Méthodes pour la veille lexicale

Mathieu Valette

Mathieu Valette « Méthodes pour la veille lexicale »,
version soumise [3 septembre 2009] et acceptée – à paraître
Actes de la journée d'étude Le dictionnaire électronique. Quelles perspectives pour les sciences humaines et sociales ? (Kénitra, le 7 décembre 2007), Leila Messaoudi éd., *Publication du laboratoire Langage et société*, Université Ibn Tofail Kénitra (Maroc)

Méthodes pour la veille lexicale

Mathieu Valette

ATILF (CNRS, Nancy)

mvalette@atilf.fr

1. Contexte

La *Veille lexicale* correspond 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, anciennement INaLF) au début des années 2000 et visait explicitement la mise à jour du *Trésor de la Langue Française* (TLF) notamment dans sa version informatisée (TLFi, 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 appellerons des *néosémies* : nouvelles acceptions, nouvelles phraséologies), et enfin, des nouveaux comportements morphosyntaxiques (changement de genre, pronominalisation, etc.). On rassemblera ces différentes manifestations sous l'appellation générale de *phénomènes néologiques*.

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 au siècle suivant, il importait davantage à une institution telle que le CNRS 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[1], etc. Ce sont donc des projets de *modélisation des phénomènes néologiques* et d'*automatisation* de la veille dont il s'agit ici. 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.). Mais pour des raisons éditoriales, nous n'aborderons ici que certains aspects liés à la modélisation, réservant les aspects ingénieriques et logiciels à des publications *ad hoc*[2].

2. Éléments pour un modèle sémantique de la néologie

Le cœur de notre proposition n'est pas étranger à l'épistémologie à l'origine du projet du TLF : le lexique ne peut être appréhendé qu'à travers des textes. La posture est banale et largement partagée dans la communauté, mais nous en radicalisons la teneur en adoptant une théorie du texte pour décrire le lexique, la sémantique textuelle (Rastier 1987, 2001).

La tradition privilégie le lexique, et plus particulièrement les groupes nominaux, dans la détermination des concepts. Or, la linguistique, depuis Saussure, pose que le versant psychique d'un signe, le *signifié*, ne se confond pas avec le concept. Un concept n'est donc pas systématiquement lié à un signe particulier, mais peut s'actualiser dans un thème ou une forme sémantique, c'est-à-dire dans un groupement de traits sémantiques (*sèmes*) non nécessairement lexicalisé. Autrement dit, à un concept ne correspond pas forcément une unité lexicale. La lexicalisation d'une forme sémantique, qui aboutit à la formation du concept, ne doit pas être envisagée exclusivement comme sa naissance, ni même comme l'aboutissement de la conceptualisation. Elle s'apparente davantage à un état de stabilisation provisoire, correspondant à un usage circonscrit d'un point de vue socioculturel et temporel.

2.1. Veille lexicale *vs.* veille terminologique

La problématique de la veille lexicale se mesure à l'aune de sa parente, la veille terminologique[3] 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 (Rastier & Valette 2009). 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 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[4], 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.

2.2. Conditions théoriques

Les principaux présupposés nécessaires à notre propos, empruntés ou inspirés de la sémantique textuelle (Rastier, *op. cit.*) sont les suivants :

(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.

(ii) Le sens se décrit en termes de traits sémantiques. La sémantique textuelle hérite de la sémantique structurale (Greimas 1966, Pottier 1974) la notion différentielle de *sème*. Le signe comprend un signifié et un signifiant. Le signifié est composé de sèmes. Les unités lexicales s'organisent en classes sémantiques structurées en fonction de traits sémantiques 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 le contexte de la veille lexicale, nous entendons donner un rôle privilégié à la notion de réseau de sèmes, en particulier à la 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.

2.2. Les phases néologiques

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 ci-dessous donne à voir l'évolution générale des formes sémantiques dans cette perspective[5].

