NEOVEILLE, SYSTÈME DE REPÉRAGE ET DE SUIVI DES NÉOLOGISMES EN SEPT LANGUES

Emmanuel Cartier, LIPN CNRS UMR 7030, Université Paris 13 Sorbonne Paris Cité

emmanuel.cartier@lipn.univ-paris13.fr

Les dictionnaires de langue décrivent les lexies et leurs sens. Mais ce vocabulaire et les usages évoluent avec le temps. Cela est évident à l'échelle historique, mais également en diachronie « courte » : on estime à environ 10% le nombre moyen de formes inconnues dans les textes de langue générale (Renouf, 2014). Une grande partie de ces mots-formes sont des noms propres, des hapax ou des erreurs typographiques, mais également des néologismes, entendus comme formes-sens nouvelles ou encore néologismes de forme. Le taux de néologismes sémantiques, c'est-à-dire de formes existantes auxquelles on attache un sens nouveau, est quant à lui aujourd'hui plus difficilement quantifiable.
La néologie - l'étude de ces évolutions-, doit nous éclairer sur la vie des formes linguistiques et des sens. Cette étude est indispensable à la linguistique générale, mais aussi à la linguistique appliquée (mise à jour de dictionnaires de langue et de ressources lexicales pour les applications TAL, par exemple).

Mots-clés : veille néologique, extraction automatique, mesure automatique de la diffusion, linguistique de corpus

Nous présentons dans cet article la plateforme Neoveille, qui est liée à un projet collaboratif international financé par la COMUE Sorbonne Paris Cité. Nous effectuons tout d'abord une revue des modèles de la néologie en linguistique et en traitement automatique des langues. Ensuite, nous décrivons exhaustivement les différents composants de la plateforme Neoveille. Nous terminons par une présentation de l'état actuel, avec les premiers résultats sur le français.

1. Éléments de modélisation des néologismes, repérage et suivi des néologismes en corpus

1.1. Modélisation des néologismes

Les mots nouveaux ont de tout temps intéressé les linguistes, mais il faut attendre le 18ème siècle pour voir apparaître, en français, les mots formés sur *néo-* et *log-* (dont *néologie* et *néologisme*) et le 19ème pour voir apparaître la sémantique comme étude de l'évolution du sens des mots[1]. Celle-ci a été récemment renouvelée, d'une part par les travaux de Rastier et Valette (2009), d'autre part par les travaux de la linguistique américaine, dans le cadre de la linguistique de corpus et des grammaires dites de construction. Nous évoquons rapidement certaines de ces approches dans les sections qui suivent, en focalisant sur les éléments qui nous seront utiles pour modéliser le phénomène néologique.

1.1.1. Tradition française

Nous partirons du dernier état du tableau des matrices lexicogéniques de Sablayrolles (2015). Ce dernier propose une typologie des procédés de formation des néologismes selon différents critères (figure 1).

[1]Il n'entre pas dans notre propos de retracer cette histoire ici et nous renvoyons à (Sablayrolles, 2000) et (Geeraerts, 2009) pour un panorama des approches.

	morpho-	construc	Affixation	**préfixation**	détatouer
m	sémantiques	tion		**suffixation**	statuesque
a				**dérivation inverse**	prester
t				**parasynthétique ?**	désidéologisé ?
r				**flexion**	ils closirent, la représaille
i			Compo-	**composition**	voiture-bélier
c			sition	**synapsie**	lanceur d'alerte
e				**composition savante**	batracianophile
				hybride	e-commerce, aquacinéaste
s			Compo-	**fracto-composition**	téléspectateur
			sition par	**compocation**	mobinaute, dircab
			amalgame	**factorisation**	optipessimiste
				mot valise	peopolitique
		imitation et		**onomatopée**	dzoing
i		déformation		**f coupe ou paronymie**	la nesthésie, infractus,
n	syntactico-	changement		**conversion**	la glisse, la gagne
		de		**conversion verticale**	un ex, le co
		fonction		**déflexivation**	le boire, le manger
t	sémantiques			**combinatoire syntax° /**	ironiser un texte
e				**lexicale**	encourir la liberté
r		changement de sens		**métaphore**	souris (inform.)
n				**métonymie**	sac à dos 'touriste'
e				**autres figures**	escorteuse 'call girl'
s	morpho-	réduction		**troncation**	blème, petit déj
	logiques	de la forme		**siglaison /acronyme**	LMD, ECUE
	phraséolo-	pragmatico-sémantique		**détournement**	faire marcher la planche à promesses
	gique	création		**création**	ne pas faire du huit megabits
matrice externe				**emprunt**	break, cool fioul, redingote

tableau 1 : matrices lexicogéniques (Sablayrolles, 2015)

Dans la lignée de Tournier (1985), il distingue ici tout d'abord les *matrices internes* à la langue d'une *matrice externe*, qui concerne exclusivement les emprunts. Dans le premier groupe, il distingue les *néologismes purement morphologiques* (dans lesquels on retrouve les troncations et les siglaisons, qui ne s'accompagnent d'aucune modification de sens), les *néologismes morpho-sémantiques* (par construction : affixation et composition, les deux mécanismes traditionnellement décrits dans les grammaires ; par imitation et déformation, opérant principalement sur des éléments phonologiques), ainsi que les néologismes phraséologiques, avec deux sous-types, la création et le détournement. Ce

sont parmi ces catégories que la forme change, tandis que dans les n*éologies syntactico-sémantiques* (changement de fonction ou de sens), seul le sens (ou l'emploi) change.

Il faut considérer cette typologie comme une catégorisation des mécanismes de base, car un néologisme peut être le résultat de plusieurs opérations successives. Il est évidemment possible d'avoir des néologismes ne faisant appel qu'à un seul procédé (*statuesque* par exemple pour la suffixation sur un radical simple, ou bien *binge-drinking* pour les emprunts), mais également des néologismes construits sur des bases elles-mêmes construites (*pré-ado* est préfixé sur un mot tronqué), ou faisant appel à des formants étrangers (*biotiful* est un amalgame, graphique, mettant en jeu un formant emprunté).

On notera que dans ce tableau ne figurent plus (comme il le faisait encore en 2000) les changements de sens par extension et par restriction de sens. C'est que J.-F. Sablayrolles, à la suite de Meillet ([1906] 1958 : 235), considère que ces évolutions sémantiques ne relèvent pas à proprement parler de la néologie, mais sont dus à la « discontinuité de la transmission du langage » et relèvent de l'histoire de la langue.

Gevaudan et Koch (2010) proposent une théorie unifiée de l'évolution lexicale appelée théorie de la filiation, qui se base sur quelques concepts-clés de la linguistique cognitive (notions de scénario et de prototype). Selon cette théorie, toute évolution lexicale (continuité lexicale et innovation lexicale) peut être décrite au moyen de trois paramètres : deux qui sont translangues, le paramètre sémantique (identité, contiguïté, similarité métaphorique, superordination, subordination, similarité cotaxinomique) et le paramètre stratique (lié à la continuité historique du vocabulaire : identité, emprunt, calque, étymologie populaire), et l'un qui est lié aux catégories propres à chaque langue, le paramètre formel, avec cependant quatre types génériques d'innovations possibles en plus de l'identité (changement de catégorie grammaticale, extension morphologique d'une forme lexicale, combinaison de formes lexicales, réduction d'une forme lexicale, intégrant l'ellipse). Cette théorisation permet de décrire précisément les différentes innovations lexicales, et peut être croisée avec l'approche plus détaillée liée aux matrices lexicogénétiques.

1.1.2. Tradition américaine : linguistique de corpus et grammaires constructionnelles

La linguistique de corpus a renouvelé les approches en proposant de quantifier les descriptions linguistiques. La fameuse formule de Firth (1957) « We can know the meaning of a word by the company it keeps » provient de la tradition distributionnelle harrissienne où l'on trouve le principe même de tous ces travaux :

...if we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution. (Harris, 1954)

Depuis les années 90, la linguistique de corpus a mis au jour des phénomènes linguistiques au travers des notions de collocations, de « multiword expression », de collostructions (Stefanowitsch and Gries, 2003), de *word sketches* (Kilgariff et al., 2004). Cette dernière notion donne une base à l'étude automatique des innovations sémantiques, en permettant de décrire automatiquement le profil combinatoire d'une lexie, par son environnement lexico-syntaxique.

L'approche distributionnelle trouve une assise théorique et cognitive nouvelle avec certains travaux de la sémantique cognitive (Langacker, 1987, 1991; Schmid, 2007, 2013). Ceux-ci ont explicité une notion, « entrenchment », qui fonde le processus d'ancrage socio-cognitif des signes linguistiques (définis comme l'association d'une forme et d'un sens), en corrélant ce processus avec la répétition des occurrences. Ils signalent par ailleurs que cet enracinement est, du fait du processus sur lequel il se base, instable dans le temps, et profondément lié à la répétition et à la variabilité des usages. Ces approches fondent l'étude statistique-distributionnelle en corpus : les répétitions de séquences linguistiques reflètent l'enracinement (ou corrélativement le déracinement) socio-cognitif des signes linguistiques. Le phénomène de répétition doit par ailleurs être replacé dans son contexte global : c'est un processus continu et évolutif, instable, qui nécessite de prendre en compte la variabilité des corpus selon l'axe diachronique, et selon l'axe socio-géographique.

Un autre courant linguistique se place clairement dans le paradigme distributionnel – dans son hypothèse initiale que la répétition de séquences en corpus nous informe sur la langue - et dans celui de la linguistique de corpus : les grammaires de construction (Fillmore et al., 1988 ; Goldberg, 1995, 2003 ; Croft, 2001, 2007). Comme la sémantique cognitive, ils considèrent que les corpus sont la

matière principale de toute étude linguistique et que le calcul des répétitions en corpus est une matière dérivée essentielle. Ils proposent un modèle linguistique qui rejette la distinction lexique (qui identifierait des mots-lexies et en décriraient les propriétés) – grammaire (qui décrirait les règles de combinaisons de lexies). Pour eux, les signes linguistiques comprennent toute unité graphique ou sonore liée à un sens, du morphème aux schémas syntaxiques[2], et ils appellent ces unités des *constructions*. L'objectif de la linguistique est alors de décrire ces constructions dans un nouvel objet qu'ils appellent *constructicon*.

Dans le cadre des grammaires de construction, plusieurs auteurs (Bybee, 2016 ; Schmidt 2007, 2008, 2015, 2016 ; Traugott et Trousdale, 2013) ont proposé des modèles du changement linguistique, basé sur ces notions d'entrenchment et de calculs des répétitions en corpus. Ils proposent également un modèle en trois phases (innovation, propagation, conventionnalisation) qui permet de passer de l'apparition des néologismes à leur assimilation éventuelle dans la langue. Ils décrivent quelques-unes des caractéristiques linguistiques de ces différentes phases qui permettent d'évaluer l'état d'un nouvel emploi dans une période p. Cette approche constructionnelle introduit le continuum dans tous les phénomènes linguistiques, et reprend la notion de prototype de la linguistique cognitive ; elle s'appuie généralement sur les hypothèses distributionnelles et de ce point de vue semble avoir une pertinence linguistique et ingénierique.

