

Détection automatique, description linguistique et suivi des néologismes en corpus : point d'étape sur les tendances du français contemporain

Emmanuel Cartier¹, Jean-François Sablayrolles², Najet Boutmgharine³, John Humbley³, Massimo Bertocci¹, Christine Jacquet-Pfau⁴, Natalie Kübler³ et Giovanni Tallarico⁵

¹ Université Paris 13, LIPN-RCLN UMR 7030 CNRS, Labex EFL, 99 avenue Jean-Baptiste Clément, 93430 Villetaneuse, France, emmanuel.cartier@lipn.univ-paris13.fr

² Université Paris 13, HTL UMR 7597 CNRS, Case postale 7034, 5 rue Thomas Mann, 75205 Paris cedex 13, France

³ Université Paris 7, CLILLAC-ARP EA 3967, Case postale 7002, 5 rue Thomas Mann, 75205 Paris, France

⁴ Collège de France et LT2D, Lexiques, textes, discours, dictionnaires, université de Cergy-Pontoise, France

⁵ Università degli Studi di Verona, Dipartimento di Lingue e Letterature Straniere, Italie

Résumé. L'objet du présent article est de faire un point d'étape sur un projet de recherche, Néoveille, commencé en 2015 et visant à confectionner une plateforme pour détecter automatiquement, décrire linguistiquement et suivre le cycle de vie des néologismes en corpus dynamique contemporain, en sept langues. Nous nous concentrons ici sur les résultats obtenus sur le français. Nous introduisons tout d'abord quelques jalons théoriques permettant de situer le phénomène néologique. Ensuite, nous présentons la plateforme web qui a été construite et qui offre aux linguistes un outil pour la détection automatique, l'analyse linguistique et de le suivi des néologismes sur corpus dynamique. Enfin, une dernière partie fait un point sur les tendances néologiques du français contemporain, suite à la collecte automatique, à la validation manuelle et à la description linguistique des innovations lexicales détectées depuis 2015 dans près de 250 sources de presse.

Abstract. Automatic Detection, Linguistic Description and Life-cycle of Neologisms in Corpus : French Contemporary Tendencies from the Neoveille Platform. This article describes a research project, Néoveille, that began in 2015, whose goal is to setup a web platform to automatically detect, linguistically analyse and follow the life-cycle of neologisms in contemporary monitor corpora, in seven languages. Here, we focus on the results obtained for the French language. We first introduce several theoretical hypotheses to define the concept of lexical innovation. We then present the web platform and its components. Finally, we detail the current results and trends for the French language, from the automatic detection, the manual validation and linguistic description of more than 20.000 neologisms in about 250 web newspapers since 2015.

Les néologismes, ou innovations lexicales, ont souvent et longtemps été considérés comme un phénomène secondaire, notamment de par la rareté de leurs occurrences en discours et leur faible impact apparent sur la structure des langues. Pourtant, ces innovations fournissent de précieuses informations sur différents mécanismes linguistiques sous-jacents, de l'emprunt à l'affixation, en passant par la composition et les phénomènes de réduction. Pour faible que soit le taux de néologisme dans les textes, il n'en reste pas moins que le mécanisme néologique est la preuve de la vitalité d'une langue et de sa créativité, et constitue une propriété essentielle des langues vivantes. Les néologismes fournissent également de précieuses informations sur les tendances sociétales, qu'il s'agisse de besoins expressifs ou encore d'utilisations liées au prestige et permettent ainsi de tracer les contours de groupes sociaux spécifiques. Avec l'avènement des technologies et corpus numériques, il est maintenant possible d'étudier ce phénomène à grande échelle. L'objet du présent article est de faire un point d'étape sur un projet de recherche, Néoveille, commencé en 2015 et visant à confectionner une plateforme pour détecter automatiquement, décrire linguistiquement et suivre le cycle de vie des néologismes en corpus dynamique contemporain, en sept langues. Nous nous concentrons ici sur les résultats obtenus sur le français. Nous introduisons tout d'abord quelques jalons théoriques permettant de situer le phénomène néologique. Ensuite, nous présentons la plateforme web qui a été construite et qui offre aux linguistes un outil pour la détection automatique, l'analyse linguistique et de le suivi des néologismes sur corpus dynamique. Enfin, une dernière partie fait un point sur les tendances néologiques du français contemporain, suite à la collecte automatique, à la validation manuelle et à la description linguistique des innovations lexicales détectées depuis 2015 dans près de 250 sources de presse. L'ensemble des données collectées et analysées sont disponibles sur la plateforme du projet¹.

