la créativité lexicale : des pratiques sociales aux textes

Sandrine Ollinger, Mathieu Valette

Sandrine Ollinger, Mathieu Valette « La créativité lexicale : des pratiques sociales aux textes »,
version soumise [22 mai 2008] et acceptée – à paraître
Actes du 1er Congrès International de néologie des langues romanes (Barcelone, 07 - 10 mai 2008), CINEO'08.

La créativité lexicale : des pratiques sociales aux textes.

Sandrine Ollinger, Mathieu Valette
ATILF (CNRS-Nancy Université)
sollinge@atilf.fr et mvalette@atilf.fr

Résumé

Nous présentons une plateforme de veille lexicale destinée à l'étude des phénomènes néologiques. La méthode générale que mise en œuvre vise à identifier les candidats à la néologie en confrontant des corpus, c'est-à-dire des archives des pratiques linguistiques, et des lexiques, considérés comme les simulations des usages lexicaux correspondants. Dans le cadre de cette étude, nous détaillons deux des modules de la plateforme : un logiciel d'acquisition automatique de candidats à la néologie formelle et catégorielle et une base de données dédiée à l'observation des candidats. Puis, nous exposons notre problématique en nous appuyant sur une étude de cas. Nous construisons à cette occasion les notion de richesse néologique et de créativité lexicale, en lien avec les genres textuels.

1. Introduction

1.1. Le contexte 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 interopérables pour collecter des textes à partir de sources différentes (fichiers, base de données textuelles, Internet) et en extraire les unités lexicales absentes de lexique de références. En bref, le projet vise *a minima* à produire du matériau pour les lexicographes, par exemple pour enrichir des lexiques existants, créer des 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 de la néologie. Cet article vise à élaborer un certain nombre de proposition conceptuelle pour l'évaluation de la créativité néologique corrélée aux genres et aux discours.

1.2. La méthode générale s'insère dans une perspective contrastive. Elle consiste à comparer *les traces de pratiques sociales* (i.e. des corpus homogènes, voir raisonnés[1]) à *des usages lexicaux simulés* (i.e. des lexiques). Notre hypothèse de départ est inspirée des propositions de la sémantique des textes (Rastier 2001) : l'activité langagière est déterminée par des pratiques sociales et on peut en observer les traces dans les discours et les genres textuels. On doit donc pouvoir traiter la néologie au prisme de ces problématiques textuelles. Nous faisons ainsi l'hypothèse que le néologisme subit les contraintes discursives et génériques exercées sur les textes dans lesquels il s'actualise. Ainsi, si tout discours est *a priori* créatif à proportion de la vitalité de la pratique sociale correspondante, les genres (ensembles de normes de production des textes), présentent un potentiel néologique variable.

[1] Un corpus raisonné est un corpus construit dans la perspective d'une *tâche* précise. Traditionnellement, les notions de discours et de genre y sont questionnées.

2. Constitution du corpus « Pouvoir d'achat »

2.1. Le corpus traite de sujets de société et d'économie (autour du thème du « pouvoir d'achat »). Il a été collecté dans une perspective comparative. Il s'agit d'observer les régimes de créativité lexicale de différents genres textuels. Toutefois, il serait abusif d'affirmer qu'il répond aux critères philologiques requis par la linguistique des textes, il demeure relativement hétérogène.

Ce *corpus de référence* se scinde en deux sous-corpus discursifs (discours politique et discours journalistique) auquel s'ajoute, à titre expérimental, un sous-corpus de blogs politiques (textes de *Versac*, *Coulisses de Bruxelles, UE*, *Plume de Presse*)[2]. Ce choix est motivé par la dynamique du genre blog. Les textes du discours politique proviennent du site officiel de l'Elysée. Le corpus journalistique est issu de la presse hebdomadaire grand public (textes de *Marianne*, le *Nouvel Observateur*, et *Le Point*). Les genres textuels du sous-corpus journalistique sont homogènes, il s'agit d'articles de presse. Les genres du sous-corpus politique sont en revanche, hétérogènes (comptes rendus, discours oralisés, entretiens, etc.) et de ce fait, nous ne l'étudierons pas dans une perspective générique. Le sous-corpus blog pose un autre type de problème : il est fondamentalement dialogique et est composé d'un *post* (lequel, en général, et compte tenu du domaine choisi, correspond à un billet d'humeur) et de commentaires en nombre, longueur et qualité variables.