1.2. Veille néologique : approches manuelles et approches automatiques

La veille néologique, avant l'avènement de l'outil informatique et des possibilités qu'il offre (accessibilité à de gros corpus, calculs automatiques, etc.), a utilisé – et utilise encore-, la méthode de la *recherche de l'aiguille dans la botte de foin*. Cette méthode traditionnelle, qui se base sur l'expertise et l'intuition linguistique, ne doit pas être sous-estimée, et il s'agit encore aujourd'hui de la méthode préférentielle de tout travail sur la néologie, au moins du point de vue de l'analyse des nouvelles formes et des nouveaux sens. Il est par ailleurs évident que les systèmes automatiques ne peuvent pas repérer toutes les occurrences de néologismes, étant donné que la créativité linguistique déborde très souvent le cadre de règles complètement automatisables. Même si aucun ouvrage méthodologique n'existe aujourd'hui en la matière, l'expertise linguistique et sa modélisation ont inspiré nombre de développements en traitement automatique des néologismes et nouveaux emplois, comme nous le verrons dans les sections suivantes.

1.2.1. Approches automatiques : le repérage des néologismes de forme

Le repérage automatique peut prendre pour objet soit les formes nouvelles, soit les nouveaux usages/sens en corpus.

La première approche consiste à utiliser une **ressource lexicographique de référence** pour repérer dans un corpus tous les mots inconnus, puis met en œuvre différents filtres afin d'identifier des candidats néologismes (Cabré *et al.*, 1995, 2003 ; Olinger et Valette, 2010; Sagot *et al.*, 2013 ; Gérard *et al.*, 2014). Cette méthode nécessite un dictionnaire de référence suffisamment couvrant, des ressources liées aux entités nommées, et des algorithmes efficaces de repérage des erreurs typographiques. Elle est donc beaucoup plus complexe à mettre en œuvre qu'il n'y paraît, pour les raisons suivantes :

- il est très difficile de disposer d'une ressource lexicographique à jour, quelle que soit la langue considérée ; pour le français par exemple, en 2015, plusieurs ressources lexicographiques ont été développées (voir Cartier *et al.*, 2015 pour une revue complète) ; citons Morfetik (Mathieu-Colas et al., 2010), Lefff (Sagot, 2013) et GLAWY (Sajous et Hathout, 2015). Ces ressources ont des mérites respectifs divers : tandis que Morfetik est le dictionnaire actuellement le plus couvrant du point de vue des formes simples, il ne dispose d'aucune

[2]« Constructions are defined to be conventional, learned form-function pairings at varying levels of complexity and abstraction (...). This definition is meant to highlight the commonality between words and larger phrasal units. » (Golberg, 2013, p.12) Et elle donne des exemples de constructions : morphèmes, mots (Iran, another, banana, V-ing) expressions figées (give the Devil his due, going great guns) et semi-figées (Jog <someone's> memory), phraséologismes (The Xer the Yer : the more you think about it, the less you understand), constructions grammaticales (Subj V Obj1 Obj2 : he gave her a fish taco, he baked her a muffin).

forme composée ; Le Lefff dispose d'une couverture moindre, mais intègre un grand nombre de noms propres, qui le rend moins utilisable en traitement automatique ; GLAWY, enfin, est sans doute la ressource la plus intéressante : il s'agit d'une ressource issue du wiktionnaire converti dans un format XML exploitable, qui comprend la plus large couverture lexicographique. Le problème ici réside dans une *trop grande* couverture, puisque ce dictionnaire comprend un certain nombre de néologismes, non marqués comme tels, ce qui rend difficilement exploitable en l'état la ressource.

- pour les langues peu dotées en dictionnaires, une telle ressource lexicographique prise comme corpus d'exclusion n'est pas disponible. Par exemple, en russe ou en tchèque, aucun dictionnaire électronique exploitable n'est à disposition de la recherche, et il est donc nécessaire de recourir à d'autres méthodes pour construire une telle ressource. Une solution consiste à récupérer les formes attestées dans un corpus suffisamment large, et à ne prendre en compte que les formes ayant une fréquence minimale supérieure à celle de l'hapax. La difficulté réside là dans le choix du corpus, puisqu'il faut alors en disposer, d'une part, et établir la couverture linguistique visée. Cependant, cette méthode de construction dynamique de la ressource linguistique paraît aujourd'hui une voix prometteuse.
- Une autre méthode consiste à utiliser des analyseurs morphosyntaxiques utilisant les méthodes d'apprentissage automatique supervisé, qui se base donc sur un corpus d'apprentissage annoté manuellement. C'est ainsi qu'il est possible d'utiliser Treetagger (Schmid, 1995) pour reconnaître automatiquement des mots inconnus dans les textes, pour différentes langues. Cette méthode présente l'avantage de ne pas avoir à construire de ressource linguistique de référence. L'une des difficultés présentée par cette approche est la reconnaissance des mots composés, puisque très peu d'analyseurs aujourd'hui permettent cette reconnaissance, alors qu'une partie des mots simples constitutifs de mots composés peuvent être inconnus (*à l'instar* par exemple).

La méthode par dictionnaire de référence reste la méthode privilégiée de tous les systèmes existants, car elle permet d'identifier dans un corpus une liste élargie de candidats néologismes. Elle est mise en place dans tous les systèmes aujourd'hui opérationnels (NeoCrawler, Logoscope, Obneo).

Elle nécessite ensuite des post-traitements afin d'éliminer de la liste des candidats :

- les noms propres : ceux-ci peuvent être repérés : sur des bases purement typographiques (mots capitalisés), mais il faut alors pouvoir distinguer ces capitales de la capitalisation du premier mot de phrase, ainsi que prendre en compte la possibilité de parties de noms propres non capitalisées (exemple : Pierre du Verger, Saône et Loire, etc.) ; sur la base d'un dictionnaire de référence, mais cette solution ne peut fonctionner que pour les noms propres les plus usuels, la catégorie des noms propres étant la plus productive dans les langues ; au moyen de règles de formations internes (exemple : Prénom + Mot capitalisé éventuellement répété), ou de règles contextuelles (*la <u>mairie</u> de Bordeaux, le <u>port</u> de Aigues-Mortes, le <u>directeur</u> de Renault,* etc.) . Le repérage des entités nommées a donné lieu à de nombreux travaux en TAL, ce qui rend cette tâche accessible.
- les erreurs typographiques : les coquilles, les erreurs d'orthographe sont une source de repérage de mauvais néologismes ; dans ce cadre, il convient d'utiliser les correcteurs orthographiques disponibles, même si les algorithmes de correction orthographique peuvent très souvent aussi tenter de corriger de vrais néologismes (exemple : *biotiful => beautiful*).
- les passages en langue étrangère, principalement dans des citations, qu'il faut donc repérer et éliminer.
- les erreurs issues des étapes précédentes du traitement automatique : lorsque le corpus provient notamment du web, l'extraction des zones de texte peut être fautive, comme l'étape préalable de segmentation des textes en mots.

Elle peut par ailleurs s'accompagner de traitements spécifiques en aval afin de catégoriser les candidats néologismes (emprunts, néologismes formels : troncation, etc.). À notre connaissance, aucun système actuel n'effectue aujourd'hui ce type de traitements de manière systématique. Deux pistes sont envisageables, et ont abouti à des prototypes.

- **Identification « interne »**, modélisant les spécificités linguistiques formelles de ces lexies et cherchant à décomposer les procédés de leur formation. Pour le repérage des emprunts (Kang et Choi, 2002 ; Alex, 2008 ; Jacquet-Pfau, 2003) détaillent les consécutions de lettres permettant de repérer dans les textes dans une langue L des « mots » en langue étrangère (spécifiquement anglicismes). Un travail similaire est mené au LDI pour repérer les mots-valises et les troncations (Vinogradova, en cours).
- **Identification « contextuelle »**, modélisant des contextes spécifiques aux néologismes, notamment au moment de leur apparition. En effet, l'apparition d'une lexie nouvelle s'accompagne dans la très grande majorité des cas d'une apparition en mention, ainsi que d'une glose définitoire-explicative (Cartier, 2011). Cette approche est un moyen complémentaire de repérage des néologismes formels et sémantiques, offrant de plus l'avantage de proposer une définition initiale de la nouvelle forme ou du nouveau sens.

La méthode par dictionnaire de référence ne permet de repérer qu'une partie des néologismes, ceux qui introduisent une forme nouvelle. Pour identifier les néologismes sémantiques et les nouveaux emplois, il est nécessaire de recourir à d'autres méthodes.

1.2.2. Approches automatiques : le repérage de la néologie sémantique

Certains chercheurs (Renouf, 1993 et 2014 ; Garcia-Fernandez *et al*., 2011 ; Cabré et Nazar, 2011, 2012 ; Cartier, 2011 ; Kerremans *et al*., 2012 ; Schmid, 2013 ; Kilgariff, 2004 ; Blumenthal, 2009) ont proposé d'utiliser une autre méthode, permettant de repérer les néologismes sémantiques et nouveaux emplois, c'est-à-dire utilisant des mots préexistants dans le vocabulaire et partant ne pouvant être repérés par la méthode précédente. Cette méthode part de la notion de « profil combinatoire » (ou *word sketch*), qui décrit les environnements typiques de la lexie étudiée sur la base de ses collocatifs, soit à partir du texte brut, soit après une analyse morphosyntaxique, soit même après une analyse en dépendances syntaxiques. C'est ainsi que *Sketch Engine* propose le profil suivant pour le nom anglais *goal* :

goal *(noun)*
ukWaC freq = 168,184 (107.82 per million)

object_of	59,154	3.20	subject_of	25,630	2.00	adj_subject_of	2,159	1.40	modifier	75,150	1.40	modifies	12,980	0.20
score	8,555	11.03	score	1,029	8.39	galore	27	7.30	ultimate	1,929	9.15	scorer	353	8.94
achieve	9,504	9.69	disallow	270	8.28	achievable	43	6.95	winning	569	7.66	kick	619	8.57
concede	1,418	9.31	concede	223	7.52	attainable	15	6.77	long-term	878	7.58	tally	127	7.58
accomplish	588	7.88	gape	76	6.48	unrealistic	14	5.59	league	638	7.19	keeper	179	6.79
reach	1,937	7.42	kick	93	5.45	intact	19	5.04	primary	1,003	7.19	drought	77	6.45
net	351	7.42	orientate	37	5.12	worthy	36	4.82	second	2,067	7.11	scramble	49	6.44
grab	413	7.31	rule	64	4.91	ambitious	21	4.63	strategic	642	7.02	cushion	52	5.99
attain	406	7.30	come	1,319	4.75	realistic	22	4.17	common	1,539	7.01	lead	264	5.86
pursue	646	7.27	cap	20	4.28	enough	92	3.98	realistic	420	6.93	setting	382	5.85
bag	193	6.66	spark	19	3.99	inevitable	14	3.82	shared	361	6.82	difference	673	5.84
pull	505	6.60	elude	13	3.89	simple	83	3.24	achievable	290	6.81	Cort	26	5.81
set	2,413	6.44	undo	14	3.87	forthcoming	9	2.85	stated	274	6.76	thriller	46	5.67

Figure 1 : profil combinatoire (word sketch) du mot anglais *goal* (source : https://www.sketchengine.co.uk/word-sketch/)

Cette photographie de la distribution de la lexie permet d'établir son usage à une période p (par exemple ici les constructions verbales dont *goal* est l'objet ou le sujet).

En généralisant cette idée, et en revenant aux sources de l'intuition distributionnelle (notamment Harris, 1986), on peut imaginer arriver à affiner ces profils en établissant les constructions typiques d'une lexie donnée, et leur regroupement, qui devrait correspondre à des sens spécifiques. Par exemple, pour un verbe, arriver à déterminer les usages typiques au niveau des groupes sujets et des constructions objets, avec regroupement des unités lexicales valides dans un sens donné, permettant d'approcher une description lexico-syntaxique des lexies et leur appariement à des sens.

