique, économique, social, et se positionne notamment par rapport à des lexies à la fois proches et concurrentes (l'expression *« simplicité volontaire »*, venue du Québec, par exemple).

2.3. Contraintes textuelles et intertextuelles

Durant ces trois phases, le processus d'innovation lexicale subit plusieurs contraintes. En premier lieu, des contraintes exercées par l'intertexte ; ces contraintes correspondent au minimum à celles évoquées précédemment : il s'agit de contraintes discursives et génériques, certains genres sont par exemple plus créatifs d'un point de vue lexical que d'autres ; une forme sémantique peut connaître une évolution différente suivant que son genre d'actualisation est plus ou moins productif. Par ailleurs, les domaines sont également d'une créativité variable. Les contraintes intertextuelles s'exercent en amont de la lexicalisation (variabilité, types de sèmes et complexité des formes sémantiques), sur la lexicalisation (constructions néologiques liées aux genres) – *« décroissance »* est plus polémique que *« simplicité volontaire »* et s'actualise volontiers dans la presse satirique ; et en aval (vitalité du domaine, spécialisation et déspécialisation du néologisme, etc.).

On discerne également des contraintes (intra)textuelles ; il s'agit de l'influence exercée par les textes dans lesquels la forme sémantique s'actualise. Ces contraintes sont liées aux réseaux sémiques desdits textes. En amont de la lexicalisation, différentes formes sémantiques sont en concurrence, elles peuvent partager des sèmes ou êtres portées par différentes isotopies du fond sémantique. Les formes sémantiques échangent des sèmes en fonction de leur proximité. Ainsi, les thèmes sémantiques de la décroissance sont en relation de cooccurrence et d'opposition avec ceux du développement durable. Une sélection des sèmes s'opère lorsqu'un thème sémantique se stabilise en signifié, par exclusion ou adoption des sèmes partagés par des unités lexicales ou thèmes sémantiques voisins.

Dans le paragraphe suivant, nous tenterons d'illustrer ces différentes propositions à partir d'analyse en corpus.

3. Contraintes intertextuelles : genres et créativité lexicale

Le contexte de la recherche que nous évoquerons ici est la réalisation d'une plateforme de veille lexicale semi-automatisée pour la production de ressources lexicographiques (attestations, mesures, contextes et sources). Il s'agit de développer des outils pour collecter des textes à partir de sources différentes (fichiers, bases de données textuelles, Internet) et d'en extraire les unités lexicales absentes de lexiques de référence ; ce, dans une perspective diachronique. En bref, on souhaite produire du matériau pour les lexicographes, par exemple pour enrichir des lexiques existants, créer les métadonnées ou encore sélectionner des contextes caractéristiques et au-delà, pour participer à la création de nouvelles pratiques lexicographiques. Mais cette plateforme constitue également un outil pour l'étude théorique de la néologie. On rapportera ici quelques propositions conceptuelles pour l'évaluation de la créativité néologique corrélée aux genres et aux discours[34].

La méthode générale s'insère dans une perspective contrastive. Elle consiste à comparer les traces de pratiques sociales (*i.e.* des collections de textes homogènes) à des usages lexicaux simulés (*i.e.* des lexiques), de manière à évaluer la richesse néologique et la créativité lexicale de différents genres textuels. Nous nous appuyons ici sur un corpus discursivement et thématiquement homogène (le pouvoir d'achat traité dans la presse magazine). Le corpus se scinde en trois sous-corpus issus de la presse hebdomadaire française grand public (*Marianne*, *Le Nouvel Observateur*, et *Le Point*) entre juillet 2004 et avril 2008. On dénombre 406 textes pour 304 727 occurrences de formes (*cf.* tableau et graphique de la figure 22).

3.1. Richesse lexicale et richesse néologique théorique

La figure 21 présente deux valeurs mesurées sur nos différents sous-corpus. La première, la richesse lexicale, correspond au rapport entre le nombre de formes et le nombre d'occurrences de formes. Cette mesure est parfois critiquée parce qu'elle dépend de la taille des textes comparés (la richesse lexicale décroît avec la taille du texte). Inquiété par la petitesse du sous-corpus du *Point* relativement aux autres, nous avons expérimenté un certain nombre d'indices pour constater une relative homogénéité quant aux résultats. Les données finalement exposées dans la figure 21 ont été calculées à partir de l'indice W proposé par E. Brunet et rapporté par Ch. Muller (1977 [1992], 196) :

[34] L'outil d'extraction de néologisme utilisé ici a été réalisé par Sandrine Ollinger (ATILF, Nancy). Pour plus d'information, on lira (Ollinger & Valette à paraître).

$$W = N^{V_{-\alpha}}$$

où N est, par convention, le nombre d'occurrences de formes, V le nombre de formes et a une constante égale à 0,172 (choix par défaut que nous avons conservé). Pour des raisons de lisibilité, le résultat présenté dans la figure 3 est $(\frac{1}{W})\times 100$.

Nous proposons ensuite de calculer l'indice de richesse néologique U suivant une équation similaire :

$$U = V^{C_{-\alpha}}$$

où V est le nombre de formes, C le nombre de candidats et a la même constante que précédemment. Le résultat présenté est $(\frac{1}{U})\times 100$.

	formes	occurrences	candidats	richesse lexicale	richesse néologique théorique
Corpus total	21 825	304 727	764	7,16	3,50
LePoint	9 030	86 772	175	9,31	2,36
Marianne	14 593	108 753	461	10,76	3,55
NouvelObs	9 860	109 202	248	9,20	2,83

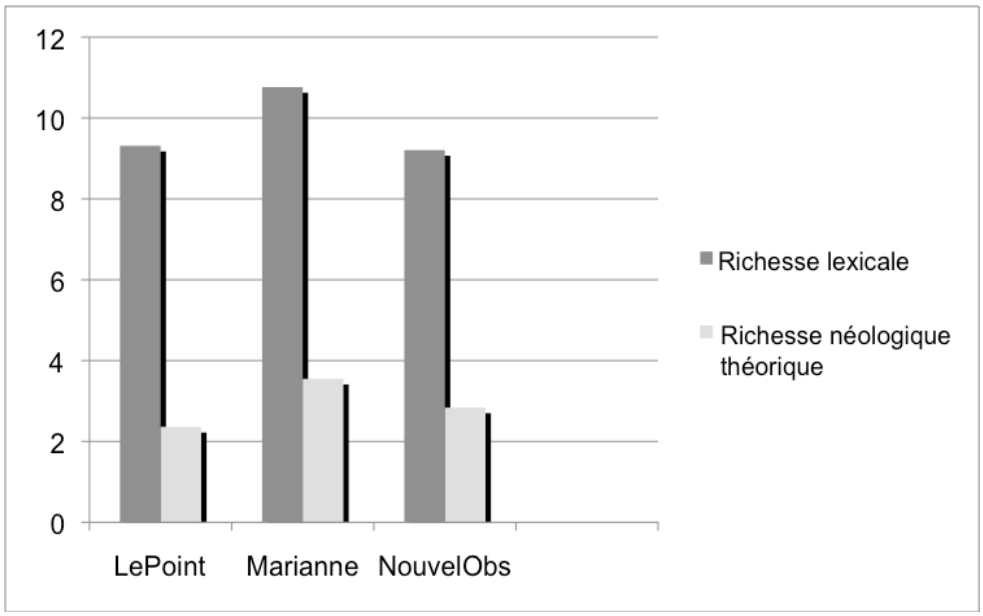


Figure 22 : Richesse lexicale et richesse néologique théorique du corpus

On parle ici de *richesse néologique théorique** dans la mesure où nous traitons des données brutes non triées. Autrement dit, certains candidats ne sont pas des néologismes –

il peut s'agir de variations idiosyncrasiques, orthographiques ou encore d'entités nommées absentes de nos lexiques ou non signalées comme noms propres par l'étiqueteur que nous avons ici utilisé (Treetagger, Schmid 1994). Le tableau et la figure 22 donnent à voir les richesses lexicales et néologiques théoriques du corpus.

3.2. Créativité et conservatisme lexicaux

On observe que la richesse lexicale et néologique du sous-corpus *Marianne* est relativement élevée comparée aux autres sous-corpus. Cela nous amène à proposer les notions de *conservatisme lexical* et de *créativité lexicale**. Le conservatisme lexical est la tendance à employer peu de néologismes proportionnellement à la variété des formes actualisées. A l'inverse, la créativité lexicale est la tendance à employer une grande variété de néologismes proportionnellement à la variété des formes actualisées. Ces concepts sont fondés mathématiquement sur le rapport entre la richesse lexicale et la richesse néologique théorique[35]. Le schéma de la figure 23 présente les taux de conservatisme lexical théorique et de créativité lexicale théorique des différents corpus.