2.2. Un ensemble d'utilitaires nous a permis de constituer ce corpus de façon semi-automatique à partir d'Internet. L'aspiration des pages web a été effectuée à l'aide de l'outil POMPADOC[3] couplé à un moteur de recherche généraliste[4]. Les critères utilisés lors de cette étape sont : (a) langue de recherche (français) ; (b) les mots-clés de la recherche (« pouvoir d'achat ») ; (c) le nombre de pages désirées ; (d) le nom de domaine sur lequel porte la recherche ; (e) le filtrage des pages similaires. A l'aide de feuilles de style XSLT[5], les pages HTML collectées ont ensuite été converties au format XML en conformité avec les recommandations de la TEI et les documents ont été purgés de leur péritexte (bannières, sommaires, hyperliens, etc.) de façon à obtenir un corpus de textes bruts (*plain text*). Pour les discours politique et journalistique, seul le texte principal fut conservé ; pour le sous-corpus « blog », le contenu du *post* et des commentaires est conservé. Un ensemble de métadonnées a également été collecté automatiquement (date, auteur, titre), renseigné (url) ou produit (identifiant numérique des paragraphes, étiquetage morphosyntaxique)[6].

2.3. À l'issue de cette phase de constitution, nous disposons d'un corpus relativement équilibré en nombre d'occurrences.

---

[2] Respectivement http://www.versac.net, http://bruxelles.blogs.liberation.fr/, http://olivierbonnet.canalblog.com/. Wikio désigne ces blogs comme trois des blogs politiques français les plus influents. La position d'un blog dans le classement Wikio dépend « du nombre et de la valeur des liens qui pointent vers lui. Ces liens sont dynamiques, c'est-à-dire qu'il s'agit de « rétroliens » (*backlinks*) ou de liens postés à l'intérieur des articles. » http://www.wikio.fr/blogs/top/politique.

[3] POMPADOC a été prototypé par Jérémie Ceintrey et Yorick Petey et est maintenu par Sandrine Ollinger.

[4] En l'occurrence, Yahoo (http://fr.search.yahoo.com).

[5] Extensible Stylesheet Language Transformations.

[6] Pour l'étiquetage morphosyntaxique des fichiers XML, nous avons utilisé TreeTagger (http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/) et XML2Tag (Etienne Petitjean, ATILF).

Discours	**Politique**	**Journalistique**			**Blog**		
Sources	Elysée	Marianne	Le Point	Nouvel Observateur	Coulisses Bruxelles, UE	Versac	Plume de Presse
Nb de textes	87	108	170	128	11	20	30
Occurrences	325 135	108 753	86 772	109 202	110 633	108 475	106 883
Formes	18 871	14 593	9 030	9 860	14 669	14 449	15 170

Discours	**Politique**	**Journalistique**	**Blog**
Nb de textes	87 textes propagandistes	406 articles	61 *posts* et commentaires
Occurrences	325135	304727	325991
Formes	18871	21825	31015

fig.1 Description du corpus

3. Dispositif de détection et d'observation de la néologie

Le dispositif expérimental que nous utilisons ici pour l'étude de la néologie dans le corpus « Pouvoir d'achat » comporte un instrument de détection, la POMPAMO[7] et un observatoire, le POAMO.

3.1. POMPAMO : outil de détection de candidats à la néologie

Nous adoptons une méthode lexicographique, basée sur l'utilisation de corpus d'exclusion. Comme Maurel (2004) a montré que moins de 4% des mots inconnus observés sont des néologismes, nous établissons des stratégies de filtrage supplémentaires destinées à améliorer la qualité des candidats détectés.