1. Jalons théoriques

Même si l'étude de l'étymologie des mots a été pratiquée depuis l'Antiquité, c'est seulement dans la dernière période de la linguistique historique qu'une réflexion théorique sur « la vie des mots » apparaît (Hermann, 1886; Darmesteter, 1887; Bréal, 1899; Meillet, 1904; Carnoy, 1927; Stern, 1931). Avec la naissance de la linguistique structurale, les considérations diachroniques passent au second plan et il faudra attendre la dernière partie du XX^{ème} siècle pour voir reparaître des travaux sur le cycle de vie des mots, d'une part dans la sphère européenne autour des travaux de (Tourmier, 1985; Sablayrolles, 2000), de (Rastier et Valette, 2009) et de la linguistique cognitive (Koch, 2000; Gévaudan et Koch, 2010); d'autre part, dans la sphère américaine et anglo-saxonne, avec la linguistique de corpus, la linguistique cognitive puis les grammaires de construction, qui se saisiront de la problématique de l'innovation lexicale. De même, la linguistique variationniste (Coseriu, 1964 ; Weinreich *et al.*, 1968 ; Labov, 1994, 2001) explicitera des notions qu'il convient de prendre en compte. Nous évoquerons ci-après les éléments qui nous ont été utiles en amont des travaux que nous avons menés en néologie.

1.1 Typologie des néologismes

Les premières typologies des "mots nouveaux" sont dues à (Carnoy, 1927; Stern, 1931), qui articulent leur typologie autour de l'opposition *onomasiologique* (des sens aux mots pour les exprimer) / *sémasiologique* (des mots aux sens qu'ils expriment). Ils distinguent la néologie formelle, combinant un sens nouveau à une forme nouvelle, et la néologie sémantique, qui combine une forme existante à un sens nouveau. Les typologies proposées reprennent les analyses classiques, concernant la formation des mots, (dérivation, composition, emprunt), et les évolutions sémantiques (extension/restriction de sens,

évolution par analogie (métaphore), évolution par contiguïté (métonymie). Dans le cadre du présent projet, nous sommes partis d'une typologie plus récente, dont l'objectif est de mettre au jour les *matrices lexicogéniques*, c'est-à-dire les mécanismes élémentaires du changement lexical (Sablayrolles, 2000 ; Sablayrolles et Pruvost, 2016). Sablayrolles propose une typologie des procédés de formation des néologismes selon différents critères (tableau 1).

Tableau 1 - Matrices lexicogéniques (Sablayrolles et Pruvost, 2016)