	Conservatisme lexical
Marianne	3,03
Nouvel Obs	3,24
Le Point	3,95

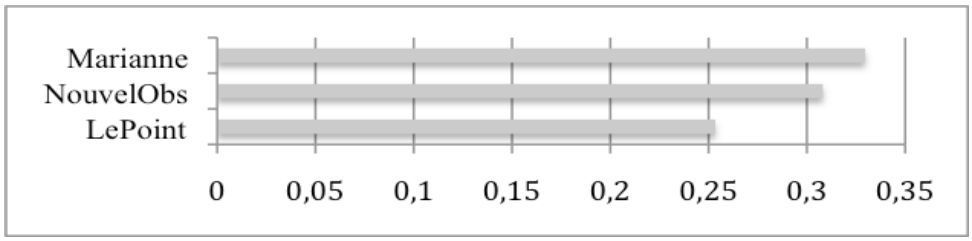


Figure 23 : Conservatisme lexical (en haut) et créativité lexicale (en bas) du corpus

Un indice de créativité lexicale élevé correspond à un recours plus important à des candidats à la néologie variés proportionnellement à la richesse lexicale mesurée. Selon cette mesure, l'hebdomadaire *Le Point* est sensiblement plus conservateur que les deux autres. *Marianne* présente le taux de créativité lexicale le plus élevé pour une richesse lexicale et une richesse néologique théorique supérieures aux autres hebdomadaires retenus.

[35] Soit, le taux de créativité lexicale :

$$\frac{1}{rich\_lex / rich\_néo}$$

Exemple : la créativité lexicale dans Marianne

Nous avons étudié les modalités de construction des candidats à la néologie dans les textes du sous-corpus *Marianne*. Une recherche par sous-chaîne de caractères nous a permis d'identifier un des modes de créativité privilégiés par cet hebdomadaire : il s'agit de la néologies dérivationnelle et suffixale. Si celle-ci n'est pas absente des autres hebdomadaires consultés, il apparaît que *Marianne* a quatre à cinq fois plus recours à ce mode de production. Le tableau de la figure 24 donne à voir sommairement les principaux types de constructions remarquables identifiées.

Nous suggérons que ces choix stylistiques d'importance sont liés à des contraintes intertextuelles ; Marianne est connu pour son ton polémique et volontiers pamphlétaire – or, des travaux récents ont montré que la néologie était un mode de stylisation courant dans le pamphlet (Jousse 2007). C'est pour nous l'indice d'une double contrainte intertextuelle conjointe, éditoriale et donc sociale d'une part (les journalistes fondent leur style dans le style du journal), générique d'autre part (on adopte les normes des genres polémiques, le pamphlet en premier lieu).

	Marianne		*Le Point*		*Le Nouvel Observateur*	
	For m.	Exemples	Form .	Exemples	Form .	Exemples
Opposition, Négation	18	anticorporatiste	6	non-annonces	4	anticoncurrentiel
Péremption	18	ex-trublion	4	ex-candidate	4	ex-travailleuse
Approximation	4	quasi-maniaque	0		1	quasi-impasse
Hyperbole	9	hypercapitalisme	2	surprofit	6	superprofits
Itération	10	refondation	3	remobiliser	0	
Agglutination	7	tactico-politiciens	2		0	
Procès (sation)		starisation	3	annualisation	1	
Dérivation d'entités nommées	26	gaudinerie, Sarkozie	1	villepeniste	9	berlusconien
Total	99		21		25	

Figure 24 : Constructions néologiques remarquables
(*Marianne* comparée au *Point* et au *Nouvel Observateur*)

Les concepts de richesse néologique (pour l'heure, « théorique »), de conservatisme lexical et de créativité lexicale que nous avons esquissés ici nécessitent bien évidemment d'être évalués et raffinés ; ils constituent toutefois des outils fonctionnels pour le développement d'une problématique générale de veille lexicale.

4. Contraintes intratextuelles : l'économie sémique

Pour aborder les contraintes intratextuelles, nous nous intéresserons à l'évolution sémantique d'un mot, ou *néosémie*$. L'étude de ce phénomène présente un important enjeu en matière de veille lexicale et donc en matière de constitution de ressources dictionnairiques, dans la mesure où seul le signifié du mot change sans que le signifiant n'en soit affecté, ce qui rend les procédures de détection délicates. Par exemple, « percuter » signifiait initialement « frapper, heurter » mais il peut aujourd'hui, dans certains discours et registres, être compris comme « comprendre immédiatement ».

La néosémie est une façon purement textuelle d'envisager le problème lexical de la polysémie. La polysémie est en effet un artefact résultant de l'isolement du mot, de sa décontextualisation. Restituer son contexte, a fortiori son contexte sémique, c'est restituer les conditions de sa sémantisation, c'est-à-dire de son interprétation comme signe. La notion de néosémie invite à considérer l'émergence d'un nouveau signifié en termes d'économie ou d'organisation sémique : la variabilité des actualisations possibles d'une lexie induit un réaménagement des sèmes composant son signifié.

On étudiera deux tendances conjointes : d'une part, certaines néosémies résultent d'une modification de l'appartenance domaniale (changement de domaine, nouvelle domanialisation, etc.), laquelle s'accompagne de variations des contraintes génériques et discursives, d'autre part, la néosémie est une reconfiguration du ou des signifiés constituant la lexie d'origine, notamment par diffusion sémique des contextes. Cette étude est développée et argumentée plus en détail dans (Rastier & Valette 2009).

4.1 La néosémie est une modification de l'appartenance domaniale

Changement de domaine

Prenons ici un exemple de lexicalisation homonymique donnant lieu à un changement de domaine. Le substantif masculin « filaire » est employé dans le domaine des télécommunications au sein de la classe sémantique des //appareils de transmission// (téléphone, modem) parce que son sème spécifique /fil/ l'oppose au sème /radio/. Or, dans le *TLF*[36], seule l'acception zoologique (substantif féminin) est retenue. La filaire est un ver au long corps rond et filiforme. On peut douter que l'usage dans la classe des //appareils de

[36] Nous prenons systématiquement le *TLF* comme dictionnaire de référence par fidélité à l'objectif lexicographique initial de la veille lexicale. Il va de soi que ce qui est néosémique par rapport au *TLF*, qui n'est plus actualisé depuis plusieurs années, ne l'est pas forcément pour les dictionnaires plus récents.

transmission// soit métaphorique *stricto sensu*. Il s'agit plutôt d'une nouvelle suffixation de 'fil' sans doute réalisée indépendamment de la forme déjà existante, vraisemblablement par métonymie (un « téléphone filaire »), ce qui explique notamment que le genre soit différent.

Pour l'heure tout du moins, l'usage néosémique du substantif « filaire » apparaît relativement aisé à repérer automatiquement. Les domaines d'actualisation sont peu nombreux et très typiques (vocabulaire technique). Comme la technologie filaire est actuellement marginalisée, on peut estimer que la lexie relève de la langue de spécialité.

Emprunt par inhibition locale

« Logiciel » est une néologie d'aménagement du domaine de l'informatique. Il date, d'après le *TLF*, de 1978. C'est un dérivé de *logique* suffixé en *-iel* (par analogie à « *matériel* »). La définition qu'en donne le *TLF* est la suivante :

> LOGICIEL, subst. fém. INFORMAT. [P. oppos. à matériel] « Ensemble des programmes, procédés et règles, et éventuellement de la documentation, relatifs au fonctionnement d'un ensemble de traitement de données » (J. O., Vocab. de l'informat., 17 janv. 1982).

Soit ; un signifié qu'on pourrait simplifier de la façon suivante

{ /INFORMATIQUE/ /programmes/ /procédés/ /règles/ /documentation// /traitement de données/ }

Figure 25 : Signifié de logiciel

Or, depuis quelques années, « logiciel » a un emploi très circonscrit dans un domaine tout autre, la politique. Il semble être une création du Parti Socialiste et est resté cantonné à ce seul parti pour finalement essaimer vers d'autres sphères technolectales. A priori, l'emploi date de la fin des années 90, avec une expansion au début des années 2000, sans doute après l'échec aux élections présidentielles du Parti Socialiste. Par exemple :

> Il faut promouvoir ces avancées par une mise en œuvre souple, décentralisée, souvent contractuelle, réactive. C'est en cela aussi que nous devons rénover le logiciel social-démocrate[37].

> La gauche radicale doit faire quelque chose qu'elle ne sait pas faire : politiser les intimités blessés par le capitalisme. Le logiciel principal reste encore la lutte contre les inégalités sociales.[38]

> Le passage de la LCR au NPA s'explique par une volonté de changer de dimension, grâce à un leader charismatique et médiatique, Olivier Besancenot. Mais le logiciel reste le même : celui du marxisme révolutionnaire, basé sur la planification de l'économie, la collectivisation des entreprises, le remplacement des patrons par des dirigeants élus par les salariés, celui de la démocratie représentative parlementaire par la démocratie directe... [39]

> Le PS révèle, sous l'icône de sa candidate, sa vraie nature : celle d'un parti au logiciel anachronique, coupé des réalités et des attentes de nos compatriotes, dépourvu d'un véritable projet d'avenir[40].