À partir de là, identifier un glissement ou une rupture de sens se ramène à identifier un collocatif nouveau, qui ne peut être ramené à une construction préexistante. Aujourd'hui, la disponibilité de gros corpus et la maturité des calculs statistiques pour identifier les collocatifs typiques doivent théoriquement permettre de le faire. Cependant, à notre connaissance, aucune description systématique ou même conséquente n'a encore été proposée.

Cette méthode complète la précédente : la première fournit des candidats néologismes formels, qui peuvent ensuite être suivis diachroniquement par la seconde, et la seconde méthode permet de plus de repérer les néologismes sémantiques et les nouveaux emplois potentiels - qui ont une modification de leur profil combinatoire.

1.2.3. Approches automatiques : le suivi des néologismes

Le suivi des néologismes repérés permet ensuite de connaître le cycle de vie des mots ou des sens nouveaux, de leur apparition à leur intégration dans le langage courant, ou à leur disparition temporaire ou définitive.

Quatre méthodes complémentaires sont utilisables pour suivre ces évolutions :

- **l'évolution fréquentielle** permet d'établir des modèles d'évolution (voir Renouf, 2014) voire des phases dans cette évolution (Valette, 2011), à partir de leur première attestation. Deux systèmes sont opérationnels : l'application Google Trends[3], qui propose des courbes de fréquence de 2004 à nos jours pour les requêtes saisies dans le moteur de recherche Google ; les travaux d'A. Renouf dans le cadre du projet AVIATOR (Renouf, 2014) établit des « modèles » du cycle de vie des néologismes à partir des courbes de fréquence. (Valette, 2011 ; Olinger et Valette, 2010) font de même.
- **l'évolution des profils combinatoires** complète le suivi fréquentiel. Les travaux jusqu'ici sont seulement programmatiques (sauf en anglais, Renouf, 2014). Nous approfondirons cette approche en mobilisant les différentes variantes de l'approche distributionnelle présentées dans (Turney et Pantel, 2010) et en reprenant (Schmid, 2013) qui présente différentes mesures pour approcher cette notion de profil combinatoire, en insistant sur les biais des différents calculs. Les approches du *data mining* (Charnois, 2009 ; 2011) pourront également être mobilisées. Un récent numéro des *Cahiers de lexicologie* (Kabatek et Girard, dir. 2012) est également consacré à ce sujet.
- **Une approche thématique liée aux types de corpus** : il est bien connu que les néologismes sont en plus grand nombre dans les discours oraux que dans les discours écrits, d'une part. Et qu'une partie non négligeable des néologismes naissent dans des groupes sociaux ou socio-économiques particuliers : c'est ainsi que par exemple nombre d'anglicismes actuels proviennent du domaine informatique puis, de par la démocratisation de ce domaine, passent ensuite pour partie dans le langage courant ; de même, nombre d'anglicismes proviennent de groupes sociaux parisiens spécifiques, dont on trouve trace dans un certain nombre de magazines (*Slate*, *GQ*, *Elle*, etc.). Il est donc tout à fait imaginable de faire une veille spécifique sur ces domaines « sensibles » à la néologie (ou à un type de néologismes), en utilisant ces sources d'informations puis en étudiant le passage ou non des néologismes repérés dans la langue générale.
- Pour ce qui concerne les emprunts, la phase d'assimilation par la langue peut être repérée lorsque le terme étranger assimile la morphologie de la langue d'accueil, et même permet des dérivations dans d'autres catégories. Ces traces sont moins évidentes pour les autres types de néologismes.

1.2.4. Plateforme de repérage et de suivi des néologismes : systèmes génériques versus systèmes dédiés

À notre connaissance, il n'existe actuellement aucune plateforme permettant d'effectuer à la fois le repérage automatique (ou semi-automatique) des néologismes en corpus et de suivre l'évolution

[3]https://www.google.fr/trends/explore

des néologismes en corpus diachronique, à l'exception des travaux menés par Renouf mais qui ne sont pas librement accessibles. Nous pouvons cependant évoquer des outils génériques permettant d'étudier les néologismes, ainsi que les systèmes dédiés.

1.2.4.1. Outils génériques pour la recherche et le suivi des néologismes.

Parmi les systèmes génériques, il faut citer deux outils régulièrement utilisés par les linguistes pour effectuer une veille néologique : les moteurs de recherche généraux ou spécialisés, et l'application Google Trends.

Les moteurs de recherche généraux (type Google ou Bing) sont des outils pratiques pour effectuer des recherches d'attestations néologiques à partir du moment où l'expert a déjà repéré des candidats : ils sont simples d'utilisation, permettent d'accéder à de très gros corpus et fournissent donc souvent nombre d'attestations. Ils présentent cependant quelques inconvénients :

- D'abord, ils effectuent une recherche sur du corpus exclusivement web, ce qui introduit un bruit non négligeable empêchant toute conclusion objective sur les contextes d'apparition ou sur le cycle de vie ; il est cependant possible de restreindre les recherches à un sous-corpus (par exemple, dans Google, restreindre aux textes en langue française, ou même aux sites appartenant à tel ou tel pays francophone) ; il est également possible d'utiliser certains moteurs de recherche spécialisés. Citons notamment, pour ce qui concerne le corpus journalistique, l'application Europresse. Mais, dans l'idéal, il faudrait pouvoir disposer d'un moteur de recherche permettant de gérer les corpus pris en compte.
- Ensuite, ils donnent des résultats sous forme d'extraits de texte, montrant généralement la seule première attestation de la lexie cherchée, alors que le linguiste souhaiterait obtenir l'ensemble des occurrences ; de ce point de vue, il existe là encore des moteurs de recherche proposant non pas des extraits de texte, mais des tables de cooccurrences ; nous pensons par exemple à WebCorp (Renouf et al., 2013) ; il existe également deux applications largement utilisées dans la communauté linguistique, IMS Corpus WorkBench (Evert et al., 2011) et SketchEngine (Kilgariff, 2004), qui proposent des moteurs de recherche de cooccurrences, mais qui nécessitent, l'un comme l'autre de construire préalablement les corpus à étudier et ne sont pas d'une prise en main aisée. Cette contrainte est un avantage, puisque cela permet de mieux contrôler les sources d'informations, et de leur associer des méta-informations. Malheureusement, les deux outils utilisent actuellement un moteur d'indexation bien plus limité que ceux proposés par des outils comme Apache Lucene ou Solr.
- Par ailleurs, les possibilités d'interrogation des corpus sont généralement limitées : dans Google, il n'existe pas, par exemple, de langage d'interrogation par expressions régulières ou encore avec le métalangage utilisé par IMS CWB ou SketchEngine, CQP. Celui-ci permet d'effectuer des recherches par expressions régulières et via les propriétés morphosyntaxiques des termes. Ce métalangage est devenu de facto la référence pour les plateformes d'étude linguistique basée sur corpus, et devrait être rendu disponible sur toute plateforme.
- Enfin, les moteurs de recherche généralistes n'effectuent aucun traitement linguistique préalable des corpus, limitant par là les possibilités d'interrogation et la précision des recherches.

Pour ce qui concerne le suivi des néologismes, Google Trends est aujourd'hui un modèle pour étudier l'évolution fréquentielle des termes au cours du temps. Même si cette application ne donne des résultats que pour les mots issus des requêtes tapées dans Google, elle donne une idée de ce qu'il faudrait implémenter pour rendre compte des évolutions d'attestations dans un moteur de recherche spécialisé en veille néologique.

1.2.4.2. Outils spécifiques pour la recherche et le suivi des néologismes.

Quatre applications de veille néologique ont été mises en place pour le français, dont deux sont accessibles en ligne : POMPAMO (http://www.cnrtl.fr/outils/pompamo/ : Olinger et Valette, 2010) et le LOGOSCOPE (http://lilpa.unistra.fr/fdt/projets/projets-en-cours/logoscope/ : Gérard *et al*., 2014). Ces deux outils, ainsi que les deux autres (Cabré, 1995), (Sagot *et al*., 2013) utilisent la méthode de repérage des néologismes par corpus d'exclusion (ou dictionnaire de référence). Pompamo

permet, à partir d'un texte étiqueté morphosyntaxiquement de repérer des candidats néologismes. Il utilise comme dictionnaire d'exclusion Morfalou ainsi que des dictionnaires additionnels (noms propres et gentilés), éventuellement fournis par l'utilisateur. Cet outil est relativement simple, car il n'effectue aucune catégorisation des candidats néologismes, et ne permet de travailler que sur de petits fichiers à télécharger par Internet, ce qui en limite singulièrement l'intérêt. Le Logoscope, pour sa part, propose directement les candidats néologismes identifiés dans des corpus de presse, avec leur contexte d'apparition. Mais, là encore, l'application est très statique, et semble peu évolutive, l'utilisateur n'accédant qu'au résultat final qui comporte une grande part de travail manuel. Selon les auteurs, les résultats du système sont humainement triés avant présentation au public. Il faut également citer les travaux menés par l'équipe de Teresa Cabré de l'IULA de l'université Pompeu Fabra de Barcelone, depuis plus de vingt ans. Ils disposent d'une plateforme, appelée OBNEO (Observatorio de la Neologia), permettant de gérer les corpus étudiés, les dictionnaires de référence utilisés, les néologismes automatiquement repérés. D'une certaine façon, le projet Neoveille, dont la présentation fait l'objet de cet article est une extension des fonctionnalités de cette dernière série d'outils.

Pour conclure, les outils actuels, en tout cas pour le français, ne fonctionnent qu'à partir de dictionnaires de référence, et proposent des fonctionnalités limitées. Pourtant, il existe, notamment en Angleterre et en Allemagne, deux systèmes plus dynamiques, proposant des candidats néologiques à la volée, sur du corpus en évolution. C'est dans cette direction que nous voudrions situer la plateforme Neoveille, que nous allons maintenant présenter.

1.2.5. Plateforme de repérage et de suivi des néologismes : les exigences d'un système idéal

Nous sommes maintenant à même d'établir les caractéristiques indispensables d'un système de repérage et de suivi des néologismes.

Tout d'abord, la **facilité d'utilisation** : il est évident qu'un tel système doit pouvoir être facilement pris en main par un expert linguiste, et dans ce cadre le modèle du moteur de recherche semble le plus adapté.

Ensuite, la possibilité de **faire interagir le système automatique et l'expertise humaine** : il est évident que les données fournies automatiquement ne sont pas fiables à 100%, et il est donc nécessaire d'inclure des possibilités de modification des résultats automatiques par l'expert linguiste, qui devra également pouvoir ajouter de l'information linguistique aux lexies extraites.

Parmi les **fonctionnalités du moteur de recherche** qui semblent primordiales, nous pouvons retenir:

- possibilité de disposer de corpus conséquent et de pouvoir *gérer* ces corpus (ajout, suppression, modification)
- possibilité de pouvoir interroger le corpus de manière simple et puissante, par exemple en disposant d'un moteur d'expressions régulières et d'un métalangage comme celui de CQP ;
- possibilité de disposer à la fois de résultats sous forme d'extraits de texte, mais également sous forme de concordancier
- possibilité de pouvoir filtrer les résultats selon différents critères : types de journaux, période temporelle, etc. De manière générique, possibilité de pouvoir analyser les données brutes de différentes façons.
- possibilité de pouvoir disposer pour une recherche donnée (et ses résultats) d'une vision diachronique sous forme de graphe montrant les évolutions fréquentielles.
- Possibilité de pouvoir exporter les lexies candidates et leurs contextes saillants, afin de pouvoir en faire une analyse linguistique externe.