[7] Une version de POMPAMO en ligne est disponible sur le site du Centre National de Ressources Textuelles et Lexicales : http://www.cnrtl.fr/outils/pompamo/

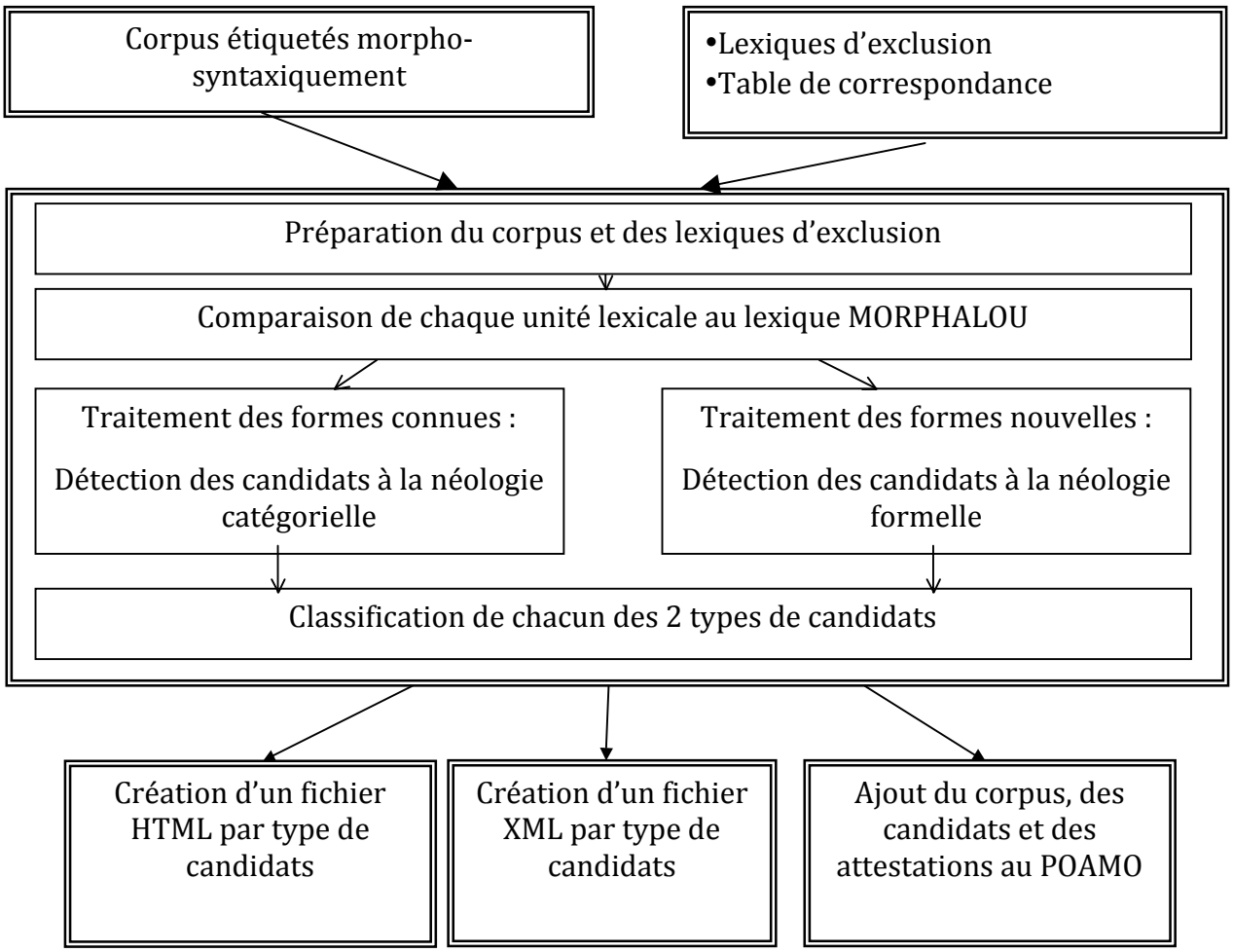

fig.2 Organisation générale de la POMPAMO

La figure 2 présente l'organisation générale de notre instrument. En entrée, on lui fournit le corpus[8] et le lexique proposé par défaut : MORPHALOU 2.0[9] (ressource lexicale à large couverture). L'utilisateur peut également choisir d'ajouter des lexiques supplémentaires d'entités nommées proposés en option (70 438 noms propres, 6 903 adjectifs toponymiques et gentilés et 140 nombres composés)[10]. Il est également possible d'ajouter jusqu'à 6 lexiques personnels, de paramétrer le filtrage des candidats et un certain nombre de métadonnées (nom du corpus, du sous-corpus, type de discours, de champ générique, de genre et de domaine).