« Logiciel social-démocrate », « logiciel du marxisme », « logiciel anachronique » ou ailleurs, *« logiciel pensé pour l'immigration »*, ce logiciel-là semble parfaitement adapté pour désigné un concept pour lequel, finalement, la définition du logiciel informatique n'est pas inadaptée. Ainsi, on pourrait imaginer que le logiciel, dans l'acception politique corresponde à une définition telle que :

> *LOGICIEL$_2$, subst. masc. POLITIQUE. « Ensemble des programmes, procédés et règles [mis en œuvre dans le cadre d'une politique] »

Il est intéressant de constater que l'informatique, qui longtemps a construit sa terminologie sur des emprunts (*« souris »*, *« fenêtre »*, *« bureau »*) soit aujourd'hui prescriptrice de terminologies et vraisemblablement, au delà des mots, des concepts correspondants. Il faut sans doute y voir une prise de conscience du rôle de l'informatique aujourd'hui, comme moyen de production au sens marxiste (*cf.* aussi l'emploi fréquent en informatique du mot *« outil »* pour désigner le programme ou le logiciel, ou encore le mot « forge » pour désigner les systèmes de gestion de développement collaboratif de logiciels)[41].

---

[37] http://www.temps-reels.net/article1111.html, [28 08 2002] 22 09 2009.

[38] http://www.lyoncapitale.fr/index.php?menu=02&article=7402, [10 03 2009] 22 09 2009)

[39] http://tempsreel.nouvelobs.com/actualites/opinions/interviews/ […], [5 02 2009] 22 09 2009)

[40] http://www.ump-clamart.info/, 22 09 2009)

[41] Maurice Ronai, ancien délégué national du Parti Socialiste pour les technologies de l'information, a envisagé de proposer le mot « forge » pour désigner les travaux de « refonte du logiciel socialiste », par

Perte de l'appartenance domaniale

Le déplacement de sens d'une lexie vers une nouvelle acception peut également la dédomanialiser (plutôt que relevant d'un ou plusieurs domaines particuliers). Ainsi, le verbe transitif « percuter », qui signifie « heurter, donner un choc » est aujourd'hui employé avec le sens de « réagir, comprendre tout de suite », ou parfois, avec la négation, « ne pas faire le rapprochement, manquer d'à-propos ».

Alors que les conditions d'énonciation du verbe original subissent de fortes contraintes domaniales et taxémiques (massivement : Transport, Mécanique et Balistique), et très peu de contraintes au niveau des genres textuels (sinon aucune), la néosémie « percuter » est non contrainte d'un point de vue domanial et taxémique, mais les genres sont – peut-être provisoirement – plutôt restreints. Quant à l'étymologie de cet usage de « percuter », il faut vraisemblablement la chercher dans l'usage déjà bien attesté de l'adjectif « percutant » : « Qui frappe par sa netteté, par son caractère imprévu, qui produit un choc immédiat » (*TLF*).

Domanialisation

Certaines lexies tendent, au contraire de l'exemple précédent, à s'enraciner dans un domaine d'usage particulier. C'est le cas par exemple du substantif féminin *« grogne »* présenté comme le dérivé du verbe *« grogner »* dans le *TLF*, et d'un usage familier et vieilli. Dans l'acception première, la grogne signifie « Mécontentement, mauvaise humeur exprimée généralement en grognant » et est synonyme de bougonnement, grognerie, pleurnicherie ; elle est aussi bien relative à un individu qu'à un groupe. Aujourd'hui, l'usage néologique de la grogne est presque systématiquement lié au mécontentement collectif, en général exprimé dans les mouvements sociaux (grèves, manifestations) d'une catégorie socioprofessionnelle particulière (infirmières, médecins, chercheurs, voire consommateurs, usagers, etc.). Si le sens premier du mot n'a guère évolué, ce sont ses conditions d'usage qui se trouvent passablement modifiées. Le substantif se trouve fortement domanialisé en //politique// (et dans le sous-domaine des //mouvements sociaux//) et actualisé dans le discours journalistique. Là, il s'intègre dans le taxème des //soulèvements populaires//, qui comprend notamment la mauvaise humeur, la révolte, la révolution.

---

analogie aux deux définitions de la forge, informatique et artisane
(*Sources* : http://www.ronai.org/spip.php?article65, [3 10 2007] 22 09 2009)

4.2. La néosémie est une reconfiguration du signifié

La reconfiguration du signifié présente un enjeu particulier parce qu'elle ne fait pas seulement appel à la notion de domaine, souvent documentée (avec plus ou moins de bonheur malgré tout) dans les dictionnaires, elle implique d'envisager une approche « sémique ». Le signifié est composé de sèmes qualifiés (sème générique, spécifique, inhérent, afférent – dans la terminologie de Rastier 1987). La reconfiguration du signifié implique une modification de la structure des sèmes d'une unité lexicale. Par exemple, lorsque les sèmes afférents prennent le pas sur les sèmes inhérents. C'est le cas de *« caviar »* dans l'expression appréciative « c'est du caviar », où le sème afférent /luxe/, caractéristique de la classe sémantique des //mets festifs//, est suractivé, au détriment des sèmes inhérents /œuf de poisson/, /hors d'œuvre/, etc. qui eux, sont complètements inhibés. Ainsi, 'caviar' quitte son domaine d'actualisation Gastronomie pour des usages strictement appréciatifs, et complètement dédomanialisés. Par exemple : « Aujourd'hui la StarAc c'est du caviar comparé à ce qu'on a déjà vécu dans la musique industrielle » (site etnoka.fr)

« Trottinette » subit actuellement une transformation par reconfiguration de son signifié intéressante également. De jouet d'enfant, elle est en phase de devenir un petit véhicule par inhibition du sème dimensionnel /enfant/ (*vs* /adulte/) :

> TROTTINETTE, subst. fém. A. 1. Jouet composé d'une plate-forme allongée montée sur deux petites roues et d'un guidon à direction articulée, que l'enfant fait avancer en s'aidant d'un pied qu'il pose régulièrement par terre pour donner l'impulsion ou en actionnant une pédale en un mouvement de va-et-vient. […]

> *TROTTINETTE$_2$, subst. fém. A. 1. *Véhicule* [~~Jouet~~] composé d'une plate-forme allongée montée sur deux petites roues et d'un guidon à direction articulée, que l'*on* [~~enfant~~] fait avancer en s'aidant d'un pied qu*e l'on* [~~il~~] pose régulièrement par terre pour donner l'impulsion ou en actionnant une pédale en un mouvement de va-et-vient. […]

4.3. L'évolution des signifiés

Certains sèmes participent de façon privilégiée aux réseaux sémiques qui parcourent un texte. Un apprentissage sur corpus peut donc faire apparaître la régularité de ces réseaux de sèmes et donner les moyens de caractériser les signifiés (en pondérant les sèmes : activation des traits spécifiques, inhibition des traits peu spécifiques ou du bruit), voire de les réorganiser en tenant compte des usages des unités lexicales dans les corpus de textes préalablement constitués. Les dictionnaires, en décontextualisant les mots, en donnent une définition typique et consensuelle qui ne correspond pas nécessairement aux instanciations.

À titre d'exemple, nous avons repris le texte de *Bouvard et Pécuchet* vu dans le chapitre 2 et cumulé les étiquetages //botanique// et /ornemental/, toujours d'après le *TLF*.

> Alors Pécuchet se tourna vers les fleurs{/bot//orn/}. Il écrivit à Dumouchel pour avoir des arbustes/bot/ avec des graines/bot/, acheta une provision de terre de bruyère et se mit à l´oeuvre résolument. Mais il planta des passiflores/bot/ à l´ombre, des pensées/bot/ au soleil, couvrit de fumier les jacinthes/bot/, arrosa les lys{/bot//orn/} après leur floraison, détruisit les rhododendrons{/bot//orn/} par des excès d´abattage, stimula les fuchsias{/bot//orn/} avec de la colle forte, et rôtit un grenadier, en l´exposant au feu dans la cuisine. Aux approches du froid, il abrita les églantiers/bot/ sous des dômes de papier fort enduits de chandelle ; cela faisait comme des pains de sucre, tenus en l´air par des bâtons. Les tuteurs des dahlias{/bot//orn/} étaient gigantesques ; – et on apercevait, entre ces lignes droites les rameaux tortueux d´un sophora{/bot//orn/}-japonica qui demeurait immuable, sans dépérir, ni sans pousser.

Cette séquence est d'une homogénéité isotopique exemplaire. Douze fois le sème //botanique// est instanciée, et six fois le sème /ornemental/, en particulier lors des séquences *« lys »* – *« rhododendron »* – *« fuchsias »* et *« dahlias »* – *« sophora »*. On peut, de ce fait, penser que le signifié d'*« églantier »*, pris entre les deux séquences, qui partage avec celles-ci le domaine //botanique//, est susceptible d'hériter du sème /ornemental/ dont il est dépourvu. De fait, l'églantier peut être utilisé dans une perspective ornementale, bien que le sème /sauvage/ présent dans son signifié puisse fonctionner comme un inhibiteur dans la classe dimensionnelle /cultivé/ *vs* /sauvage/. Nous faisons toutefois l'hypothèse que si la cooccurrence du sème /ornemental/ et de la lexie *« églantier »* est statistiquement significative sur un corpus homogène, le signifié d'*« églantier »* est susceptible d'accueillir ce nouveau sème.