M A T R I C E S I N T E R N E S	morpho- sémantiques	construction	Affixation	préfixation	<i>détatouer</i>
				suffixation	<i>statuesque</i>
				dérivation inverse	<i>Turbuler, prester</i>
				parasynthétique ?	<i>désidéologisé ?</i>
		Composition	flexion	<i>ils closirent, la représsaille</i>	
			composition	<i>voiture-bélier</i>	
			synapsie	<i>lanceur d'alerte</i>	
			composition savante	<i>batracianophile</i>	
			hybride	<i>e-commerce, aquacinéaste</i>	
	Composition par amalgame	fracto-composition	<i>téléspectateur</i>		
		compocaption	<i>mobinaute, dircab</i>		
		factorisation	<i>optimessimiste</i>		
		mot valise	<i>peopolitique</i>		
	syntactico- sémantiques	imitation et déformation	changement de fonction	onomatopée	<i>dzoing</i>
				fausse coupe ou paronymie	<i>la nesthésie, infractus,</i>
				conversion	<i>la glisse, la gagne</i>
				conversion verticale	<i>de rejuvénation, un ex</i>
				déflexivation	<i>le boire, le manger</i>
changement de sens		combinatoire syntaxique / lexicale	<i>ironiser un texte encourir la liberté</i>		
		métaphore	<i>souris (inform.)</i>		
		métonymie	<i>sac à dos 'touriste'</i>		
		autres figures	<i>escorteuse 'call girl'</i>		
morpho- logiques	réduction de la forme	troncation	<i>blème, petit déj</i>		
		siglaison /acronyme	<i>LMD, ECUE</i>		
phraséologique	pragmatico-sémantique	détournement	<i>être les dindons de la crise, faire marcher la planche à promesses</i>		
		création	<i>ne pas faire du huit megabits</i>		
MATRICE EXTERNE			emprunt	<i>break, cool fioul, redingote</i>	

S'inspirant des travaux de Tournier (Tournier, 1985; 1991), Sablayrolles distingue tout d'abord les matrices internes à la langue et la matrice externe, qui permet de rendre compte des phénomènes d'emprunts. Dans le premier groupe, il distingue :

- les mécanismes morphologiques : ils emportent une modification morphologique des lexies, sans s'accompagner d'aucune modification de sens, avec deux cas principaux, les troncations et les siglaisons;
- les mécanismes morpho-sémantiques : ils combinent une modification morphologique et une création de sens. Il en existe deux types principaux : par construction, d'une part, comprenant l'affixation et la composition, les deux mécanismes traditionnellement décrits; par imitation et déformation, opérant principalement sur des éléments phonologiques;
- les mécanismes phraséologiques : ils opèrent au niveau d'une séquence de lexies, emportent également une création de sens, et comprennent deux sous-types, la création et le détournement;
- les mécanismes syntactico-sémantiques : ils opèrent une modification au niveau syntactico-sémantique, et s'accompagnent soit d'un changement de fonction soit d'un changement de sens.

Cette typologie catégorise les mécanismes élémentaires, mais un néologisme peut être le résultat de plusieurs opérations successives : certains néologismes ne font appel qu'à un seul procédé (*statuesque* par exemple pour la suffixation sur un radical simple, ou bien *binge-drinking* pour les emprunts), d'autres sont 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).

1.2. Approches distributionnelles

Le distributionnalisme harrisien a développé une méthode de description des langues ne faisant appel ni au sens, ni à l'histoire, mais faisant reposer toute l'analyse sur le seul matériau qui nous soit accessible, les productions linguistiques. La linguistique de corpus a repris cette approche en proposant de quantifier les descriptions linguistiques, tout d'abord en comptant les occurrences des unités rencontrées dans les textes.