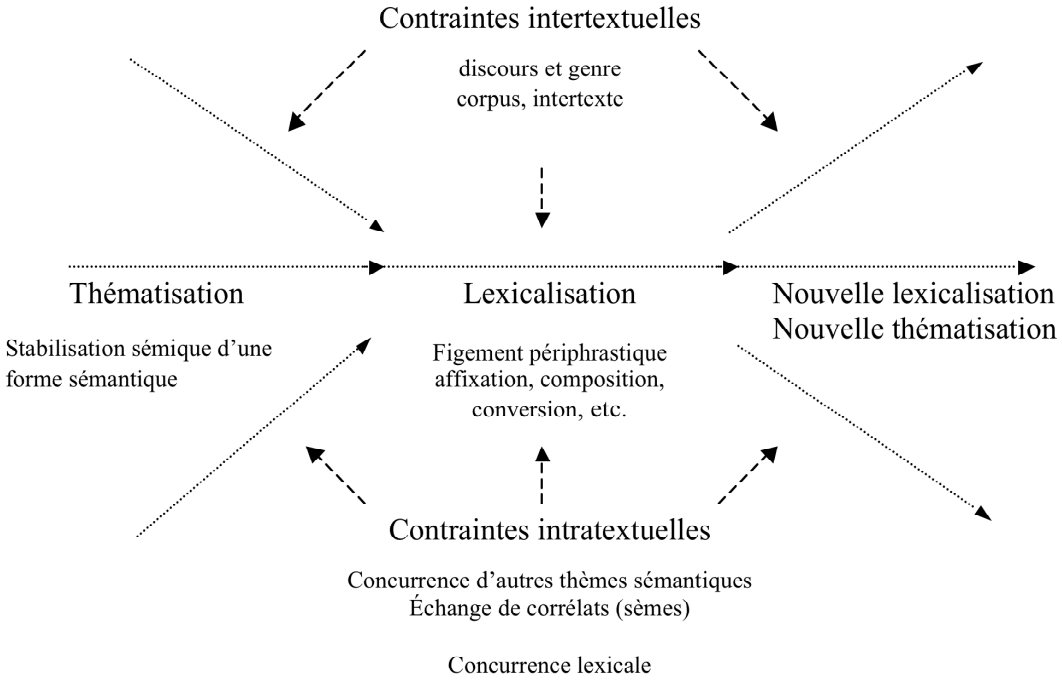

Fig. 1 : Formes sémantique et lexicalisation

Distinguons trois phases dans le développement d'un néologisme. 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 thème(s) en cours de structuration. Elle se caractérise par une instabilité sémantique et une certaine complexité textuelle. Elle est enchâssée dans un réseau complexe d'expressions et de phraséologies. 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 termes de *« post-productivisme »*, *« convivialité »*, etc.).

La deuxième phase, la *lexicalisation*, correspond au figement lexical de la forme sémantique stabilisé. 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 parler de néologisme. 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.

La troisième phase consiste en l'altération du néologisme. C'est typiquement à ce moment-là que l'on peut parler de *néosémie* (cf. paragraphe 4). 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, sociale, 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.

Voilà esquissées nos propositions générales. Elles nécessitent bien évidemment d'être approfondies et validées. Les deux paragraphes ci-dessous présentent différentes réalisations et outils mis en œuvre dans le but d'éprouver notre modèle. Dans le paragraphe suivant, nous présenterons quelques résultats d'une recherche menée sur la détection des néologismes de formes, ce qui nous permettra d'illustrer la question des contraintes intertextuelles. Puis, dans un autre paragraphe, nous exposerons quelques propositions pour la détection de la néologie sémantique (ou néosémie) ; ce sera l'occasion d'illustrer la question des contraintes intratextuelles exercées sur la formation des néologismes.

Certains des travaux qui seront présentés ci-après ont été réalisés en collaboration avec Sandrine Ollinger et Etienne Petitjean. Ils ont bénéficié, bénéficient ou bénéficieront également, à différents degrés, de la participation de Susanne Alt, Alexander Estacio-Moreno, Bertrand Gaiffe, Mick Grzesitchak, Evelyne Jacquey, Etienne Petitjean, Jean-Marie Pierrel, Egle Ramdani, François Rastier, Coralie Reutenauer.

3. Cas de contraintes intertextuelles : genres et créativité lexicale

Le contexte ingénierique de cette recherche 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, il s'agit de 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.

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 / fig. 2 *infra*). Pour une description plus détaillée de l'outil et plus approfondie de l'étude dont il est question ici, on se reportera à Ollinger & Valette *à paraître*.

3.1. Richesse lexicale et richesse néologique théorique

La figure 2 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 3 ont été calculées à partir de l'indice W proposé par E. Brunet et rapporté par Ch. Muller (1977 [1992], p. 196) :

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

où N est, par convention, le nombre d'occurrences de formes, V le nombre de formes et α 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 $\left(\frac{1}{W}\right)\times 100$.