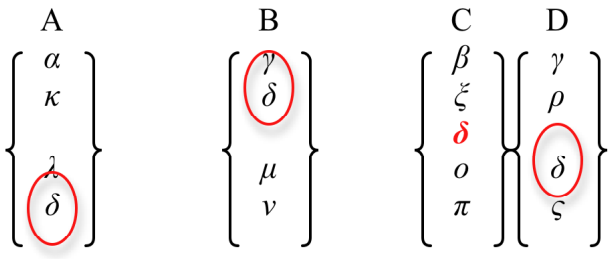


Figure 26 : C héritant du sème δ (/ornemental/) par propagation des signifiés voisins

Cette proposition ouvre également des possibilités pour l'analyse assistée des signifiés, voire pour la constitution automatique de signifiés dans le cas des néologismes nouvellement identifiés et auxquels aucun contenu sémantique n'a encore été alloué. Ainsi, un candidat à la néologie pourrait se voir attribuer automatiquement les sèmes présents avec une certaine régularité statistique dans son cotexte, – minimalement, les sèmes

permettant d'identifier le domaine (isotopie domaniale), mais aussi les sèmes de ses cooccurrents privilégiés.

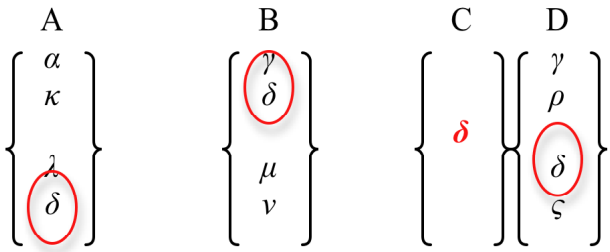


Figure 27 : C sans signifié héritant du sème δ par propagation des signifiés voisins

L'affaire /Outreau/

C'est précisément ce genre de recherche que rapportent (Reutenauer *et al.*, en préparation). Les auteurs évaluent l'hypothèse que l'émergence d'un concept s'accompagne de déformations de son environnement sémique, avec apparition et disparition de sèmes jusqu'à la stabilisation du signifié. L'étude, qui exploite le dictionnaire de sèmes décrit dans le chapitre précédent, se focalise sur un mot pour dont le signifié a évolué rapidement dans le temps, *« Outreau »*. Une étude diachronique a été auparavant réalisée par (Lecolle 2007). L'objectif est de mettre en évidence cette évolution au moyen de corpus annotés en sèmes. Le corpus, réalisé par (Lecolle, 2007), porte donc sur l'affaire judiciaire dite d'Outreau. Il est constitué d'articles de presse de novembre 2001 à avril 2006 comportant au moins une occurrence du mot *« Outreau »*. Il est divisé en cinq sous-corpus correspondant à cinq périodes :

- 2001-2002 : découverte d'un réseau pédophile, arrestation de notables ;
- mai-juin 2004 : procès de Saint-Omer
- 1er et 2 juillet 2004 : attente du verdict de Saint-Omer
- 3 au 8 juillet 2004 : verdict du procès
- décembre 2005 à avril 2006 : procès en appel à Paris ; suite et conséquences (commission d'enquête parlementaire)

(Lecolle, 2007) observe que le sens d'Outreau évolue de toponyme à « l'erreur judiciaire par excellence ». Dans l'étude de (Reutenauer *et al.* en prép.), le corpus se présente sous deux versions parallèles : une version lexicale de 400 000 occurrences de formes (issue de Lecolle, 2007) ; une image sémique de 10 millions de candidats obtenue de la même façon que lors des études rapportées dans le chapitre 3).

Comme il s'agit d'un nom propre, *« Outreau »* est absent du *TLF*i en tant qu'entrée. Il n'a donc pas plus de signifié dans le dictionnaire de sèmes qui en émane. En plus du sème réflexif /Outreau/, une définition lexicographique transformée en signifié a été produite :

> OUTREAU
> 1. Ville française du Pas-de-Calais
> 2. Erreur judiciaire liée à la découverte et croyance en l'existence d'un réseau pédophile puis à la réfutation publique de cette croyance.

Soit, le signifié :

Figure 28 : signifié d'*« Outreau »*

Afin de mesurer l'évolution diachronique de l'image sémique d'Outreau, (Reutenauer *et al.*, en prép.) cherchent à quantifier le degré de surreprésentation ou de sous-représentation de chaque candidat-sème à une période donnée. Ainsi, pour chaque période, le calcul des spécificités est appliqué aux candidats-sèmes sur le sous-corpus sémique des paragraphes de la période concernée contenant *« Outreau »*. Chaque candidat-sème se voit ainsi affecter un coefficient par période.

Les résultats mettent en évidence l'évolution des candidats-sèmes d'une période à l'autre ou encore leur poids respectif au sein d'une même période. Ainsi, des ensembles de candidats, qui semblent correspondre à des réseaux sémiques émergent de façon statistiquement significative sur certaines périodes. Par exemple, le candidat-sème /ville/ voit sa spécificité décroître au fil du temps, tandis que /judiciaire/ ou /procès/, absents en période 1, s'imposent aux périodes suivantes. Autrement dit, Outreau est de moins en moins une ville, et de plus en plus une affaire judiciaire. Pour analyser les résultats statistiques, (Reutenauer *et al.*, en prép.) ont choisi d'effectuer une évaluation manuelle indépendante du processus de traitement automatisé.

Image sémique issue de la définition théorique

(Reutenauer *et al.* en prép.) ont élaboré un protocole d'observation en trois parties :

(A) l'étude de la pertinence des candidats-sèmes dans le signifié où le caractère prédicatif de certains candidats (/découverte/, /existence/, /croyance/ et /réfutation/) pose des problèmes d'ambigüité.

(B) l'estimation manuelle de l'activation par période de chaque candidat *sans connaissance des résultats statistiques* puis la confrontation des listes établies manuellement et statistiquement. Les candidats sont classés manuellement suivant qu'ils sont « activé », « non activé » ou « indécidable » pour chaque période. Puis le résultat est comparé aux spécificités statistiques. Les valeurs de spécificités négatives ou faibles (inférieures à 2) correspondent à une non-activation du candidat-sème, et les spécificités supérieures à 2, à son activation. A l'exception des cas d'ambigüité mentionnés, (Reutenauer *et al.* en prép.) constate une convergence parfaite en période 1 (figure 29) et sur l'essentiel de la période 2. En revanche, la convergence est médiocre aux périodes 3 à 5, mais, les listes manuelle et automatique s'accordent au niveau des spécificités les plus élevées.

(C) l'évolution observée sur les cinq périodes à candidat fixé a été jugée cohérente 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.

Evaluation humaine

	période				
candidat-sème	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 candidats-sème ;
"non" : non-activation du candidat ;
"-" : candidat inclassable

Résultats numériques (spécificités)

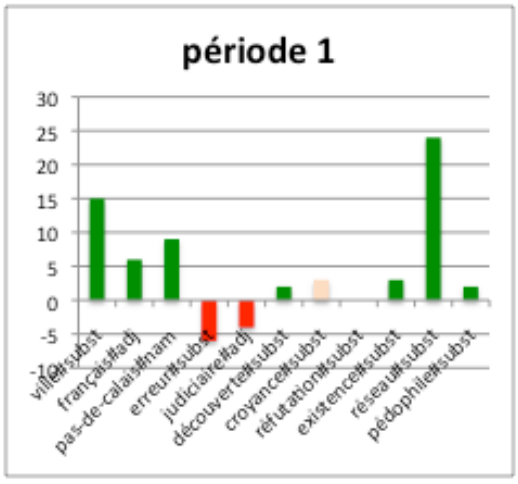


Figure 29 : Confrontation des résultats manuels et statistiques en période 1
(extrait de Reutenauer *et al.*, en préparation)

Image sémique de résonance

(Reutenauer *et al.* en prép.) appellent *l'image sémique de résonance* une image sémique où ne sont retenus que les candidats-sèmes qui correspondent aux spécificités lexicales, c'est-à-dire aux spécificités dans la version lexicale du corpus. En d'autres termes, seuls les sèmes réflexifs constituent cette image. Si on écarte les candidats de faible spécificité (entre -2 et 2), le taux de convergence entre image sémique et image lexicale atteint 89% au total, et est supérieur à 80% pour chaque période (figure 30). Mieux encore, il semble que l'image sémique de résonance réorganise pertinemment les spécificités lexicales.

	période					
	1	2	3	4	5	total
taux de convergence	87%	93%	89%	94%	83%	89%

Figure 30 : Proportion de candidats-sèmes pour lesquelles données numériques et évaluations humaines s'accordent (extrait de Reutenauer *et al.*, en préparation)

Une autre approche implique la production manuelle et indépendante de classes sémantiques. Elle consiste à confronter l'émergence par période des classes d'après les données statistiques et une analyse manuelle indépendante. A titre d'exemple, la classe //ville_et_habitants//, considérée comme très saillante en période 1 et non saillante aux autres périodes, présente un profil de spécificités conforme aux attentes (figure 31).