Finalement, la plateforme doit comporter trois **ressources en interaction, gérées par l'expert linguiste** :

- gestionnaire des corpus
- gestionnaire de dictionnaires de référence
- gestionnaire des néologismes

Plusieurs interactions se produisent entre chacun des composants :

- des outils permettent d'extraire des néologismes dans les corpus choisis ;
- un moteur de recherche et d'analyse permet ensuite à l'expert linguiste d'étudier à la fois l'apparition de nouveaux néologismes et de suivre leur évolution ;
- les néologismes automatiquement repérés et validés par l'expert linguiste peuvent ensuite être stockés dans le gestionnaire de néologismes pour une description linguistique complète, puis éventuellement, après constatation d'une assimilation par la langue générale, dans les dictionnaires de référence.

Avec ces exigences, nous allons pouvoir exposer les propriétés du système Neoveille.

2. Présentation de la plateforme Neoveille

Ce projet collaboratif, financé pour trois ans (juin 2015 - juin 2018) par la COMUE Sorbonne Paris Cité, regroupe plusieurs laboratoires de Sorbonne Paris Cité (LIPN, LDI, CLILLAC-ARP, ERTIM), les acteurs du groupe EMPNEO et l'université de Sao Paolo (USP).

Le projet vise à :

- mettre en place une **plateforme multilingue de veille et de suivi des néologismes** à partir de corpus contemporains de très grande taille dans sept langues (français, grec, polonais, tchèque, portugais du Brésil, chinois et russe) ;
- utiliser cette plateforme pour mener une étude des **emprunts** (notamment mais pas exclusivement anglicismes) dans différentes langues (français, grec, polonais, tchèque -langues du groupe EmpNéo-, portugais du Brésil, chinois et russe) ;
- utiliser cette plateforme pour étudier la **notion d'innovation sémantique** et pour proposer de nouvelles procédures d'identification de nouveaux emplois.

2.1. Architecture générale

L'architecture générale du système est présentée dans la figure 2.

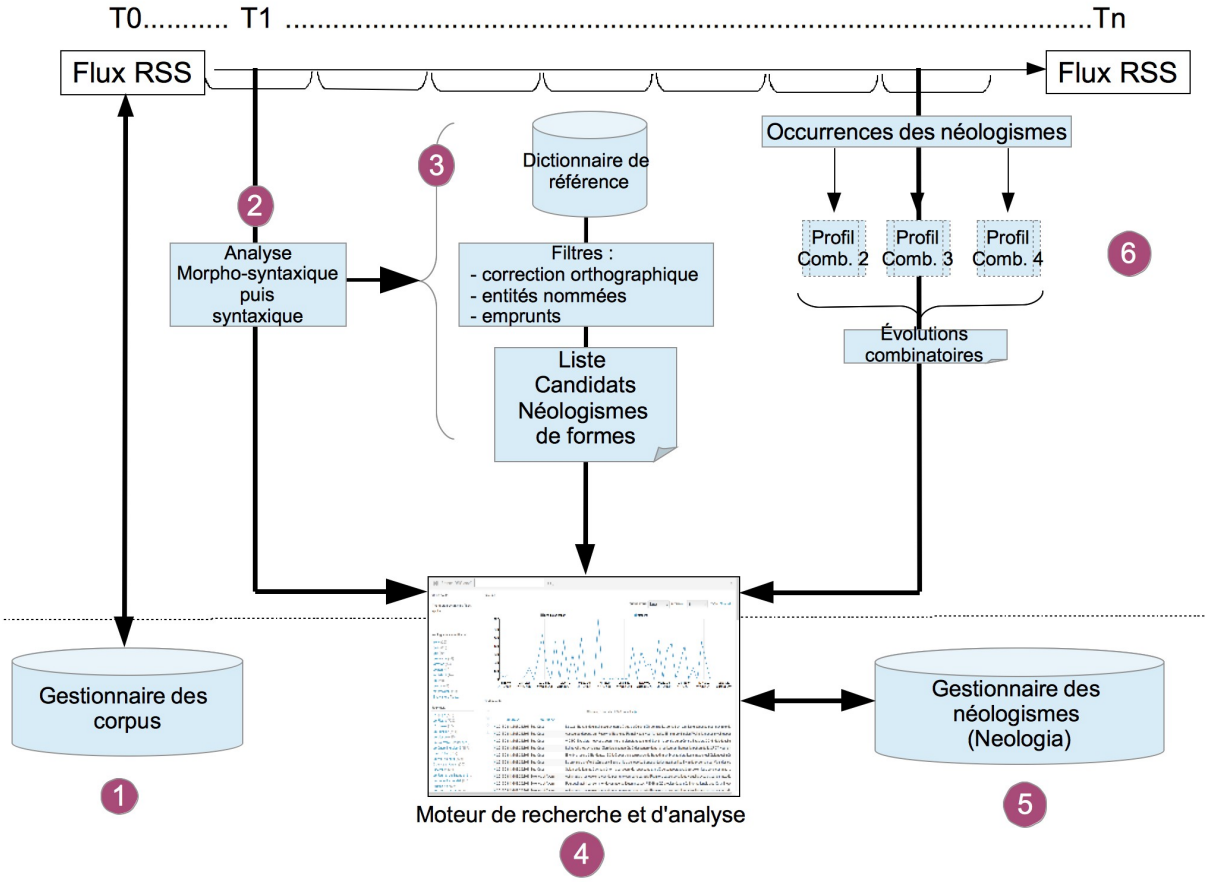

Figure 2 : architecture générale du projet Neoveille

Dans cette architecture, le trait horizontal sépare les composants où l'expert linguiste pourra intervenir (partie basse) des composants où il n'aura pas accès (domaine de l'expert linguiste informaticien).
On distingue ainsi six grands modules :

1. **Le gestionnaire de corpus** : l'expert linguiste peut déterminer (ajouter, supprimer, modifier) les corpus qu'il souhaite faire analyser par le système, actuellement soit un fil RSS, soit un site web. Il peut expliciter par ailleurs un certain nombre de méta-informations : nom du journal, url d'entrée, catégorie des informations fournies (presse générale ou spécialisée à l'heure actuelle), domaine (informatique, santé, économie, mode, etc.), langue (parmi les sept langues du projet), pays du journal (cette information pourra servir ultérieurement à étudier des différences néologiques par pays pour une même langue), type de la ressource (site web ou fil RSS actuellement), fréquence de parution. Ces informations sont associées à chaque unité d'information (« article ») qui sera récupérée et pourront permettre de filtrer les résultats dans le moteur de recherche. Nous présenterons dans la troisième partie les corpus actuellement utilisés.
2. **La récupération des fils RSS, des articles liés et leur analyse linguistique** : ce module permet d'effectuer la récupération régulière des articles de presse explicités dans les fils RSS et les pages web et d'effectuer différents traitements linguistiques : segmentation en mots, analyse morphosyntaxique puis syntaxique. Ce module permet d'ajouter à chaque fil de presse des éléments de contenu : titre de l'article, description de l'article (dénotant soit un résumé du contenu, soit une accroche), contenu de l'article lui-même, contenu étiqueté morphosyntaxiquement, lemmes du document (restreints aux catégories nom, verbe et

adjectif), noms propres du document. Nous détaillerons l'état actuel de ce module dans la section 2.2.

3. **Le repérage automatique de néologismes par la méthode du dictionnaire de référence pris comme corpus d'exclusion** : ce module permet, à la suite de l'analyse morphosyntaxique, de ne conserver que des candidats néologismes après plusieurs filtres : noms propres, erreurs typographiques, puis pré-catégorisations des néologismes candidats en emprunts et néologismes internes (voir figure 1). Nous préciserons l'état actuel dans la section 2.3.
4. **Le moteur de recherche et d'analyse des néologismes** : cette interface permet de fouiller les résultats obtenus par les étapes précédentes via un moteur de recherche comprenant différentes propriétés précisées en 2.4.
5. **Le gestionnaire de néologismes** est une base de données préexistante au projet développée en collaboration avec Jean-François Sablayrolles au LDI. Nous renvoyons à (Cartier et Sablayrolles, 2010) pour le détail de ce module. Neologia est en interaction avec le moteur Neoveille de deux façons principales : d'une part, les néologismes présentés et leurs contextes peuvent être directement exportés dans la base Neologia ; d'autre part, il est toujours possible d'obtenir des informations sur le cycle de vie des néologismes après son insertion dans Neologia, par retour au moteur Neoveille.
6. **Le repérage des néologismes sémantiques par la méthode du profil combinatoire** est lancé sur les lexies cibles et sera également disponible dans l'interface de recherche et d'analyse. Elle ne sera pas décrite dans le présent document.

Nous examinons dans les parties suivantes les modules 2, 3 et 4.

2.2. Module de récupération des fils RSS, des articles liés et leur analyse morphosyntaxique puis syntaxique

Ce module comprend lui-même plusieurs étapes : à partir de la liste des sources contenues dans le gestionnaire de corpus, une récupération régulière (dépendant de la fréquence de parution) des informations textuelles est lancée. Pour les fils RSS (« unité d'information »), des méta-informations sont récupérées lorsqu'elles sont explicitées : titre du document, description du document (court résumé ou accroche sur le contenu du document), date de publication du document, catégorie du document, mots-clés associés au document. De plus, le lien vers l'article complet est utilisé pour récupérer le document complet, à l'aide d'un outil de zonage permettant de ne conserver que les parties textuelles pertinentes sur la page.

À la suite de cela, une analyse linguistique en trois étapes est lancée :

- **segmentation en mots** : à cette étape est générée la liste des mots ; un calcul de leur fréquence est également effectué ; à partir de ces résultats, pour les langues dont nous ne disposons pas d'analyseur morphosyntaxique de niveau suffisant, une correction orthographique est lancée, puis il est procédé à la recherche des mots inconnus du dictionnaire de référence/d'exclusion. C'est aussi dans cette phase qu'est effectuée l'analyse distributionnelle pour retrouver des profils combinatoires ; il est prévu également dès cette étape une reconnaissance des lexies composées[4].
- **analyse morphosyntaxique du texte** : cette étape concerne les langues pour lesquelles on dispose d'un tel outil et permet de récupérer un texte annoté avec les parties du discours, et une indication « mot inconnu » pour les mots non-reconnus ; cette analyse permet de dégrossir la liste des candidats néologismes, et permet également, sur des bases heuristiques, de récupérer la liste des noms propres et des lemmes appartenant aux catégories verbe, nom ou adjectif. À partir du résultat, la correction orthographique puis la recherche par dictionnaire de référence/d'exclusion est lancée, ainsi que l'analyse distributionnelle pour retrouver des profils combinatoires ;
- **analyse syntaxique** : cette analyse permettra de travailler sur un texte doté des fonctions syntaxiques des groupes de mots, et permet de lancer une analyse distributionnelle de plus haut niveau.

[4]Ce module sera implémenté dans la prochaine livraison de la plateforme.

-
- Dans ce cadre, nous avons identifié les dictionnaires électroniques disponibles ainsi que les analyseurs morphosyntaxiques utilisés pour chacune des sept langues. Nous évoquerons ces points dans la section 3.2.

2.3. Module de repérage des néologismes par la méthode d'un dictionnaire de référence pris comme corpus d'exclusion (désormais méthode DRE)

Le module de repérage des néologismes de forme par la méthode DRE peut être lancé soit à partir du texte segmenté, soit à la suite de l'analyse morphosyntaxique. Actuellement, pour le français, nous utilisons TreeTagger, y récupérons les mots inconnus, puis utilisons plusieurs filtres pour écarter les noms propres, un certain nombre d'erreurs orthographiques et des erreurs issues des traitements précédents (voir section 3. Premiers résultats, pour une analyse). Pour les autres langues, nous sommes en phase d'évaluation des différents analyseurs morphosyntaxiques.