Selon la fameuse formule de (Firth, 1957:11), « You shall know a word by the company it keeps ». Depuis les années 90, la linguistique de corpus a mis au jour des phénomènes linguistiques, au travers des notions de collocations (Firth, 1957), de « multiword expression » (Constant *et al.*, 2017), de *collostructions* (Stefanowitsch et Gries, 2003), de *collocational profile* (Sinclair, 1991), de *profil combinatoire* (Blumenthal, 2005) ou *behavioral profile* (Gries, 2010). Ces trois dernières notions donnent une base à l'étude automatique des innovations sémantiques, en proposant de les décrire sur la base d'une évolution de leur usage-sens entendu comme une combinatoire d'emplois prototypiques. Par exemple, en comparant les emplois de *tsunami*, entre les années 1900 et les années 2000², on constate plusieurs évolutions : au départ, le nom emprunté (de *tsu*, vague, et *nami*, de port) se combine exclusivement avec des adjectifs dénotant une propriété de l'événement concret (*catastrophique, soudain, puissant...*), avec, d'ailleurs, un sens restreint au phénomène tel qu'il se manifeste en Asie du sud-est (*Le tsunami, le raz-de-marée des japonais, Le Petit Caporal*, 07/01/1909). A partir des années 1970, on voit apparaître de nouvelles combinaisons, qui dénotent un sens figuré (*un tsunami social, architectural, etc.*), puis, après le tsunami de 2004, qui va permettre au terme de diffuser dans la plupart des langues, une troisième combinatoire apparaît (*un tsunami d'applaudissements, de libido, etc.*) transformant le nom en déterminant complexe conservant les traits du sens concret initial (*soudain, violent, massif*). Les néologismes sémantiques peuvent donc être détectés sur des bases formelles, non pas au niveau de leur forme interne, mais au niveau de la combinatoire qui les caractérise.

De façon complémentaire, Harris évoque très tôt une seconde façon d'exploiter la distribution des lexies :

« ...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, 1954:43)

À partir de cette intuition, un nouveau champ, la *sémantique distributionnelle* (Baroni et Lenci, 2010) verra le jour, ainsi que des applications pratiques (Mikolov *et al.* 2013), permettant, à partir de larges corpus, d'identifier les lexies en relation de similarité sémantique, et donc, en diachronie, d'étudier l'évolution de cette signature sémantique. Par exemple, (Hamilton, 2016) montre comment l'adjectif *gay* est sémantiquement similaire à *daft, tasteful, sweet, pleasant*, dans les années 1900, puis sémantiquement similaire à *bisexual, homosexual, lesbian* à partir des années 1990. Cette approche, si l'on dispose de corpus conséquents diachroniques, est complémentaire de la précédente, permettant de détecter, sur la base de contextes partagés, les lexies en relation de similarité sémantique, c'est-à-dire, notamment, en relation de synonymie, d'hyponymie et d'hyperonymie, ce qui permet de suivre le sens de ces lexies.

1.3. Apport de la linguistique cognitive

L'approche distributionnelle trouve une assise théorique et cognitive nouvelle avec certains travaux de la sémantique cognitive (Langacker, 1987 et 1991 ; Schmid, 2015). Ceux-ci ont explicité la notion d'*entrenchment* (*enracinement*), qui fonde le processus d'ancrage socio-cognitif des signes linguistiques, en corrélant ce processus avec la répétition des occurrences. Ces approches fondent une 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 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.

1.4. Grammaires de construction

Les grammaires de construction (voir (Goldberg, 2013) pour une présentation des principales variantes) considèrent également, généralement, que les corpus sont la matière principale de toute étude linguistique et que le calcul des répétitions des formes-sens en corpus est une information essentielle pour suivre la vie des lexies.

Plusieurs auteurs (Bybee, 2010 ; Bybee, 2016 ; Traugott et Trousdale, 2013) considèrent que le cycle de vie des lexies comprend trois phases saillantes (émergence, diffusion, conventionnalisation) : l'émergence correspond à l'apparition de la construction nouvelle (nouvelle forme et/ou nouvel usage) ; la diffusion correspond à la répétition de la construction nouvelle au-delà du cercle socio-démographique initial ; la conventionnalisation (ou encore lexicalisation, ou stabilisation) correspond à la phase de l'adoption par la communauté linguistique de la nouvelle construction. 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.

Signalons également la théorie de l'évolution lexicale proposée par (Koch, 2000 ; Gévaudan et Koch, 2010), 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 comme innovation lexicale) peut être décrite au moyen de trois paramètres : deux qui sont translangues, le paramètre sémantique (avec plusieurs opérations qui s'appliquent temporellement : 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 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 peut être croisée avec l'approche plus détaillée liée aux matrices lexicogéniques.