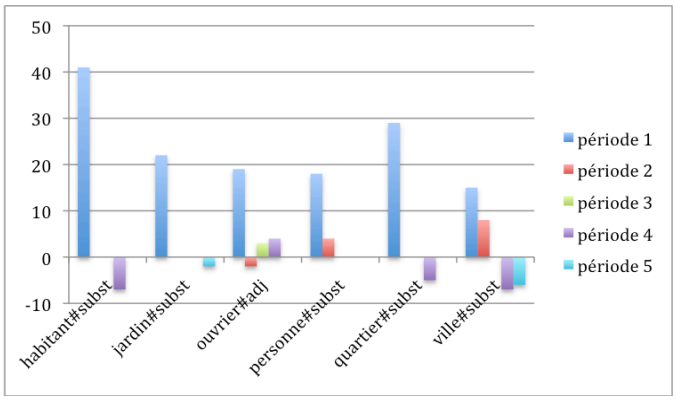


Figure 31 : Profil de spécificités des candidats de la classe //ville_et_habitants// (extrait de Reutenauer *et al.*, en préparation)

En bref, la méthode présentée par (Reutenauer *et al.* en prép.) met en place une représentation sémique quantifiée d'*« Outreau »*. Celle-ci permet d'observer une évolution diachronique de candidats-sèmes, de façon isolée ou en groupements sémiques.

Finalement, le phénomène de disparition ou apparition de classes sémantiques, assimilables aux taxèmes de la sémantique textuelle et sensible statistiquement, ouvre des perspectives en termes d'automatisation, et donc de veille conceptuelle.

5. Directions de recherches

5.1. Veille lexicale et veille conceptuelle

La *Veille lexicale* correspond initialement à un objectif lexicographique : la mise à jour de dictionnaire. L'expression est née, à notre connaissance, à l'ATILF (*Analyse et Traitement Informatique de la Langue Française*) au début des années 2000 et visait explicitement la mise à jour du *Trésor de la Langue Française* notamment dans sa version informatisée (Dendien & Pierrel 2003). Achevé dans les années 90, le *TLF* repose en partie sur le dépouillement et l'analyse d'une vaste base de données textuelles, valorisée par la mise en place de FRANTEXT, dont l'empan couvre plusieurs siècles mais s'achève aux environs des années 70. Le *TLF* rend principalement compte de la langue des XIXe et XXe siècles. En annonçant un programme de veille lexicale, il s'agissait donc d'enrichir le dictionnaire à la fois en mots nouveaux (néologie formelle : nouvelles constructions, nouveaux figements), de rendre compte des nouveaux usages (ou innovations sémantiques, ce que nous appelons des *néosémies* : nouvelles acceptions, nouvelles phraséologies), et enfin, des nouveaux comportements morphosyntaxiques (changement de genre, pronominalisation, etc.).

D'une certaine façon, la veille lexicale hérite de la version papier du *TLF* son approche sémasiologique. Elle consiste à repérer dans des corpus de textes les phénomènes néologiques attestés ou susceptibles de mener à de nouvelles attestations. Toutefois, les objectifs lexicographiques initiaux ont été reconsidérés dans le courant des années 2000 : le *TLF* constituait le projet lexicographique du XXe siècle parce qu'il s'est adossé à la mécanisation puis à l'informatisation des ressources (corpus) et des procédés (concordanciers et tris morphosyntaxiques). Plutôt que de prolonger cette expérience, il importait davantage de préparer le dictionnaire du siècle suivant. C'est donc dans cette intimidante perspective qu'il convient de situer aujourd'hui la veille lexicale. Il ne s'agit plus de maintenir un dictionnaire, mais de participer au projet lexicographique du XXIe siècle.

Fort heureusement, il reste 90 ans avant l'ultime échéance et plutôt que de s'égarer dans de vaines conjectures, on se posera la question d'une lexicographie localisée à court ou à très moyen termes, c'est-à-dire de la lexicographie telle qu'elle peut se développer dans l'environnement social, économique et culturel européen aujourd'hui dominé par une

technophilie radieuse. L'informatisation massive des différents secteurs de la société laisse en effet entrevoir à grands traits l'évolution à court terme de la production lexicographique : disparition du support papier, exploitation par l'ordinateur (*machine readable*) des dictionnaires, automatisation des tâches préparatoires à la production lexicographique, voire à la production lexicographique elle-même[42], etc. Ce sont donc des projets de *modélisation des phénomènes néologiques* et d'*automatisation* de la veille dont il s'agit ici. Comme on l'a laissé entrevoir, l'objectif en est notamment la mise en place d'une plateforme de veille lexicale alimentée à intervalles réguliers de textes (collectés par exemple sur Internet) dont la fonction est d'étudier le lexique en diachronie (néologismes, nouveaux emplois, etc.).

Les néologismes, comme tout phénomène linguistique, sont le fruit de libertés et de contraintes. Liberté qu'offrent les langues de créer de nouveaux mots par la lexicalisation de thèmes sémantiques émergents, liberté que donnent certaines situations de production et d'interprétation faiblement contrôlées (on songe principalement à la conversation, qu'elle soit entendue au sens classique ou moderne, et électronique – forum, *chat*, etc.). Mais contraintes également ; contraintes cognitives, certes (on conçoit), mais contraintes linguistiques : lexicologiques, sémantiques et textuelles. Ce sont ces contraintes-là que nous avons évoquées ici.

Les néologismes non techniques (c'est-à-dire hors langue de spécialité) apparaissent le plus souvent dans des situations peu contraintes, mais pour qu'un mot intègre la langue, la tradition lexicographique impose qu'une autorité la valide. Si la machine se substitue au lexicographe, il nous faut trouver une autre forme d'autorité. Elle peut être de deux ordres : (a) Une autorité *statistique* : *i.e.* la fréquence d'une forme nouvelle, sa stabilisation à la fois orthographique (*« bloggeur »*, *« blogueur »*) mais aussi dans ses usages – même si ceux-ci peuvent être variés ; (b) une autorité éditoriale : Internet bouleverse les normes en matière de sanction éditoriale car s'improviser éditeur et mettre en ligne les textes est à la portée de beaucoup et est, à l'heure actuelle, valorisé par les initiatives du « Web participatif ». De fait, les index des moteurs de recherche généralistes n'intègrent pas de règles d'autorité éditorialement valides (Google s'appuie sur la popularité et le liage des pages), à l'inverse toutefois des moteurs de recherche spécialisés (Google Scholar pour les publications académiques, Cismef pour les publications médicales, etc.).

Pour une recherche en veille lexicale, il importe donc de statuer sur les ressources possibles et probablement d'exclure les sources sans autorité éditoriale, à contre-courant de

[42] On peut objecter que le modèle dominant actuellement est, certes informatisé, mais de nature humaine et collaborative (*wiki* ; *wikipedia*, *wiktionary*, etc.).

la tendance actuelle qui consiste à réduire les textes au statut préscientifique de ressource (voir l'introduction). Car l'occurrence d'un néologisme dans un commentaire de blog n'a pas le même poids ni la même validité qu'une occurrence dans un article de presse ou dans le billet dudit blog, si son auteur bénéficie d'une autorité auctoriale. Dès lors, il est probable, par exemple, que le commentaire de blog soit à exclure (peut-être provisoirement) des recherches en veille lexicale lorsque celles-ci ont vocation lexicographique.

L'application lexicographique d'une lexicologie textuelle pourrait, dans la continuité méthodologique du *TLF*, être un observatoire du lexique dont nous établissons ici trois aspects :

1*) Aller aux sources*. Il s'agit de procéder à la collecte automatique et périodique de sources textuelles variées et sanctionnées (dans la presse notamment parce que, heur ou malheur, les médias sont largement prescriptifs en matière d'usage aujourd'hui) ; enrichissement de ces sources à des fins d'analyse (étiquetage morphosyntaxique, syntagmatique, annotation sémique)[43].

2) Prédire la néologie. Il s'agit de repérer des formes sémantiques stabilisées ou en cours de stabilisation, qui peuvent constituer des « candidats-concepts ». Le caractère transdicursif et transgénérique d'une forme sémantique, ou ses variations sémiques en genre, en discours, permet de la qualifier et est susceptible de donner des indications sur son destin néologique. Par exemple, une forme sémantique transgénérique est potentiellement un concept du domaine à la différence d'une forme sémantique propre à un genre. Il est donc possible de localiser de manière statistique des formes sémantiques stabilisées dans un genre et, de surveiller leur diffusion dans d'autres genres du même domaine. Il est possible que la lexicalisation de la forme sémantique ou son caractère imminent soit détectable de façon automatique par des indices phraséologiques ou grammaticaux. D'ailleurs, la lexicalisation peut vraisemblablement modifier sensiblement l'environnement textuel de la forme sémantique. Le concept, en contractant une forme, crée un vide sémantique à investir et donne lieu à un renouvellement de l'environnement sémique. Par ailleurs un concept, une fois lexicalisé, est davantage susceptible de manipulation (mise en relation avec d'autres concepts, migration vers d'autres problématiques, ou d'autres thématiques, etc.).

[43] Dans ce cadre, le CNRTL (http://www.cnrtl.fr) a récemment passé un accord avec l'*Est Républicain* pour une collecte annualisée du quotidien lorrain.