2.4. Moteur de recherche et d'analyse des néologismes

Le moteur de recherche[5] comporte un certain nombre de fonctionnalités, dont certaines encore en développement. On trouvera en figure 3 une prise d'écran de l'interface actuelle.

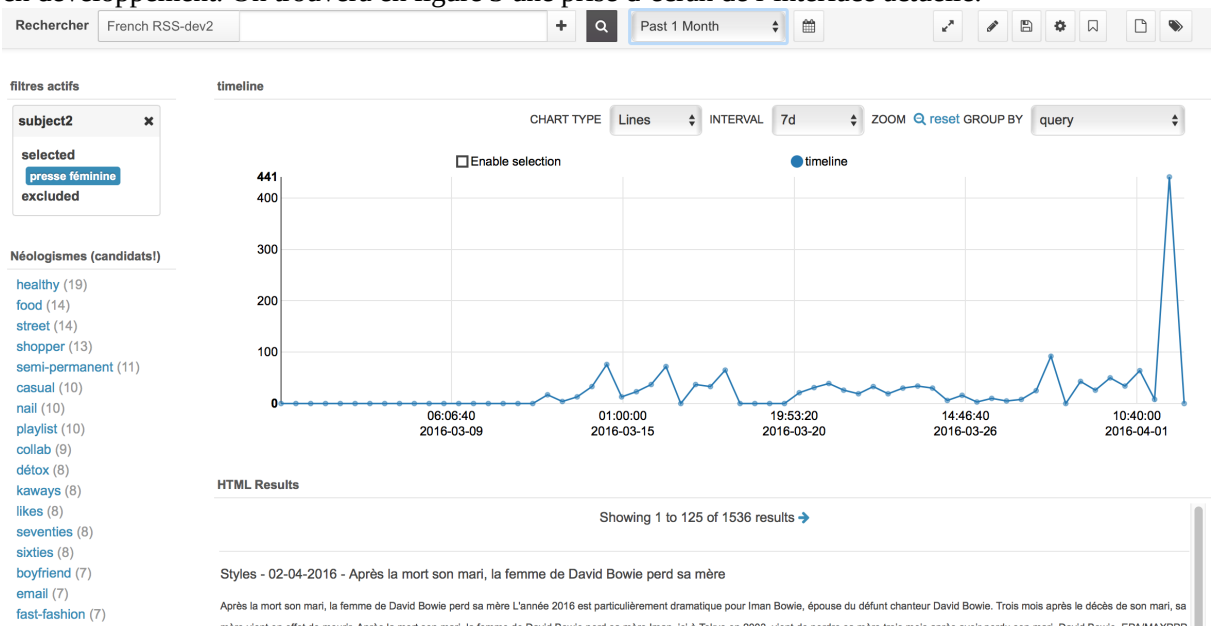

Figure 3 : capture d'écran de l'interface de recherche et d'analyse (presse féminine)

L'écran se présente sous la forme générique d'un moteur de recherche avec la zone de recherche (en haut à gauche) et les résultats (en bas). Mais plusieurs fonctionnalités additionnelles sont présentes : d'une part, la liste des néologismes candidats repérés automatiquement (sur la gauche ainsi que d'autres filtres (permettant de restreindre le corpus à un type de journal, à un domaine, etc.) ; une vision temporelle (au-dessus des résultats) qui présente l'évolution des occurrences.

Nous présentons ci-après dans le détail les différentes fonctionnalités proposées.

2.4.1. Recherche simple et complexe

[5]L'interface du moteur de recherche est basée sur Hue (www.gethue.com)

Le moteur de recherche, qui est basée sur l'outil Apache Solr[6], propose des requêtes simples ou plus complexes. Il est ainsi possible de rechercher un mot simple, un mot composé (entre guillemets), mais également un mot tronqué. Par exemple, on peut rechercher tous les mots commençant par le préfixe *anti-* (anti*), ou encore toutes les formes finissant par *-èle* (*èle). Ce type de recherche est très utile pour retrouver des formes nouvelles issues de la productivité préfixale ou suffixale, ou même des composés.

Une recherche par expression régulière est également disponible, en mettant la chaîne recherchée entre slashes droits (//). Par exemple pour retrouver les néologismes commençant par le fracto-lexème *e-* on peut saisir /^e-/.

Il est également possible de restreindre la recherche à certains journaux, à certaines catégories d'informations et de croiser les critères de multiples façons.

2.4.2. Listes des candidats néologismes

Une liste de candidats néologismes est proposée sur la gauche de l'écran, triés par défaut par fréquence d'apparition. Chaque néologisme est cliquable et permet de visualiser les documents dont il est issu, et de surligner les occurrences elles-mêmes, comme le montre la figure 4. À terme, il sera également possible de visualiser l'évolution fréquentielle des néologismes, dès que nous aurons une couverture suffisante en diachronie courte.

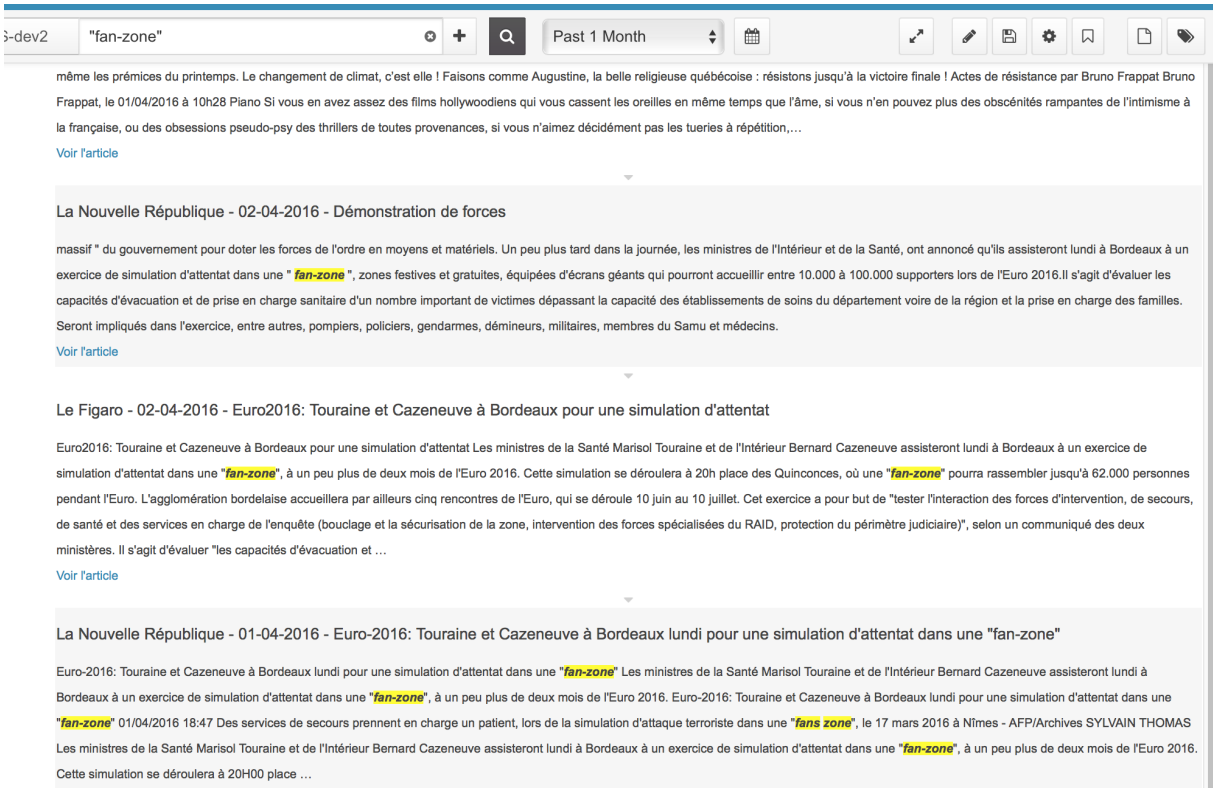

Figure 4 : surlignage des occurrences à partir de la requête « *fan-zone* »

Les candidats néologismes automatiquement repérés peuvent être édités par l'utilisateur, comme le montre la figure 5. En effet, la reconnaissance automatique n'est pas (ne peut pas être ?) complètement fiable : parmi les mots « inconnus », on rencontre des mots simples ou composés de la langue générale ou de langues spécialisées, des erreurs typographiques non corrigées, des erreurs de segmentation des mots, ainsi que de vrais néologismes. Il faut donc donner à l'expert linguiste la

[6]Apache Solr (http://lucene.apache.org/solr/) est un logiciel open Source sous licence Apache basé sur Apache Lucene.

possibilité de sélectionner les lexies de cette liste et de décider à quelle catégorie elles appartiennent. Cette procédure permet :

- d'ajouter des lexies aux dictionnaires de référence (mots simples, mots composés de la langue générale, dictionnaire spécialisé) ; il s'agit donc d'une procédure permettant d'améliorer la couverture lexicographique des dictionnaires de référence ;
- d'ajouter des mots au dictionnaire additionnel des erreurs typographiques et autres erreurs ; ce cycle permettra à terme d'améliorer les résultats du repérage des néologismes ;
- d'exporter les vrais néologismes vers Neologia, la base de données de description linguistique des néologismes.

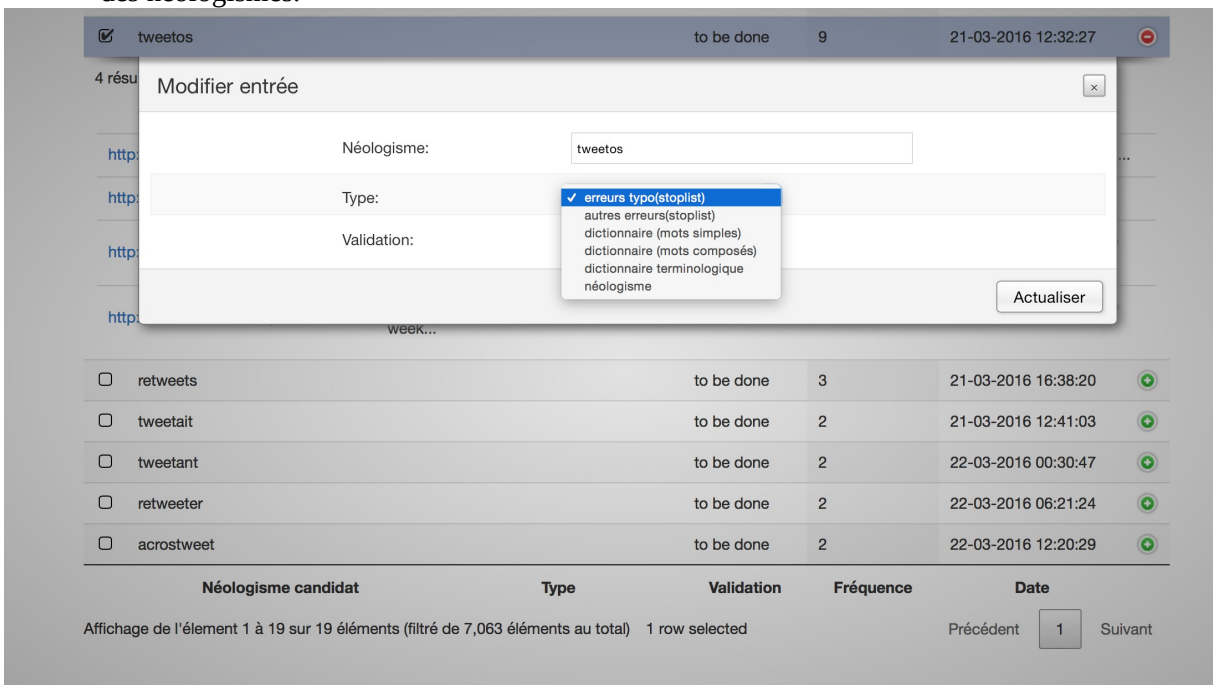

Figure 5 : formulaire d'édition et de catégorisation des néologismes candidats

2.4.3. Filtres

En dehors de la liste des néologismes, le moteur propose des filtres permettant de restreindre les corpus étudiés par domaine, par journal, ou même selon d'autres contraintes. Il suffit pour cela de sélectionner par exemple les journaux à étudier. Cette possibilité permet un filtrage intéressant pour étudier les zones de propagation d'un néologisme.