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 α la même constante que précédemment. Le résultat présenté est $\left(\frac{1}{U}\right)\times 100$.

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, Université de Stuttgart). Le tableau et la figure 2 donnent à voir les richesses lexicales et néologiques théoriques du corpus.

	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

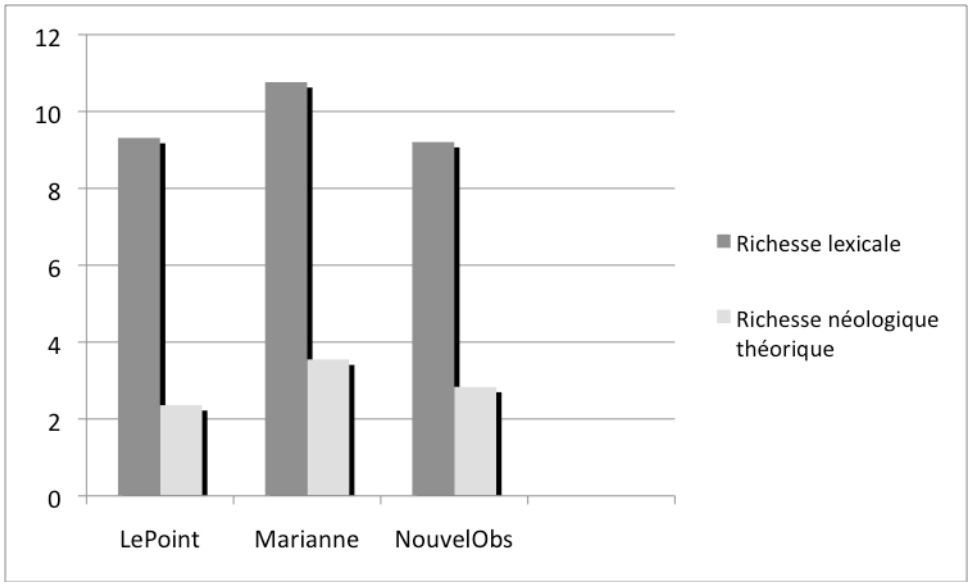

Fig. 2 : Richesse lexicale et richesse néologique théorique du corpus

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[6]. Le schéma de la figure 3 présente les taux de *conservatisme lexical théorique* et de *créativité lexical théorique* des différents corpus.

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

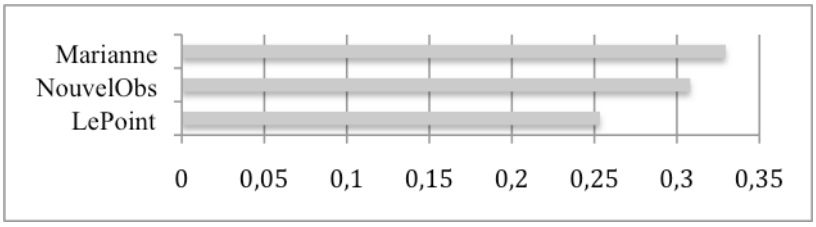

Fig. 3 : Conservatisme lexical (à gauche) et créativité lexicale (à droite) du corpus

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.

3.3. 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 l'hebdomadaire *Marianne* : 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 4 donne à voir sommairement les principaux types de constructions remarquables identifiées.

	Marianne		**Le Point**		**Le Nouvel Observateur**	
	Form.	**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 (-iser, -isation)	7	*starisation*	3	*annualisation*	1	
Dérivation d'ent. nom.	26	*gaudinerie, Sarkozie*	1	*villepeniste*	9	*berlusconien*
Total	**99**		**21**		**25**	

*Fig. 4 : Constructions néologiques remarquables (*Marianne *comparée au* Point *et au* Nouvel Observateur*)*

Nous suggérons que ces choix stylistiques d'importance sont liés à des contraintes intertextuelles ; *Marianne* est connu pour son ton polémique et volontiers pamphlétaire (à l'inverse notamment du *Point*, réputé conservateur) – 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).

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

4.1.1 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*[7], 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 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.

4.1.2. 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[8] (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).

4.1.3 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*.