3) Colliger des passages. Pour suivre l'évolution sémantique d'une unité lexicale, la solution est de rassembler des *passages** dans lesquels elle est actualisée, par période et par discours, de manière à en étudier les variations (*cf.* les études préparatoires rapportées *supra*, Reutenauer *et al.* 2009, Reutenauer *et al.* en préparation). On peut sélectionner ces passages en fonction de leur similarité à des types et ainsi construire des typologies d'emplois. Parce que, comme le disait (Guiraud, 1960 : 19), un mot « se définit finalement par la somme de ses emplois », les passages sont susceptibles de fournir le matériau sémique pour la construction de signifié et le matériau textuel pour la rédaction de définition lexicographique, de la même manière que les *concordances* en usage lors de l'élaboration du *TLF* (il n'est d'ailleurs pas impossible que l'assortiment des deux permette la réalisation automatique ou semi-automatique des définitions). Les collections de passages, organisées, structurées en fonction des sources et de leur variation par rapport à un type, pourraient constituer un nouveau matériau de *lexicographie textuelle*, ni dictionnaire, ni corpus.

5.2. Péremption du dictionnaire de sèmes

Il y a un paradoxe apparent, que nous avons souligné à plusieurs reprises dans ce mémoire. Partisan d'une linguistique des textes pour laquelle la langue générale n'existe pas et qui privilégie une approche variationniste (il y a différentes pratiques sociales auxquelles correspondent différentes pratiques linguistiques), nous avons toutefois illustré notre propos à partir d'études qui, pour la plupart, ont parmi leurs principaux protagonistes un dictionnaire de sèmes intrinsèquement généraliste, puisqu'il est dérivé du *TLF*. Il ne faut pas y voir un biais dans le raisonnement mais la manifestation d'un principe de réalité empirique. Le dictionnaire de sèmes n'est pas une fin en soi, il n'est qu'un moyen, le moyen d'étudier les objectivations sémantiques en corpus suivant une approche sémique. Les sèmes d'un corpus sont dans une très large mesure contraints par l'intertexte discursif et domanial, le fameux corpus réflexif qui comprend les clés interprétatives d'un texte. Si les annotations effectuées dans les études signalées dans ce chapitre et le précédent (Grzesitchak *et al.* 2007, Reutenauer *et al.* 2009, Reutenauer *et al.* en préparation), ne correspondent pas à cet intertexte, on peut affirmer néanmoins qu'il s'agit d'études préparatoires destinées à évaluer les potentialités d'une annotation sémique et à déterminer les moyens de construire les dictionnaires *ad hoc*. En d'autres termes, le dictionnaire de sèmes utilisé permet de réaliser une image sémique initiale dont le destin est d'être périmée au plus vite par différentes méthodes[44], lesquelles reste à développer

[44] Une collaboration a été amorcée en 2006 avec le LIP6 (Paris). Le LIP6 a en effet réalisé un prototype de système de segmentation thématique qui semblait adapté à la recherche des formes sémantiques : « Ce prototype utilise un algorithme de partitionnement sur les sèmes, l'algorithme des X-moyennes (Pelleg *et al.*, 2000). Il permet de trouver les différentes classes de sèmes dans l'espace des documents.

(identification et suppression du bruit, sélection des candidats-sèmes pertinents, neutralisation des candidats-sèmes impertinents, enrichissement des signifiés par propagation contextuelle, etc.). C'est pourquoi nous avons à un moment préféré parler d'une *collection de dictionnaires* plutôt que d'un dictionnaire. Tant que la procédure permettant de générer rapidement (c'est-à-dire automatiquement) le dictionnaire sémique correspondant à un corpus ou à une archive n'est pas dessinée, la question de l'applicabilité de ces recherches à des projets industriels (par exemple en Recherche d'Information, en détection d'opinion ou en profilage de textes) ne semblent pas pouvoir se poser. Car à l'opposé des « sémantiques légères » (Perlerin 2004) qui proposent des outils ciblés pour des applications précises, les recherches prospectives présentées *supra* relèvent d'une sémantique *enrichie*, complexe à mettre en œuvre et qui nécessite des temps de traitement incompatibles avec les exigences industrielles. Mais l'enjeu, à termes, est de mieux comprendre les mécanismes nécessaires à la production et à l'interprétation des textes et d'en tirer partie pour des applications ciblées. Ces recherches ne sont pas achevées, ce mémoire en fait un bilan intermédiaire. Elles doivent notamment être éprouvées sur d'autres langues[45].

C'est une extension de l'algorithme classique des K-moyennes dans lequel le nombre de classes est estimé au lieu d'être fixé par l'utilisateur. Les sèmes appartenant aux mêmes ensembles (les classes de l'algorithme de partitionnement) auront la même représentation dans le nouvel espace. L'algorithme commence avec deux partitions et décide itérativement quand il convient de couper une classe en utilisant un critère d'information bayésien (CIB). Si en partitionnant, ce critère décroît, on garde alors l'ancienne partition. L'hypothèse sous-jacente est que les sèmes qui cooccurrent fréquemment sont sémantiquement proches, c'est une hypothèse courante en recherche d'information, que l'on retrouve également dans d'autres méthodes de réduction de la dimension comme la méthode de projection *Latent Semantic Analysis*. Par rapport à cette dernière qui est basée sur des combinaisons linéaires de mots (ici des sèmes) difficilement interprétables, notre méthode permet de conserver pour chaque classe les mots représentatifs » (Extrait du projet DIXEM$_1$, soumis à l'appel « Corpus et outils de la recherche en sciences humaines et sociales » de l'ANR, édition 2006 ; paragraphe rédigé par Massih Amini).

[45] Un projet applicatif ciblé et « léger », inspiré de ces recherches, est actuellement à l'étude.

Ouverture

La description du lexique constitue un enjeu pour la linguistique à l'heure du document numérique. Par tradition, mais aussi parce qu'il s'instancie dans des « chaînes de caractères », c'est-à-dire dans de l'immédiatement perceptible, il est privilégié par les informaticiens. Son appréhension théorique est logiquement dominée par un paradigme ontologique et universalisant incompatible avec la variété des textes — variété en pratiques et en langues. Il importe que la linguistique adapte ses objets et ses théories aux nouvelles pratiques du texte (Internet, dématérialisation) et à la diversité des langues et des usages culturels. Mais il ne s'agit pas là de soumettre la linguistique au dictat de la technophilie ambiante. Il ne s'agit pas de ne penser les sciences du langage qu'en termes d'applications et de la Recherche & Développement. Il s'agit de la confronter aux textes modernes, à la textualité contemporaine de façon à ce qu'elle assume son rôle de description des observables. La linguistique doit être en mesure de décrire aussi bien les articles de la presse arabe, la poésie italienne du quattrocento, le blog en vogue au Japon ou le journal des frères Goncourt, parce que c'est dans cette diversité que la langue se révèle — c'est dans la diversité des *textes*, qu'ils soient oraux, ou imprimés, matériels ou immatériels.

C'est dans cette perspective que nous avons souhaité, dans ce mémoire, aborder le lexique. Le mot — le signe, puissance de la langue, réceptacle de la connaissance, du concept, est emblématique du phénomène langagier, et nous avons tenté d'illustrer la thèse qu'une description du mot par le texte comme trace des pratiques est non seulement possible et nécessaire.

Des textes au mot. Analyse sémantique pour l'accès à la connaissance

Car le mot est aussi redevable au texte de son contenu sémantique que l'inverse. L'identification et la caractérisation des phénomènes de conceptualisation, qu'elle soit création lexicale ou renouvellement du contenu sémantique d'une unité lexicales existante,

a plus d'une incidence. Incidence en linguistique générale (décrire le lexique), mais aussi incidence dans une perspective cognitive (comprendre la conceptualisation) et enfin, incidence sur les applications possibles de la linguistique, sa réponse à la demande sociale, et donc sur sa pérennité en tant que science qui a quelque chose à dire sur le monde. En cela, l'ingénierie des connaissances constitue peut-être un des terrains d'aventure que la linguistique a, jusque là trop négligés. Dans les conférences organisées par l'Association Françaises d'Intelligence Artificielle (AFIA), qui réunissent des centaines de chercheurs, les linguistes se comptent sur les doigts d'une main. Mais l'expérience montre qu'ils sont écoutés, que leur connaissance du phénomène linguistique, de sa complexité, de sa diversité intéresse et trouve un écho dans les préoccupations, les échecs rencontrés par les ingénieurs de la connaissance. Le monde des objets (ontologiques) qui est le leur se heurte d'une part à la doxa, aux opinions, aux valeurs, au caractère éminemment humain de l'interprétation et d'autre part, à la variété des langues et des pratiques, des façons de percevoir le monde, des façon de *construire les connaissances* d'un bout à l'autre de la planète, quand même l'on voudrait imposer l'anglais comme modèle de la pensée et de la connaissance, celui-ci se fragmenterait, se créoliserait, se pidginiserait incessamment.

C'est donc là un travail des années à venir : réinscrire la linguistique, et plus particulièrement la sémantique des textes en corpus, dont le potentiel descriptif et explicatif a été montré, dans le champ des sciences de la connaissance et de leurs applications contemporaines et surtout à venir ; enfin, et c'est là le moyen autant que la fin, expliquer l'élaboration des connaissances, la formation des concepts dans la variation des textes, des pratiques et des langues.