Ces quelques fonctionnalités sont secondées par d'autres calculs et visualisations qui seront implémentés, par exemple pour obtenir des tableaux croisés dynamiques permettant de visualiser la courbe fréquentielle d'un néologisme selon le domaine, ou selon les types de journaux.

3. Premiers résultats, premières analyses

Dans cette section, nous présentons les premiers résultats du projet, qui a commencé en juin 2015.

3.1. Gestion des sources d'informations : état des lieux

Les corpus utilisés sont, à l'heure actuelle, exclusivement des fils RSS de presse généraliste, dans les sept langues. Le tableau 2 synthétise la répartition par langue.

Langue	Nombre de fils de presse

Chinois	16
Français	77
Grec	37
Polonais	26
Portugais du Brésil	20
Russe	37
Tchèque	24
	237

Tableau 2 : répartition des fils de presse par langue

Ces fils de presse sont récupérés deux fois par jour. Depuis le début du projet en juin 2015, nous avons récupéré plus de 1,5 million d'articles de presse.

Pour ce qui concerne le français, les fils de presse se répartissent en presse nationale (60) / presse régionale (17), ainsi que par domaines : généraliste (44) pour plus de la moitié, spécialisés pour les autres (économie, sport, santé, politique, société industrie, sciences (biologie, nature, high-tech, espace, physique), informatique, presse féminine (cuisine, mode, beauté, lifestyle), ces trois derniers domaines correspondant à des travaux spécifiques dans le cadre du projet.

Pour les autres langues, tous les fils sont actuellement de type généraliste.

Il est à noter que ces sources d'informations peuvent d'ores et déjà être gérées (ajout, modification, suppression) via le gestionnaire de corpus.

3.2. Repérage automatique des néologismes de formes

Le repérage des néologismes de forme utilise la méthode DRE. Pour chacune des langues, nous avons identifié les différents dictionnaires électroniques, les différents analyseurs morphosyntaxiques ainsi que les corpus de référence disponibles. En effet, trois méthodes peuvent être utilisées : soit on identifie les formes manquantes directement après une segmentation en mots, à partir du ou des dictionnaires de référence et d'exclusion ; soit on utilise un analyseur morphologique identifiant par lui-même les mots inconnus, qui devront ensuite être filtrés pour distinguer les noms propres, les erreurs typographiques et d'autres erreurs. Soit aucune ressource n'est disponible ou de qualité suffisante, et nous pouvons utiliser un corpus de référence permettant d'extraire les formes attestées sur une période antérieure contemporaine. Après identification des différentes sources, nous avons opté pour Treetagger pour le français et le portugais du Brésil, car c'est seulement pour ces deux langues que la qualité d'analyse est suffisante. Encore faut-il prévoir un post-traitement, car cet outil n'est pas aujourd'hui basé sur des corpus d'apprentissage suffisamment récents. Pour les autres langues, nous avons opté pour une segmentation en mots puis pour une technique utilisant des dictionnaires open source de correction orthographique issus de l'outil Hunspell[7], très largement utilisé dans les applications bureautiques et sur Internet.

Nous nous concentrerons ici sur le traitement du français. L'outil Treetagger génère pour les formes non reconnues, une étiquette lemme <unknown>, ce qui permet de restreindre les post-traitements à ce sous-ensemble des lexies.

Ensuite, nous avons le choix entre deux dictionnaires de référence de couverture diamétralement opposée : le dictionnaire Morfetik (Cartier et Grezka, 2015), et le dictionnaire GLAWY (Sajous *et al.* 2015). Morfetik comporte la nomenclature contemporaine la plus étendue, pour les seules lexies simples ; et GLAWY comporte la nomenclature la plus étendue toutes lexies confondues, qu'il s'agisse de mots composés, de noms propres et même de néologismes. Il n'est pas évident d'utiliser cette dernière ressource, d'autant que plusieurs tests (Cartier, 2015) ont montré que sa couverture en lexies simples est certes plus étendue que Morfetik, mais comporte un très grand nombre de lexies très rares, désuètes ou encore spécialisées, ce qui a pu être démontré en comparant sa couverture par rapport au corpus Wikipedia et un corpus de *dix ans du Monde*. Dès lors, nous avons opté pour une solution permettant de construire progressivement une ressource lexicale issue de

[7]http://hunspell.github.io

Morfetik, mais progressivement plus couvrante : après analyse par Treetagger, élimination des noms propres (mots qui débutent par une majuscule, sauf en première position de phrase), élimination des lexies présentes dans Morfetik, et élimination des erreurs typographiques[8], nous obtenons une liste plus réduite de candidats néologismes que nous présentons à l'expert linguiste qui pourra décider s'il s'agit d'un vrai néologisme, d'une lexie simple manquante dans le dictionnaire, d'une lexie complexe manquante dans le dictionnaire, d'une lexie de dictionnaire spécialisé ou encore d'une erreur typographique ou liée aux étapes précédentes du traitement automatique. Ce système, combinant expertise humaine et traitement automatique, permet, au fur et à mesure, d'améliorer la ressource de référence, de construire une liste d'erreurs typographiques fréquentes et de conserver les vrais néologismes. Au fur et à mesure, les extractions se font ainsi plus précises. En six mois, la ressource des lexies simples a ainsi été complétée de près de mille termes, et de plus de trois mille termes pour ce qui concerne la ressource des lexies composées (limitée aux mots à trait d'union). Nous pensons effectuer un travail similaire pour les autres langues, étant donné la relative rapidité de la construction des dictionnaires de référence et d'exclusion.

3.3. analyse des listes de néologismes candidats

Le corpus français constitué jusqu'ici comporte près de deux millions d'occurrences de lexies différentes (noms propres compris). Sur ce total, environ 10 000 lexies sont proposées comme candidats néologismes, soit un pourcentage de 0,5% (sans noms propres, et avant correction orthographique). Le tableau 3 présente les différents types de néologismes-candidats, ainsi que des exemples et le pourcentage sur l'ensemble des mots inconnus :

Type	Exemples	Nombre	%
Erreur typographique	Aïgue, déplait, goutez, gustave, maroc, lequels ; kaways, prefectures...	1238	11,48%
Autres erreurs[9]	Http, play-offsà, occées, nucléairespréventives...	1540	14,28%
Dictionnaire général (mot simple)	Addictif, pancetta, antiterrorisme, narrativement, ingénieure, ndlr...	1967	18,25%
Dictionnaire général (mot composé)	Longs-métrages, peut-être, bande-dessinée	3400	31,54%
Dictionnaire spécialisé	Géocroiseurs, coronavirus, épicatéchine	231	2,14%
Néologismes	Instragrameuse, décret-socle, biotiful, féminicide, spectable, shopper...	2405	22,31%
		10781	100%

Tableau 3 : type des candidats néologismes et répartition sur un corpus de 10 000 lexies inconnues

On remarque que :

1. le contingent le plus important concerne les mots composés à traits d'union, car le dictionnaire de référence actuel est en cours de construction de manière incrémentale ;
2. ensuite viennent les vrais néologismes, que nous étudierons dans la section suivante ;
3. le dictionnaire des mots simples pourra également être amélioré avec le temps ; il est à noter que parmi les lexies non reconnues, un grand nombre de gentilés sont présents, ce qui nous a

[8]Le filtre de correction typographique est limité à deux opérations, c'est-à-dire une distance maximale de 2, ce qui permet d'éliminer la très grande majorité des coquilles liées à l'absence d'une lettre, le doublement d'une lettre, ou encore l'interversion de lettres)

[9]Il est à noter que les autres erreurs proviennent essentiellement d'erreurs dans les traitements précédents, spécialement de la segmentation en lexies. Cette validation manuelle permet ainsi d'identifier les problèmes de ce traitement.

amené à ajouter un dictionnaire de référence spécifique, permettant de couvrir la grande majorité des cas, en extrayant de GLAWY toutes les lexies avec l'indication « gentilés » ;

4. les erreurs typographiques ont ensuite été corrigées à plus de 95% des cas avec un seuil d'édition de 2 (exemples : *goutez*, *gouutez* (*goûtez*) sont corrigés, mais pas *guoutez*) ; mais il y a un risque d'exclure parfois des néologismes.
5. les autres erreurs proviennent d'une mauvaise segmentation, qui peut avoir de multiples sources (structure du document de départ, en général) ; cependant, près de 70% des cas sont aujourd'hui résolus.

Parmi un échantillon de 400 néologismes choisis au hasard (sur les 2405 repérés), nous présentons dans le tableau 4 leur répartition en types, en reprenant les catégories pertinentes de la typologie de (Sablayrolles, 2015) :

Type	**Exemples**	**nombre**	**%**
Préfixation	anti-blasphème, néopatron[10],	92	23,00%
Suffixation	Accidentalité, ailières, orbiteur, organelles, paparazziesque...	52	13,00%
Parasynthétique[11]	déghéttoïsation	13	3,25%
Composition	Anarcho-sataniste, panier-average, panthère-carreaux, photoshoot, robot-bavardeur, paléoartiste, afro-euphorisme	84	21,00%
Composition savante	Acromioclaviculaire	3	0,75%
Mot-valise	Afroptimistes, aoûthlétisme, nymphirmières	13	3,00%
Onomatopée	Ahhh, ahah, ahlala, ouhhhh, adooore	5	1,25%
Troncation	Accro, actu, applis, négo, perf	52	13,00%
Emprunt	barzakh, bashing, buzz...	86	21,50%
totaux		400	100,00%

Tableau 4 : répartition des néologismes par type

On remarque que :

- près de 40% des néologismes sont issus de préfixation/suffixation très productives (morphologie constructionnelle) (*anti- : anti-réac, antiracaille, antisyndicale ; après-, auto(-), avant- ; co(-), cyber(-), demi-, e-, ex-, hyper(-), inter(-), ...*; nous détaillons ces résultats dans la section suivante ;
- la composition est également très présente (*bobos-gogos, boboécolo, café-femmes, camions-épaves, chômeur-directeur, ...*)
- les troncations sont en nombre également conséquent ;
- les emprunts (anglais, mais aussi italiens dans le corpus : *amaretto, spritz, bloggueur, bodybuildés, booktubeuses, boostéer, shopper, borderline, bricogirl, burnout, burn-out, food-truck, cheesecake, coaching, coming-out...*) sont également très présents, mais en plus grande concentration dans les journaux informatiques et la presse féminine, et dans des états d'assimilation plus ou moins avancés.