Dans son état actuel, la POMPAMO permet la détection de deux types de candidats à la néologie : (a) des candidats à la néologie formelle (formes entièrement inconnues du lexique, formées par dérivation, composition, abréviation ou variation graphique) ; (b) des candidats à la néologie catégorielle (formes connues, mais sous une autre catégorie syntaxique).

---

[8] La POMPAMO accepte les sorties des étiqueteurs Cordial Analyseur (http://www.synapse.com) ou TreeTagger au format TXT.
[9] Lexique des formes fléchies du français, http://www.cnrtl.fr/lexiques/MORPHALOU/
[10] Ces lexiques d'entités nommées ont été réalisés à partir de données proposées par J. Morel (Tours), Th. Poibeau (LIPN-Paris 13) – que nous remercions – et S. Ollinger (ATILF-Nancy).

Trois formats de sortie sont générés par la POMPAMO. L'utilisateur peut se faire une première idée des différents candidats détectés dans son corpus à l'aide de fichiers HTML créés et affichés à l'issue du traitement pour chaque type de candidat. Chacun de ses fichiers contient des informations minimales. Il se compose de deux tableaux, un tableau de candidats et un tableau d'attestations. Des identifiants sont insérés et liés par liens hypertextes afin de faciliter la navigation entre ces sous-parties. L'ensemble des informations fournies pour chaque candidat à la néologie repéré est assez proche des informations contenues dans un dictionnaire.

Par ailleurs, nous avons fait le choix de nous appuyer sur les recommandations TEI et LMF[11] en la matière pour établir la structure d'un proto-dictionnaire de format XML. Enfin, les étiquettes morphosyntaxiques pouvant être vues comme un concentré d'informations hermétiques, nous avons choisi de les traduire en un ensemble d'informations claires à l'aide de *tables de correspondance*[12].

Afin d'enrichir notre représentation du lexique de la langue française et de son évolution à travers les différentes études effectuées, nous avons mis en place une base de données relationnelle MySQL[13], appelée POAMO. Nous disposons d'une version interne de la POMPAMO, nourrissant cet observatoire de créativité lexicale.

3.2. POAMO : observatoire de créativité lexicale

Pour centraliser l'ensemble des candidats à la néologie détectés et faciliter leur observation, nous avons mis en place une base de données relationnelle interrogeable à travers une interface web. Cette base de données s'organise en 5 tables :

(a) Une table de candidats / type (formelle | catégorielle). Un candidat s'y définit en 4 éléments : un identifiant numérique, une forme graphique, une hypothèse d'étiquette morphosyntaxique et une hypothèse de lemmatisation.

(b) Une table d'attestations / type (formelle | catégorielle). Une attestation s'y définit en 4 éléments : l'identifiant du candidat qu'elle atteste, l'identifiant du document dans lequel elle se trouve, l'extrait, étiqueté, du document et la localisation du candidat dans le document XML, en nombre de mots, phrases et paragraphes.

(c) Une table de description des documents, Un document s'y définit en 6 groupes d'éléments : un identifiant numérique, une appartenance à un corpus et à un sous-corpus nommés, un titre, un auteur et une date, une typologie en discours, domaine(s) et genre(s), un nombre de mots, phrases et paragraphe, ainsi que des taux de néologie global, formelle et catégorielle.