Glossaire

Candidat-sème : étiquette sémantique résultant de traitements automatiques. En validant le candidat-sème, le linguiste peut lui octroyer un statut de *sème*. La notion de sème-candidat est évoquée dans (Valette 2008) sous l'appellation « sème potentiel » et explicitée à partir de (Reutenauer *et al.* 2009). Les candidats-sèmes y sont notés *étiquette#catégorie* (par exemple, chômage#subst).

Classe sémantique : domaine, taxème, dimension.

Concept : forme sémantique stabilisée et lexicalisée de façon synthétique (lire chap. 3, § 2.1). Avant la lexicalisation, on peut parler de préconcept, ou de notion.

Conservatisme lexical : Contraire de la créativité lexicale. Voir cette entrée.

Créativité lexicale : rapport entre la richesse lexicale (Muller 1977) et la richesse néologique. Un indice de créativité lexicale élevé correspond à un recours plus important à des néologismes variés proportionnellement à la richesse lexicale mesurée.

Dimension : classe souvent binaire exprimant une opposition générale (féminin/ *vs* /masculin/, /animé/ *vs* /inaminé/, /mélioratif/ *vs* /dépréciatif/, etc.).

Discours : ensemble d'usages linguistiques codifiés attachés à un type de pratique sociale ((Rastier 2001).

Document numérique : document poly-sémiotique dématérialisé destiné à être lu sur un écran et non sur papier. Ex. Un *blog* est un document numérique.

Domaine : ensemble de taxèmes correspondant à une pratique déterminée

Fond sémantique : ensemble des isotopies d'un texte

Forme sémantique : regroupement syntagmatique de sèmes stabilisé en corpus

Genre : programme de prescriptions qui règlent la production et l'interprétation d'un texte (Rastier 2001).

Isotopie : récurrence d'un sème de nature et d'empan variés (de la collocation à l'intertexte).

Néosémie : néologie sémantique**.**

Notion : forme sémantique stabilisée non lexicalisée (équivalent au préconcept). Voir aussi Concept.

Objectivation sémantique : structure cohérente de sèmes correspondant à une unité sémantique (signifié, fond sémantique, forme sémantique)

Passage : association stabilisée d'un fond et d'une forme sémantique caractéristique d'un usage lexical, d'un texte ou d'un corpus (*cf.* Rastier 2007 pour une définition *in extenso*).

Préconcept : voir notion

Réseau sémique : Fond ou forme sémantique (i.e. les objectivations sémantiques à l'exception du signifié).

Ressourcisme : position académique et scientifique visant à la constitution de ressource sans objectifs fonctionnels déterminés ni conditionnés par une application ou une gamme d'applications. Les critères d'évaluation de la ressource sont dominés par une vision formelle : disponibilité, couverture, maintenance (pérennité, évolutivité). D'un point de vue scientifique, le respect de ces critères donne lieu à des initiatives de normalisation telles que la *Text Encoding Initiative (TEI)* ou *Lexical Markup Framework (LMF)* et d'un point de vue académique, à la *Common Language Ressources and Technology Infrastucture (CLARIN)*. (Valette 2008).

Richesse néologique : Mesure de la néologie d'un texte fondée sur le rapport entre le nombre de néologismes et le nombre de formes. La richesse néologique est dite

théorique lorsque l'on traite des candidats à la néologie. Voir aussi créativité lexicale.

Sème : propriété sémantique d'ordre métalinguistique résultant d'une validation par le linguiste. Les regroupements paradigmatiques de sèmes constituent des fonds et des formes sémantiques. Les regroupements syntagmatiques des sémèmes ou des signifiés. Voir aussi candidats-sèmes. Ex. Dans le signifié de 'Sophora', /ornemental/ est un sème.

Sème générique : sème partagé par les signifiés d'une même classe sémantique dont il assure la cohésion. Ex. /domestique/ dans la classe //animaux de compagnie//

Sème spécifique : sème propre à un signifié et qui en assure l'originalité dans une classe sémantique. Ex. /domestique/ pour *« chien »* dans la classe //canidés//.

Sème réflexif (ou candidat-sème réflexif) **:** candidat-sème correspondant au signifiant du mot considéré, attribué par défaut à tous les mots par la plateforme d'annotation. Par exemple, /chien/ est le sème réflexif de *« chien »*.

Sémème : voir signifié

Sémie : voir signifié

Signifié : contenu sémantique d'un signe, exprimé en collection de sèmes (*sémème* et *sémie* sont deux manières de qualifier le signifié)

Stable : dont la structure sémique varie peu (notamment statistiquement) dans un corpus donné – dans un genre, un discours ou un domaine.

Taxème : Petite classe sémantique correspondant à une situation pratique précise. La cohésion de la classe est assurée par les sèmes génériques.

Zonage isotopique : contiguïté de plusieurs isotopies participant à l'identification de passages.

Bibliographie

Beust, P. et Th. Roy (2006) Prendre en compte la dimension globale d'un corpus dans la contextualisation du sens : expérimentations en informatique linguistique, *GLOTTOPOL, 8, Traitements automatisés des corpus spécialisés : contextes et sens*, 52-72.

Blumenthal, P., Hausmann, F.J. éds. (2006), Collocations, corpus, dictionnaires, *Langue française*, Paris, Larousse.

Bourigault D., Slodzian M. (2000), « Pour une terminologie textuelle », in *Terminologies Nouvelles*, n° 19 : 29-32.

Bourion É (2001), *L'aide à l'interprétation des textes électroniques*. Thèse de doctorat, Université Nancy 2 ; publié sur Texto ! Textes et cultures (http://www.revue-texto.fr).

Cabré, MM. T. (1999) *La terminología. Representación y comunicación*, IULA, UPF, Barcelona.

Caillet, M., Pessiot, J.F., Amini, M.-R., Gallinari P. (2004) « Unsupervised Learning with Term Clustering for Thematic Segmentation of Texts », *Proceedings of the 7th Recherche d'Information Assistée par Ordinatieur (RIAO 2004)* pp. 648—656, Avignon, France

Condamines, A., éd. (2005) *Sémantique et corpus*, Paris, Hermès.

Corblin, F. Gardent, C., éd. (2005) *Interpréter en contexte*, Traité IC2, Série Cognition et traitement de l'information, Hermès science publications.

Cruse, D-A (1986). *Lexical Semantics*. Cambridge : Cambridge University Press.

Cruse, D-A (1995). *Polysemy and relted phenomena from a cognitive linguistic viewpoint.* Dans Saint-Dizier P. and Viegas E. (eds).

Crestan É, El-Bèze M., de Loupy Cl. (2003) « Peut-on trouver la taille de contexte optimale en désambiguïsation sémantique? », in *TALN 2003*, Batz-sur-Mer, 11-14 juin 2003.

Daille, B. (1994) *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. Thèse de Doctorat en Informatique Fondamentale. Université Paris 7.

Dendien, J. & Pierrel, J.-M. (2003) « Le Trésor de la Langue Française informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence », *TAL*, 44/2, 11-37.

Desclès, J-P (2004). *Reasonning and aspecto-temporal calculus. Reasonning and Language*. D. Van Der Veken, Kluwer Pusblisher.

Duteil-Mougel, C. (2004) « Introduction à la sémantique interprétative », *Texto ! Textes et cultures*, http://www.revue-texto.net/.

Enjalbert, P. (2005) *Sémantique et traitement automatique du langage naturel*, Paris, Hermès.

Fellbaum C. (1998), *WordNet: an Electronic Lexical Database*, Cambridge MA, MIT.

Fillmore Ch. J. (1976) "Frame semantics and the nature of language", in *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, Volume 280: 20-32.

Fillmore Ch. J., Baker C. F., Sato, H. (2002) "The FrameNet Database and Software Tools", in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*, Las Palmas: 1157-1160.

Fillmore, G. (1968). *The case for case. Universals in Linguistic Theory*. Bach and Harms. p.1 - 90

Fillmore, C.J., Jonhson, C., Petruck, M.R.L. (2003), "Background to FrameNet", *International Journal of Lexicography*, 16, 1, 235-250.

Fuchs, C., Habert, B., éd. (2004). Traitement automatique et ressources numérisées pour le français, *Le français moderne*, Vol. 72, n°1.

Godard, D et Jayez, J (1995) Types nominaux et anaphores : le cas des objets et des événements. Dans *Anaphores Temporelles et (in-)cohérence*. Cahiers Chronos.

Greimas, A.J. (1966) *Sémantique structurale*, Paris, PUF.

Gross G. (1975), *Méthode en syntaxe*, Hermann, Paris.

Gross G. (2005), « Un dictionnaire électronique des adjectifs du français ». *Cahiers de Lexicologie*, n°86 : 11-33.

Grzesitchak, M., Jacquey, E. Valette, M. (2007) « Systèmes complexes et analyse textuelle : Traits sémantiques et recherche d'isotopies », *ARCo'07 – Cognition, Complexité, Collectif., Acta-Cognitica*, 227-235.

Guiraud, Pierre, (1960) *Problèmes et méthodes de la statistique linguistique*, Paris, PUF.