Répartition des préfixes : étude préalable pour un suivi des affixes et des fractolexèmes les plus productifs

[10]Nous considérons ici que *neo* s'est grammaticalisé, en suivant les listes proposées par *Le Petit Robert* 2015

[11]L'existence de cette catégorie combinant simultanément les deux procédés de préfixation et de suffixation a été fortement mise en cause par Danielle Corbin (1980) avec une argumentation convaincante. Mais il n'est pas impossible que ce procédé existe néanmoins dans certains cas précis.

Nous avons vu que l'affixation est le principal processus néologique utilisé. Ce mécanisme productif fonctionne, selon les grammaires traditionnelles, à partir d'une liste finie de préfixes ou de suffixes. On peut toutefois rapprocher de l'affixation les compositions avec fracto-lexèmes, qui se situent à la frontière entre l'affixation et la composition. Notre corpus montre en effet que la répartition des préfixes/fracto-lexèmes exhibe un certain nombre de fracto-lexèmes très productifs, à tel point qu'on peut se demander si certains formants ne sont pas en cours de grammaticalisation (par exemple *bio* ou même *e-*) Parmi les vingt premiers, on notera (*e-*), ainsi que (*franco-*). Il conviendrait de suivre l'évolution de ces répartitions (voir Cusin-Berche, 2003 ; Makri-Morel 2015 sur ces questions), pour identifier des tendances de la langue française, et des autres langues étudiées.

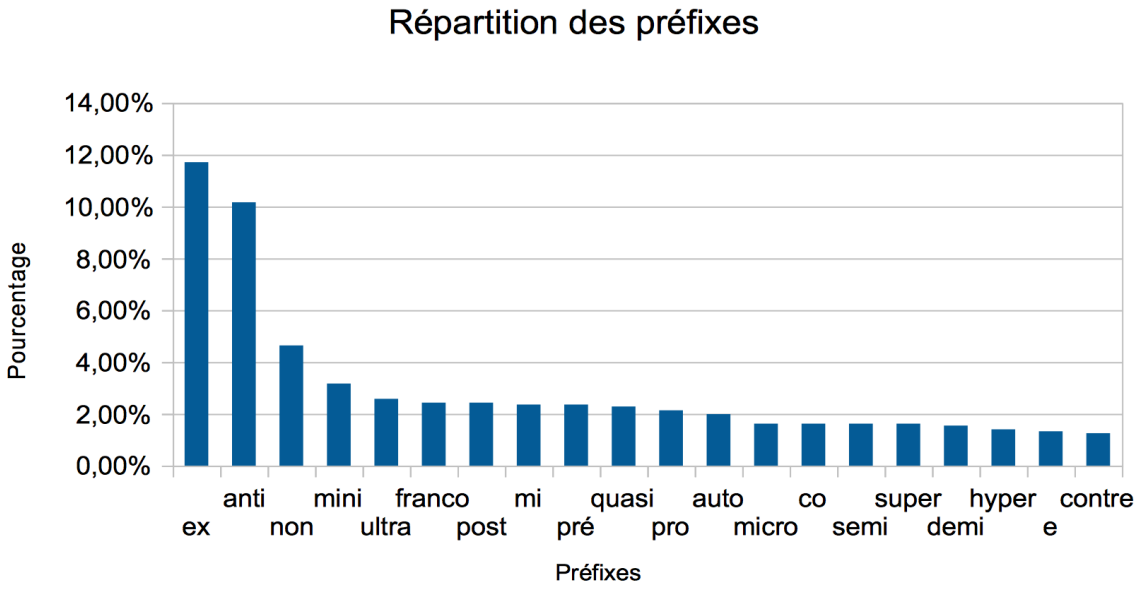

Figure 6 : répartition des préfixes/fractolexèmes utilisés dans le corpus étudié

Répartition par journaux et par catégorie, exemple presse féminine

Un autre type d'étude consiste à scruter la répartition des néologismes, et des types de néologismes, selon le type de corpus. Nous avons ainsi pu effectuer une comparaison entre le corpus « presse générale » et presse féminine (voir figure 7 : pour chaque catégorie, la colonne de gauche concerne la presse générale, la colonne de droite la presse féminine), qui montre clairement la prédominance des emprunts dans la presse féminine, l'affixation et la composition restant très productives.

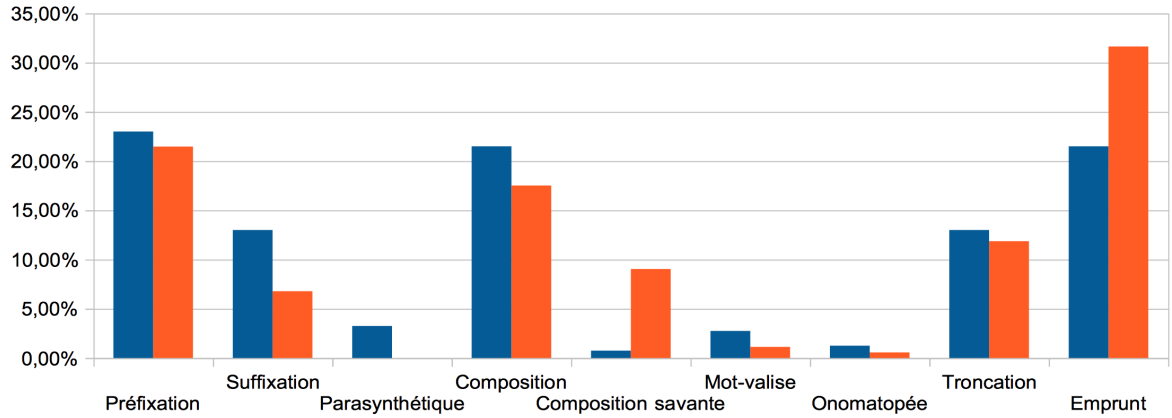

Figure 7 : répartition des néologismes par type, selon le type de corpus

Diffusion des néologismes : exemple de quelques néologismes empruntés à l'anglais dans le domaine informatique

Nous avons vu, dans la première section, que l'une des voies pour évaluer l'assimilation des néologismes est la présence de nombreux dérivés du néologisme initial, ainsi que l'assimilation morphologique du néologisme, notamment lorsqu'il est emprunté. Nous donnons ci-après, issus de notre corpus, les mots de la famille de *tweet* qui ont été rencontrés, qui montre bien que ce terme est désormais tout à fait assimilé par le français, même si subsiste une hésitation entre les graphies -ee- et -i-.

lexie	fréquence
tweetté	13
tweettait	1
tweets	152
tweetos	11
tweetée	13
tweete	461
tweetant	2
tweetait	2
retweets	3
retweeter	2
retweetées	13
retweetée	49
retweeté	25
retweetant	17
retweet	13
macro-tweeting	14
live-tweet	15
acrostweet	11

Tableau 5 : lexies dérivées du néologisme *tweet*

Conclusion

Dans cet article, nous avons essayé de présenter les caractéristiques « idéales » d'un système de repérage automatique et de suivi des néologismes. Après un rapide état de la question sur les différentes aspects de la néologie, d'un point de vue linguistique, puis du point de vue du traitement automatique des langues, nous avons présenté l'architecture de Neoveille et les premières réalisations de cette plateforme de veille néologique qui vise à répondre à trois grands principes : la simplicité d'utilisation ; le paramétrage par l'expert linguistique des différentes ressources linguistiques (corpus, dictionnaires de référence, dictionnaire de néologismes) ; le potentiel d'étude et d'analyse des néologismes en corpus.

Cette plateforme est encore aujourd'hui en développement, mais les partenaires linguistes du projet ont déjà commencé à l'exploiter et à faire des remarques pour son amélioration. Elle est désormais disponible par identifiant et mot de passe à l'adresse suivante : http://tal.lipn.univ-paris13.fr/neoveille (ou http://www.neoveille.org). Un accès libre sera prochainement mis à disposition, qui présentera les néologismes apparus chaque semaine.

Bibliographie

Alex B. (2008), « Comparing corpus-based to web-based lookup techniques for automatic English inclusion detection ». In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, pages 2693–2697. Marrakech, Morocco

Blumenthal, P. (2009), « Éléments d'une théorie de la combinatoire des noms », in Blumenthal, P. / Petit, G. (éds.) : *Cahiers de lexicologie* 94 (2009-1), 11-29.

Bréal, M. (1897), *Essai de sémantique : science des significations*, Paris, Hachette.

Bybee J. (2016), *Language Change*, Cambridge Textbook in Linguistics, Cambridge University Press.

Cabré M.T. et de Yzaguirre L. (1995), « Stratégie pour la détection semi-automatique des néologismes de presse », *TTR : traduction, terminologie, rédaction*, vol. 8, n° 2, p. 89-100.

Cabré, M. T., Domènech, M., Estopà, R., Freixa, J., and Solé, E. (2003), « L'observatoire de néologie: conception, méthodologie, résultats et nouveaux travaux ». In Sablayrolles Jean-François, *L'innovation lexicale*, p. 125–147.

Cabré, T. and Nazar R. (2011), « Towards a new approach to the study of neology » In *Neology and Specialised Translation* 4th Joint Seminar Organised by the CVC and Termisti.

Cabré, T. et Nazar R. (2012), « Towards a New Approach to the Study of Neology », *Neologica 6*, p 63-80.

Cartier E. (2011) « Utilisation des contextes dans le cadre dictionnairique : état des lieux, typologie des contextes, exemple des contextes définitoires », in *Actes des Huitièmes Journées scientifiques du Réseau de chercheurs Lexicologie, terminologie, traduction*, Lisbonne, 15-17 octobre 2009 p.619-632.

Cartier E., Grezka A. et Mathieu-Colas M. (2015), « Dictionnaires morphologiques du français contemporain : présentation de Morfetik, éléments d'un modèle pour le TAL », *TALN* 2015, Caen, 18-23 June 2015.

Cartier, E. (2012), « Néologie et Traitement Automatique des Langues », Numéro Spécial « la néologie », *Langages*, 183, 2012.

Cartier, E. et Sablayrolles, J.-F (2010), « Neologia, une base de données pour la gestion des néologismes », in Teresa Cabré, Ona Domènech, Rosa Estopà, Judit Freixa, Mercè Lorente (eds.), *Actes del I Congrés Internacional de neologia de les Llengües Romàniques (Barcelone (7-10 mai 2008)*, Barcelone, Université Pompeu Fabra, IULA, sèrie activitats 22, p. 759-767.

Cartier, E. et Sablayrolles, J.-F (2011), « Nouvelles technologies, nouveaux modèles linguistiques et néologie », dans Ponchon Thierry et Laborde-Milaa Isabelle (éds), *Sciences du langage et nouvelles technologies*. Actes du colloque 2009 de l'Association des Sciences du Langage, Limoges, Éditions Lambert-Lucas, 2011, p. 53-59.