L'interface d'interrogation, développée en PHP[14] permet des interrogations croisées en combinant les différents critères de définition du corpus de travail et des candidats, auxquels s'ajoute la possibilité d'interroger sur début ou fin de forme graphique. Elle offre également

---

[11] Text Encoding Initiative et Lexical Markup Framework (ISO TC37).
[12] Ces tables de correspondance traduisent les étiquettes de format prioritaire des différents étiqueteurs en un ensemble d'informations morphosyntaxiques en anglais, dans un format similaire à celui utilisé dans MORPHALOU
[13] http://dev.mysql.com/doc/refman/5.0/fr/index.html
[14] PHP: Hypertext Preprocessor

la possibilité de visualiser une description de l'ensemble des documents présents dans le POAMO et la répartition d'un candidat particulier dans un ensemble de documents.

4. Analyses et propositions

4.1. Richesse lexicale et « richesse néologique théorique »

Si la POMPAMO permet d'étudier tant la néologie catégorielle que formelle, cette étude se focalisera sur la seule néologie formelle. Nous avons effectué un certain nombres de mesures simples afin d'évaluer, en première approximation, la validité de nos hypothèses préliminaires.

La figure 3 présente deux valeurs mesurées sur nos différents sous-corpus (discours politique, journalistique et « blog »). La première, la *richesse lexicale*, correspond au rapport entre le nombre de formes et le nombre d'occurrences de formes. Cette mesure est souvent critiquée parce qu'elle dépend de la taille des textes comparés (la richesse lexicale décroît avec la taille du texte). Dans notre cas, le problème ne se pose pas vraiment dans la mesure où les corpus sont de tailles très semblables (*Le Point* présente toutefois un déficit sensible, cf. figure 1). Nous avons cependant expérimenté un certain nombre d'indices pour constater une relative homogénéité quant aux résultats. Les données finalement exposées dans la figure 3 ont été calculé 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 candidat et α la même constante que précédemment. Le résultat présenté est $\left(\frac{1}{U}\right) \times 100$.

On prendra soin de parler ici de richesse lexicale *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 étiquetées comme nom propre par Treetagger.

	formes	**occurrences**	**candidats**	**richesse lexicale**	**richesse néologique théorique**
Blog	31 015	325 991	2 746	9,51	8,85
Presse	21 825	304 727	764	7,16	3,50
Politique	18 871	325 135	523	5,80	2,77

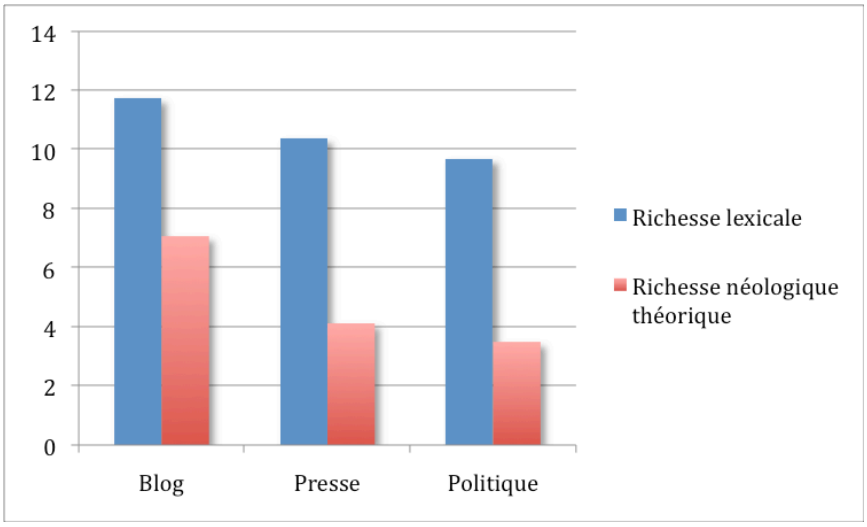

Fig. 3 Richesse lexicale et richesse néologique théorique du corpus de référence

	Formes	Occurrences	Candidats	Richesse lexicale	Richesse néologique théorique
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
Plume de Presse	15 170	106 883	964	10,96	5,22
Coulisses Bruxelles	14 669	110 633	917	10,75	5,14
Versac	14 449	108 475	1244	10,73	6,01