Habert, Benoît (2005). « Portrait de linguiste(s) à l'instrument », *Revue Texto,* vol. X, n°4.

Hanks, Patrick, James Pustejovsky (2005) « A Pattern Dictionary for Natural Language Processing", *Revue française de linguistique appliquée*, 2005, X, 2 (63-82).

Ide N., Véronis J. (1998), « Word Sense Disambiguation: The State of the Art", *Computational Linguistics*, 24/1: 1-40.

Jousse A.-L., Polguère A. (2005), Le DiCo et sa version Dicouèbe, *Document descriptif et manuel d'utilisation, Version du rapport 1.0 – 19 avril 2005*, Observatoire de linguistique Sens-Texte (OLST), Université de Montréal.

Kerbrat-Orecchioni C. (1996), « Texte et contexte » in Schmoll, P. (éd.) *"Contexte(s)", Scolia* 6 : 39-60.

Kleiber, G. (1990). *La sémantique du prototype*, Paris, PUF.

Kleiber G. (1997), « Sens, référence et existence : que faire de l'extra-linguistique ? », *Langages*, Vol.127 : 9-37.

Kleiber, G. (1999). Problèmes de sémantiques : la polysémie en question. *Sens et structure*. Villeneuve d'Asq : Presses Universitaires du Septentrion.

Labbé C., Labbé D. (2003), « La distance intertextuelle », *Corpus*, 95-118.

Lafon, P. (1984) *Dépouillements et statistiques en lexicométrie*, Genève-Paris, éd. Slatkine – Champion.

Lebart, L. et Salem, A., 1994, *Statistique textuelle*, Paris, Dunod.

Lecolle, M., (2007) « Polysignifiance du toponyme, historicité du sens et interprétation en corpus. Le cas de Outreau », *Corpus n*°6, 101-125

L'Homme, M.C. 2004. *La terminologie : principes et techniques*, Montréal : PUM.

Loiseau S. (2005), « Thématique et sémantique contextuelle d'un concept philosophique », in La Linguistique de corpus. *Actes des deuxièmes journées de la linguistique de corpus,* G. Williams (éd.), Rennes, PUR : 129-140.

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

Martin R. (2001) *Sémantique et automate*. Paris : PUF.

Mayaffre Damon (2002), « Les corpus réflexifs : entre architextualité et hypertextualité »,

C*o*rpus, 1, p. 51-69.

Mayaffre Damon (2005), « Rôle et place des corpus en linguistique : réflexions introductives », Actes des Journées d'Etude TOUlousaines JETOU 2005, Rôle et place des corpus en linguistique, Toulouse, 2005, p. 5-17.

Missire Régis (2006) *Sémantique des textes et modèle morphosémantique de l'interprétation*, thèse de doctorant Université de Toulouse II Le Mirail, Texto ! Textes et cultures, http://www.revue-texto.net/.

Muller, Charles (1964) *Essai de statistique lexicale. L'illusion comique de Pierre Corneille*, Paris, Klincksieck.

Namer, F. Dal, G. Hathout, N. (2004). « Morphologie constructionnelle et traitement automatique des langues: le projet MORTAL ». *Lexique 16,* Pierre Corbin (éds). Villeneuve d'Ascq: Presses Universitaires du Septentrion, pp.199- 229.

Ollinger, S., Valette, M. (à paraître) « La créativité lexicale : des pratiques sociales aux textes », *Actes du 1er Congrès International de néologie des langues romanes (Barcelone, 07 - 10 mai 2008) CINEO'08*.

Pelleg D., Moore A. (2000), « X-means: Extending K-means with Efficient Estimation of the Number of Clusters », *Proceedings of 17th International Conf. on Machine Learning*, 727-734.

Perlerin, V. (2004*) Sémantique légère pour le document. Assistance personnalisée pour l'accès au document et l'exploration de son contenu*, Thèse de doctorat (informatique).

Pincemin, Bénédicte (1999) « Sémantique interprétative et analyses automatiques de textes : que deviennent les sèmes ? », *Sémiotiques*, n° 17, décembre 1999, pp. 71-120.

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

Pottier. B. (1974) *Linguistique générale. Théorie et description*. Paris, Klincksieck.

Pottier, B. (1987) *Théorie et analyse en linguistique*, Hachette, Paris (rééd. 1992).

Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press.

Ramdani, Egle (2007) *Du dictionnaire de langue au lexique TAL – la construction d'une ressource pour l'annotation sémantique des textes*, Mémoire de Master 2, M. Valette (dir.), INaLCO, Paris.

Rastier F. (1989), *Sens et textualité*. Paris : Hachette.

Rastier F. éd. (1995), *L'Analyse thématique des données textuelles : l'exemple des sentiments*. Paris : Didier.

Rastier, F. (1998) « Le problème épistémologique du contexte et le statut de l'interprétation dans les sciences du langage », *Langages*, 129, 97-111.

Rastier F. (2001), *Arts et sciences du texte*. Paris : PUF.

Rastier, F. (2004) « Ontologie(s), *Revue des sciences et technologies de l'information, série : Revue d'Intelligence artificielle*, 2004, vol. 18, n°1, 15-4.

Rastier, F. (2007) « Passages », *Corpus*, B. Pincemin, éd., 6, 125-152.

Rastier (2008) « Sémantique du web *vs* sémantic web – Le problème de la pertinence », *Pour une science des textes instrumentée,* M. Valette, éd., *Syntaxe & Sémantique,* n°9.

Rastier, François, Cavazza, Marc, Abeillé, Anne (1994), *Sémantique pour l'analyse. De la linguistique à l'informatique*, Paris, Masson.

Rastier F., Valette M. (2009) « De la polysémie à la néosémie », *Le français moderne*, S. Mejri, éd., *La problématique du mot*, 77, 97-116.

Reutenauer, C., Valette, M. Jacquey, E. (2009) « De l'annotation sémique globale à l'interprétation locale : environnement et image sémiques d' "économie réelle" dans un corpus sur la crise financière », *Colloque ARCo'09, Interprétation et problématiques du sens*, 9-11 novembre 2009, Rouen.

Reutenauer C., Lecolle, M., Jacquey E., Valette, M. (en préparation), « *Outreau* en n sèmes, *Outreau* en cinq temps. Diachronie de la représentation sémique d'une unité lexicale », *Du thème au terme. Emergence et lexicalisation des connaissances,* M. Slodzian, M. Valette, éd., *TIA 2009*, Toulouse, 20 novembre 2009.

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.

Schank R. C. & Abelson R. P. (1977), *Scripts, Plans, Goals and Understanding*. Lawrence Erlbaum, Hillsdale, New Jersey.

Schmid G. (1994) « TreeTagger – a language indépendant part-of-speech tagger », *Procee-dings of EACL-SIGDAT 1995*, Dublin, Ireland., 44-49.

Slodzian M. (1999), WordNet et EuroWordNet – Questions impertinentes sur leur pertinence linguistique. *Sémiotiques*, n°17, 51-70.

Tartier, A. (2004) *Analyse automatique de l'évolution terminologique : variations et distances.* Thèse de doctorat en informatique, U. de Nantes.

Tognini-Bonelli, E. (2001) *Corpus Linguistics at Work*, Amsterdam, John Benjamin's Publishing.

Valette, M. (2004) « Sémantique interprétative appliquée à la détection automatique de documents racistes et xénophobes sur Internet », *Approches Sémantiques du Document Numérique, CIDE.7*, P. Enjalbert et M. Gaio, eds, 215-230.

Valette, M. (2008) « A quoi servent les lexiques sémantiques ? Discussion et proposition », *Cahiers du CENTAL*, n°5, M. Constant, A. Dister, L. Emirkanian & S. Piron, éd., 43-58.

Valette M., Estacio-Moreno A., Petitjean É., Jacquey É. (2006), « Éléments pour la génération de classes sémantiques à partir de définitions lexicographiques. Pour une approche sémique du sens », in *Verbum ex machina*, *TALN 06*, vol. 1 : 357-366.

Véronis, Jean (2000) « Annotation automatique de corpus : panorama et état de la technique ». In J.-M. Pierrel (Ed.*), Ingénierie des langues* (pp. 111-129). Paris: Editions Hermès.

Véronis, Jean (2004). Quels dictionnaires pour l'étiquetage sémantique ? *Le Français moderne*, Vol. 72, n°1, *Traitement automatique et ressources pour le français*, C. Fuchs, B. Habert, éd., 27-38.

Viprey, Jean-Marie (2005) « Corpus et sémantique discursive : éléments de méthode pour la lecture des corpus », *Sémantique et corpus*, Anne Condamines, eds, Hermès, Paris, pp. 245-276.

Vossen, P. (Ed.) *Eurowordnet : A Multilingual Database With Lexical Semantic Networks.* Dordrecht: Kluwer Academic Publishers.

Zweigenbaum P., Habert B. (2004), « Accès mesurés aux sens », *in* Mots. Les langages du politique, n°74 : 93-106.

Zweigenbaum, Pierre & Jean Charlet (2007) « Ressources termino-ontologiques pour le domaine médical. », L. Depecker, V. Dubois, and Chr. Roche éds, Terminologie et ontologie : descriptions du réel, Le savoir des mots, Paris, 83-110.