Charnois, T. (2011), *Accès à l'information : vers une hybridation fouille de données et traitement automatique des langues*. Habilitation à Diriger des Recherches, Université de Caen, 1er décembre

2011
Charnois, T., Plantevit M., Rigotti C. and Crémilleux B. (2009), « Fouille de données séquentielles pour l'extraction d'information dans les textes ». *TAL*, 50(3) : 59–87.
Cook P. and Stevenson S. (2010), « Automatically identifying the source words of lexical blends in English". *Computational Linguistics*, 36(1):129–149.
Cook, P. and Hirst G.(2011), "Automatic identification of words with novel but infrequent senses ». In *Proceedings of the 25th Pacific Asia Conference on Language Information and Computation (PACLIC 25)*, pages 265–274. Singapore, page 265–274.
Corbin D., « Contradictions et inadéquations de l'analyse parasynthétique en morphologie dérivationnelle », *Théories linguistiques et traditions grammaticales*, A.-M. Dessaux-Berthonneau, coll. " Linguistique ", Lille, 1980, p. 181-224.
Croft, W. (2000), *Explaining Language Change. An Evolutionary Approach*. Harlow:Pearson Education.
Croft, W. (2007) *Construction grammar*. In Geeraerts and Cuyckens, eds., 463–508.
Darmesteter, A. (1886). *La vie des mots*. Paris : Delagrave.
Evert, S. and Hardie, A. (2011), « Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium ». In *Proceedings of the Corpus Linguistics 2011 conference*, University of Birmingham, UK.
Falk I., Bernhard D., Gérard C. (2014) , From Non Word to New Word: Automatically Identifying Neologisms in French Newspapers *LREC - The 9th edition of the Language Resources and Evaluation Conference*, May 2014, Reykjavik, Iceland. 2014, Proceedings of the International Conference on Language Resources and Evaluation
Fillmore, C.J., Kay P., O'Connor C. (1988), "Regularity and idiomaticity in grammatical constructions: the casr of let alone", *Language* 64,3 p. 501-538.
Firth, J. R. (1957), « A synopsis of linguistic theory 1930–1955 ». In *Studies in Linguistic Analysis*, pp. 1–32. Blackwell, Oxford.
Garcia-Fernandez A., Ligozat A.-L., Dinarelli M., and Bernhard D. (2011), « Méthodes pour l'archéologie linguistique: datation par combinaison d'indices temporels » *Actes du septième DÉfi Fouille de Textes*.
Geeraerts D. (2009) *Theories of Lexical Semantics*, Oxford University Press.
Gérard C., Falk I., Bernhard D. (2014), « Traitement automatisé de la néologie : pourquoi et comment intégrer l'analyse thématique ? », *CMLF 2014*, Jul 2014, Berlin, Germany pp.2627 – 2646.
Gérard C., Kabatek J. (2012), « La néologie sémantique en question : quelles conceptions pour quelles méthodes ? » *Cahiers de lexicologie*, 2012, 1 (100), p. 11-36.
Gérard Christophe, Falk Ingrid, Bernhard Delphine (2014), « Traitement automatisé de la néologie : pourquoi et comment intégrer l'analyse thématique ? » *Actes du 4e Congrès Mondial de Linguistique Française (CMLF 2014)*, Jul 2014, Berlin, Allemagne. 8, pp.2627 - 2646, 2014
Gevaudan P. et Koch P. (2010), « Sémantique cognitive et changement sémantique », *Grandes voies et chemins de traverse de la sémantique cognitive*, Mémoire de la Société de linguistique de Paris, XVIII, p. 103-145.
Goldberg, A.E. (1995), *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.
Goldberg, A.E. (2003), « Constructions: A new theoretical approach to language ». *Trends in Cognitive Sciences* 7: 219–224.
Goldberg, A.E. (2013), « Constructionist Approaches, » in *The Oxford Handbook of Construction Grammar*, Edited by Thomas Hoffmann and Graeme Trousdale, Oxford University Press.
Harris, Z. (1954), « Distributional structure ». *Word*, 10(23), 146–162. Traduction 1970.
Harris, Z. (1988), *Language and Information*. New York: Columbia University Press, ix, 120 pp.
Hildenbrand Z., Kaprzak A. et Sablayrolles J.-F éd. (2016), *Emprunts néologiques. Études interlangues*, Limoges, Éditions Lambert-Lucas, collection « La Lexicothèque ».
Jacquet-Pfau C. (2003), « Du statut de l'emprunt en traitement automatique des langues », in J.-F. Sablayrolles (dir.), *L'innovation lexicale*, Honoré Champion, p. 79-97.
Janssen, M. (2012), « NeoTag: a POS Tagger for Grammatical Neologism Detection ». In *LREC 2012*, page 2118–2124
Kabatek, J. et Christophe G. (dir.) (2012), *Néologie sémantique et analyse de corpus*, *Cahiers de lexicologie* 100 (2012)

Kang B.-J. and Choi K.-S. (2002), « Effective foreign word extraction for Korean information retrieval ». *Information Processing and Management*, 38(1):91–109.
Kerremans, D., Stegmayr S. and Schmid H.-J. (2012), « The NeoCrawler: identifying and retrieving neologisms from the internet and monitoring on-going change ». In: Kathryn Allan and Justyna A. Robinson, eds., *Current methods in historical semantics*, Berlin etc.: de Gruyter Mouton, 59-96.
Kilgarriff, A., Pavel R., Pavel S., and Tugwell D. (2004), « The Sketch Engine ». In *Proceedings of Euralex*, pages 105–116, Lorient.
Langacker, R. W. (1987), *Foundations of cognitive grammar. Vol. 1, Theoretical prerequisites*. Stanford, CA: Stanford University Press.
Langacker, R. W. (1991), *Foundations of cognitive grammar. Vol. 2, Descriptive application*. Stanford, CA: Stanford University Press.
Langages n° 183, (2011), « Néologie ; nouveaux modèles théoriques et NTIC », co-dirigé par Jean-François Sablayrolles et Salah Mejri, Armand Colin.
Lau J. H., Cook P., McCarthy D., Newman D., and Baldwin T. (2012), « Word sense induction for novel sense detection ». In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, page 591–601.
Makri-Morel J., (2015) « Mots-valises : quand les segments communs se font la malle », *Neologica* n° 9, p. 61-79.
Meillet, A. (1906), « Comment les mots changent de sens » in *L'Année sociologique*, repris dans *Linguistique historique et linguistique générale*, Paris : Champion.
Ollinger S. and Valette M. (2010), « La créativité lexicale : des pratiques sociales aux textes ». In *Actes del I Congrés Internacional de Neologia de les llengües romaniques*, volume Publicacions de l'Institut Universitari de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra (UPF), pages 965–876, Barcelona, Spain.
Pruvost J. et Sablayrolles, J.-F. (2003), *Les néologismes*, Que sais-je ? n° 3674, PUF, rééd. 2012
Rastier F., Valette M. (2009), « De la polysémie à la néosémie », La problématique du mot, S. Mejri, éd., *Le français moderne*, 77, p. 97-116.
Renouf A. (2010), « Identification automatique de la néologie lexicologique et sémantique : questions soulevées par notre méthode ». Cabré, M.T.; Domènech, O.; Estopà, R.; Freixa, J.; Lorente, M. (eds.). *Actes del I Congrés Internacional de Neologia de les Llengües Romàniques*. Barcelona: Institut Universitari de Lingüística Aplicada; Documenta Universitaria, 2010. 129-141.
Renouf, A & A. Kehoe (2013), « Filling the gaps: Using the WebCorp Linguist's Search Engine to supplement existing text resources ». *International Journal of Corpus Linguistics*, 18(2): 167-198 (John Benjamins).
Renouf, A. (1993): « Sticking to the Text: a corpus linguist's view of language » in *ASLIB Proceedings*, Vol. 45/5, May 1993.
Renouf, A. (1994): « Corpora and Historical Dictionaries », in *Early Dictionary Databases*, eds. Ian Lancashire, & Russon Wooldridge, Univ. of Toronto, Oct 1-8, 1993.
Renouf, A. (1996): with Baayen, Harald, « Chronicling the Times: Productive Lexical Innovations in an English Newspaper », *Language*, 72.1, pp. 69-96.
Renouf, A. (2014), « Semantic Neology: the challenges for automatic identification » *Neologica* 8, p. 185-220.
Sablayrolles J.-F. (2000), *La néologie en français contemporain*, *Examen du concept et analyse de productions néologiques récentes*, coll. Lexica mots et dictionnaires, Paris, Champion,
Sablayrolles J.-F. (2002), « Fondements théoriques des difficultés pratiques du traitement des néologismes », *Revue française de linguistique appliquée*, vol. VII-1. / juin 2002 « Lexique : recherches actuelles », pp. 97-111.
Sablayrolles J.-F. (2006), « Métaphore et évolution du sens des unités lexicales », *Cahiers du CIEL 2000-2003*, Université Paris 7, septembre 2006, pp. 109-124.
Sablayrolles J.-F. (2012), « Extraction automatique et types de néologismes : une nécessaire clarification », Actes du colloque Néologie sémantique et corpus, une rencontre de méthodes, Université de Tübingen, 29-30 avril 2010, organisée par Christophe Gérard et Johanes Kabatek *Les Cahiers de lexicologie* n° 100, juillet 2012, p. 37-53
Sablayrolles J.-F. (2015), « Quelques remarques sur une typologie des néologismes : Amalgamation ou télescopage : un processus aux productions variées (mots valises, détournements…) et un tableau hiérarchisé des matrices », Actes de CINEO II, Sâo Paulo, 5-8 décembre 2011, *Neologia das linguas romanicas*, Ieda maria Alves et Eliane Simões Pereira éd., Sâo paulo, Humanitas, p. 187-218.

Sagot, B., Nouvel, D., Mouilleron, V., & Baranes, M. (2013). « Extension dynamique de lexiques morphologiques pour le français à partir d'un flux textuel ». In *TALN 2013* (pp. 407–420).
Sajous F. et Hathout N. (2015). « GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary. » *Proceedings of eLex 2015 conference*, pp. 405-426, Herstmonceux, England.
Schmid H. (1995), « Improvements in Part-of-Speech Tagging with an Application to German ». *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.
Schmid H.-J. (2007), « Entrenchment, salience and basic levels » In: Dirk Geeraerts and Hubert Cuyckens, eds., *The Oxford Handbook of Cognitive Linguistics*, Oxford: Oxford University Press, 117-138.
Schmid H.-J. (2008), « New words in the mind: Concept-formation and entrenchment of neologisms. » *Anglia* 126 (1), 1-36.
Schmid H.-J. and Küchenhoff H. (2013), « Collostructional analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems and cognitive underpinnings». *Cognitive Linguistics* 24(3), 531-577.
Schmid, H.-J. (2015), « A blueprint of the Entrenchment-and-Conventionalization Model», *Yearbook of the German Cognitive Linguistics Association* 3, 1-27.
Schmid, H.-J., ed. (to appear 2016), Entrenchment, memory and automaticity. The psychology of linguistic knowledge and language learning. Boston: APA and Walter de Gruyter.
Stefanowitsch A. and Gries S. (2003), « Collostructions : investigating the interaction of words and constructions », *International Journal of Corpus Linguistics* 8/2, 209-243.
Tournier J. (1985), *Introduction descriptive à la lexicogénétique de l'anglais contemporain* , Paris-Genève, Champion- Slatkine,1985.
Tournier, J. (1991), *Structures Lexicales de l'anglais. Guide Alphabétique*, Nathan, Paris.
Traugott E.C. and Trousdale G. (2013), *Constructionalization and Constructional Changes*. Oxford: Oxford University Press.
Turney P. and Pantel P. (2010), « From Frequency to Meaning: Vector Space Models of Semantics ». Journal of Artificial Intelligence Research (JAIR), 37(1):141-188. AI Access Foundation.
Valette, M. (2010), « Méthodes pour la veille lexicale » in Actes de la journée d'étude *'Le dictionnaire électronique'. Quelles perspectives pour les sciences humaines et sociales*?' (Kenitra, le 7 décembre, 2007). Messaoudi, Leila (ed.). Publication du laboratoire Langage et société. Université Ibn Kenitra, Maroc. µhttp://hal.archives-ouvertes.fr/hal-00438627/§