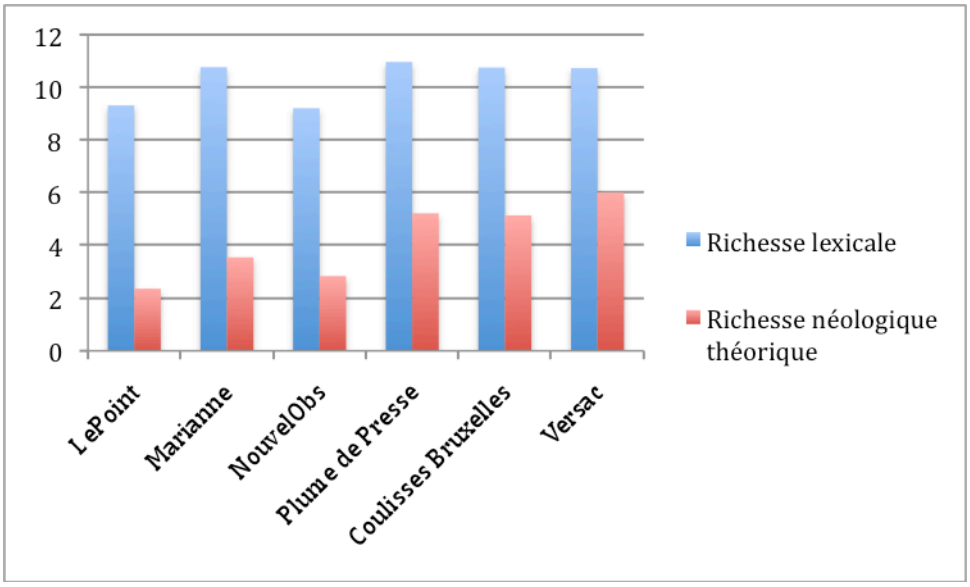

Fig. 4 Richesse lexicale et richesse néologique théorique du corpus journalistique et du corpus « Blog »

La richesse néologique des blogs apparaît sensiblement plus importante que celle des articles de presse (discours journalistique). Elle oscille entre 5,14 et 6,01 suivant notre indice calculé,

quand celle du sous-corpus journalistique varie entre 2,35 et 3,54. A la différence des articles de presse, dont l'orthographe et, dans une moindre mesure, le style, sont contrôlés, les blogs sont des publications sans *sanction normative* – le bloggeur peut certes supprimer des commentaires mais il semble que les corrections orthographiques soient rares. Par exemple, on a relevé manuellement que sur le blog *Versac*, le plus « riche », plus de 60% des candidats à la néologie sont soit des fautes d'orthographes (redoublement d'une console, etc.), soit des omissions d'accents. Si on corrige à partir de cette estimation les trois blogs, on obtient un indice moyen de 3,27, ce qui ramène la richesse néologique théorique des blogs dans le haut de la fourchette des valeurs observées pour le sous-corpus journalistique. Nous reviendrons, au moment de conclure, sur les problèmes posés par ce biais important.

4.2. Mesure de la créativité lexicale

On aura noté, sur la figure 3, que les richesses lexicale et néologique du sous-corpus politique (*Elysée)* sont relativement basses comparée à celle du sous-corpus journalistique. Ce taux est peut-être à rapprocher de la rhétorique du discours politique (fondée sur la répétition, le ressassement), en particulier dans le cas de documents collectés en fonction d'un mot-clé qui, dans ce contexte, peut passer pour un slogan ou un leitmotiv. Cela nous amène à proposer les notions de *conservatisme lexical et de créativité lexicale*. Le schéma de la figure 5 présente le taux de *créativité lexicale théorique* fondé sur le rapport entre la richesse lexicale et la richesse néologique théorique[15].