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)

4.3. L'évolution des signifiés

Certains sèmes participent de façon privilégiée aux réseaux de sèmes 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 mis en italique, dans le court texte suivant extrait de *Bouvard et Pécuchet* de G. Flaubert, tous les mots étiquetés dans le *TLF* comme relevant du domaine de la Botanique et nous avons signalé par un exposant les mots dont le signifié comprend le sème /ornemental/, toujours d'après le *TLF*.

> Alors Pécuchet se tourna vers les *fleurs*$^{/o/}$. 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* à l´ombre, des *pensées* au soleil, couvrit de fumier les *jacinthes*, arrosa les *lys*$^{/o/}$ après leur floraison, détruisit les *rhododendrons*$^{/o/}$ par des excès d´abattage, stimula les *fuchsias*$^{/o/}$ 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*$^{/o/}$ étaient gigantesques ; – et on apercevait, entre ces lignes droites les rameaux tortueux d´un *sophora*$^{/o/}$-japonica qui demeurait immuable, sans dépérir, ni sans pousser.

Cette séquence est d'une homogénéité évidemment exemplaire. Douze fois l'étiquette domaniale 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. Nous faisons donc 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.

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 (cas, par exemple, des néologismes détectés dans l'étude rapportée dans le troisième paragraphe *infra*, cf. aussi Ollinger & Valette, *op. cit.*). 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. Plusieurs propositions ont été faites (Valette, Estacio-Moreno *et al*. 2006, Grzesitchak *et al.* 2007, Valette 2008) pour la réalisation d'un

dictionnaire de sèmes pour l'annotation de corpus. L'exploitation de ce dictionnaire dans la perspective de la veille lexicale est actuellement à l'étude.

5. Pour conclure

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, *tchat*, etc.). Mais contraintes également ; ces contraintes sont linguistiques : lexicologiques certes, mais aussi sémantiques et textuelles. Ce sont ces contraintes-là que nous avons choisi d'évoquer 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. Une tendance actuelle est de considérer les textes non pas comme objet de science mais de les réduire, 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. Or, 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 celui-ci est journalistique 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.

6. Références bibliographiques

Bourigault, D. Slodzian, M. (1999) « Pour une terminologie textuelle », *Terminologies nouvelles*, 19, p. 29-32.

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

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.

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

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.

Jousse, A.-L. (2007) « La néologie dans le pamphlet », *Neologica*, *Revue Internationale de Néologie*, 1, 227 p.

Muller, Ch., (1977) *Principes et méthodes de statistique lexicale*, Paris, Hachette (rééd. Champion 1992).

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*.

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

Rastier, F. (1987) *Sémantique interprétative*, Paris, PUF.

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

Rastier, F. (2005) « Pour une sémantique des textes théoriques », *Revue de sémantique et de pragmatique*, 17, 151-180.

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

Valette, M. (2008) « A quoi servent les lexiques sémantiques ? Discussion et proposition », *Description linguistique pour le traitement automatique du français*, M. Constant, A. Dister, *et al.* (éds), *Cahiers du CENTAL*, n°5 – décembre 2008, PUL, 43-58.

Valette, M., Estacio-Moreno, A., Petitjean, E., Jacquey, E. (2006) « Éléments pour la génération de classes sémantiques à partir de définitions lexicographiques. Pour une approche sémique du sens », *Verbum ex machina (TALN 06)* P. Mertens *et al.* (éds). *Cahiers du CENTAL,* 2.1, UCL PUL Volume 1, pp. 357-366.

---

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

[2] On lira par exemple Ollinger & Valette, à paraître.

[3] 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.

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

[5] On trouvera dans Rastier 2005 une représentation de l'évolution des concepts assez similaire.

[6] Soit, le taux de créativité lexicale :

$$\frac{1}{rich\_lex / rich\_néo}$$

[7] 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.

Mathieu Valette « Méthodes pour la veille lexicale »,
version soumise [3 septembre 2009] et acceptée – à paraître
Actes de la journée d'étude Le dictionnaire électronique. Quelles perspectives pour les sciences humaines et sociales ? (Kénitra, le 7 décembre 2007), Leila Messaoudi éd., *Publication du laboratoire Langage et société*,
Université Ibn Tofail Kénitra (Maroc)

---

[8] Nous empruntons la notion de *taxème* à F. Rastier : le taxème est une classe sémantique de petite taille.