	Conservatisme lexical
LePoint	3,95
Marianne	3,03
NouvelObs	3,24

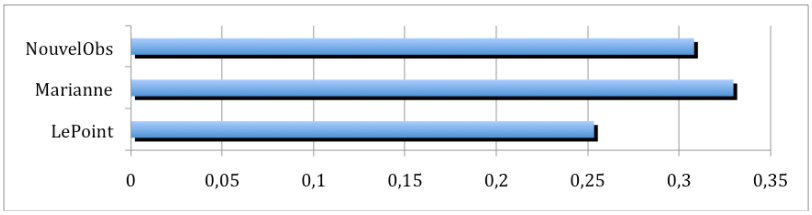

Fig. 5 Conservatisme lexical (à gauche) et créativité lexicale (à droite) du corpus journalistique

Plus l'indice de créativité lexicale est élevé, moins les textes sont conservateurs, c'est-à-dire moins ils ont recours à des néologismes (plus précisément, à des candidats à la néologie) proportionnellement à leur richesse lexicale. En nous focalisant sur le sous-corpus journalistique, on observe par exemple que l'hebdomadaire *Le Point* est sensiblement plus conservateur que les deux autres. *Marianne* présente à la fois le taux de créativité lexicale le plus élevé pour une richesse lexicale et une richesse néologique théorique supérieures aux autres hebdomadaires retenus.

4.3. La créativité lexicale dans Marianne

Nous avons étudié les modalités de construction des candidats à la néologie des textes du sous-corpus *Marianne*. Une recherche par sous-chaîne de caractères, rendue possible par l'interface d'interrogation du POAMO, nous a permis d'identifier un des modes de créativité privilégiés par l'hebdomadaire *Marianne* : il s'agit de la néologies dérivationnelle et suffixale.

---

[15] Soit, la richesse lexicale $\frac{1}{rich\_lex / rich\_néo}$.

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 6 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**	

Il ne sera pas question de discuter ici des raisons de ces choix stylistiques d'importance. On notera toutefois que des travaux récents ont montré que la néologie était un mode de stylisation courant dans le genre pamphlétaire (Jousse 2007). De fait, l'hebdomadaire Marianne est connu pour son ton polémique.

6. Perspectives

6.1. Comme toute activité langagière, la néologie est déterminée par les pratiques sociales. En matière d'observation linguistique, celles-ci sont identifiables au moyen des genres et des discours qui leur sont attachés. Nouvelle lexicalisation d'un thème sémantique stabilisé, le néologisme subit donc les contraintes discursives et génériques exercées sur les textes dans lesquels il s'actualise. Nous avons vu que la créativité lexicale ne répondait pas aux mêmes modes de formation du mot selon les genres. Certains genres sont peut-être des ateliers de créativité lexicale (le pamphlet), d'autres seront réputés conservateurs ou institutifs (la loi-programme, le discours de politique générale) – ce qui explique peut-être l'indice de conservatisme lexical du corpus politique (0,36, sensiblement supérieur à celui du corpus journalistique) que nous avons calculé.

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 évaluer et raffiner ; ils constituent toutefois des outils fonctionnels pour le développement d'une problématique générale de veille lexicale.

6.2. Par ailleurs, des questions relatives à la constitution semi-automatisée des corpus, se posent : les textes des blogs sont-ils de bons indicateurs pour la veille lexicale ? Certes, les néologismes naissent 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, à l'heure actuelle, valorisé par les initiatives du « Web participatif ». De fait, les index des moteurs de recherche généraliste 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 exclure les sources sans autorité éditoriale. Une tendance actuelle est de considérer les textes non pas comme objet de science mais à les réduire, par défaut, au statut préscientifique de ressource – un matériau brut dont la qualité est déterminée par la seul présence, après raffinage, de l'objet étudié. Or, l'occurrence d'un néologisme dans un blog n'a pas le même poids ni la même validité qu'une occurrence dans un article de presse. Dès lors, il est possible que, pour des raisons aussi bien techniques qu'éditoriales, le blog soit à exclure (peut-être provisoirement) des recherches en veille lexicale, si celle-ci ont vocation lexicographique.

7. Bibliographie

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

Maurel D. (2004) « Les mots inconnus sont-ils des noms propres ? » in *JADT 2004 : 7ème journées internationales d'Analyse statistique de Données Textuelles*, 776-784

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

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