Les travaux de la linguistique cognitive et des grammaires de construction permettent de revisiter les dichotomies structuralistes langue/discours et synchronie/diachronie. La langue représente la "mémoire collective" des constructions (et de leur organisation) d'une communauté linguistique donnée. Cette mémoire collective permet l'intercompréhension entre locuteurs compétents d'une langue, et rend possibles les discours, qui exploitent cette mémoire partagée. Le phénomène de stabilisation des constructions est quant à lui lié à l'*entrenchment*, qui lui-même se base principalement sur la répétition en discours des constructions. D'un autre côté, tout discours est une innovation intrinsèque, puisque lié à un contexte nouveau, et alimente la mémoire partagée tout en la modifiant par l'énoncé de combinaisons nouvelles. Langue et discours, conçus par la linguistique structurale comme deux objets distincts et ordonnés, sont donc en réalité deux faces d'une même réalité. Cette

conception intégrative oblige également à revisiter l'opposition classique synchronie (où la langue peut être décrite par saisie d'un état) / diachronie (où intervient le discours, comme événement). En réalité, la synchronie est une abstraction, seuls existent les événements discursifs *et* leur mémorisation dans l'esprit des locuteurs. La langue est donc intrinsèquement évolutive. Dans ce cadre, le phénomène néologique est partie intégrante du système linguistique, chaque construction (ou lexie, en termes plus classiques) évoluant au cours du temps. Ces évolutions peuvent être suivies par le biais des propriétés de ces constructions : fréquence, modification des propriétés morphologiques, combinatoires et distributionnelles.

1.5. Langue standard et variations linguistiques

L'étude des évolutions lexicales doit également être mise en regard des phénomènes de variation linguistique. La linguistique variationniste a en effet mis au jour l'évidence des variations d'emploi et l'idéalisation fautive que constitue la notion de langue unique. En réalité, chaque locuteur a la compétence de plusieurs strates de connaissances linguistiques, et plusieurs variétés cohabitent au sein d'une communauté linguistique. Nous devons à (Cosieriu, 1964) les premières intuitions sur les paramètres permettant d'identifier ces strates de savoirs linguistiques : variation diatopique (variation dans l'espace : français des régions, français des pays francophones, etc.), variation diastratique (variation sociale : liée à l'existence de sous-groupes aux usages linguistiques différenciés), variation diaphasique (variation dans la situation communicative, registres et styles). (Koch et Oesterreicher 1985) introduiront un autre axe, l'axe de la *distance communicative* permettant de rendre compte, sur un continuum, des différentes situations, de l'immédiat (l'oral) à la distance (l'écrit). Une modélisation adéquate des variétés est capitale pour caractériser et suivre le cycle de vie des innovations lexicales.

Ces différents éléments théoriques fournissent les fondements sur lesquels l'architecture de la plateforme Néoveille a été conçue et développée, et sur lesquels les descriptions linguistiques ont pu être menées.

2. Plateforme Néoveille de repérage, d'analyse et de suivi des néologismes en sept langues

La plateforme est le résultat d'un projet collaboratif mené avec plusieurs partenaires français et internationaux. Le projet vise notamment à :

- mettre en place une plateforme multilingue de détection et de suivi des néologismes à partir de corpus contemporains dynamiques de très grande taille, dans sept langues (français, chinois, grec, polonais, portugais du Brésil, russe et tchèque), d'autres langues ayant été ensuite ajoutées (italien, allemand, néerlandais);
- mettre en œuvre des algorithmes et programmes pour détecter automatiquement les néologismes de forme;
- utiliser cette plateforme pour étudier la notion d'innovation sémantique et pour proposer de nouvelles procédures d'identification des nouveaux emplois.

La plateforme est aujourd'hui accessible et opérationnelle, comprenant une partie publique et une partie privée pour l'édition des données.

2.1. Architecture et modules de la plateforme

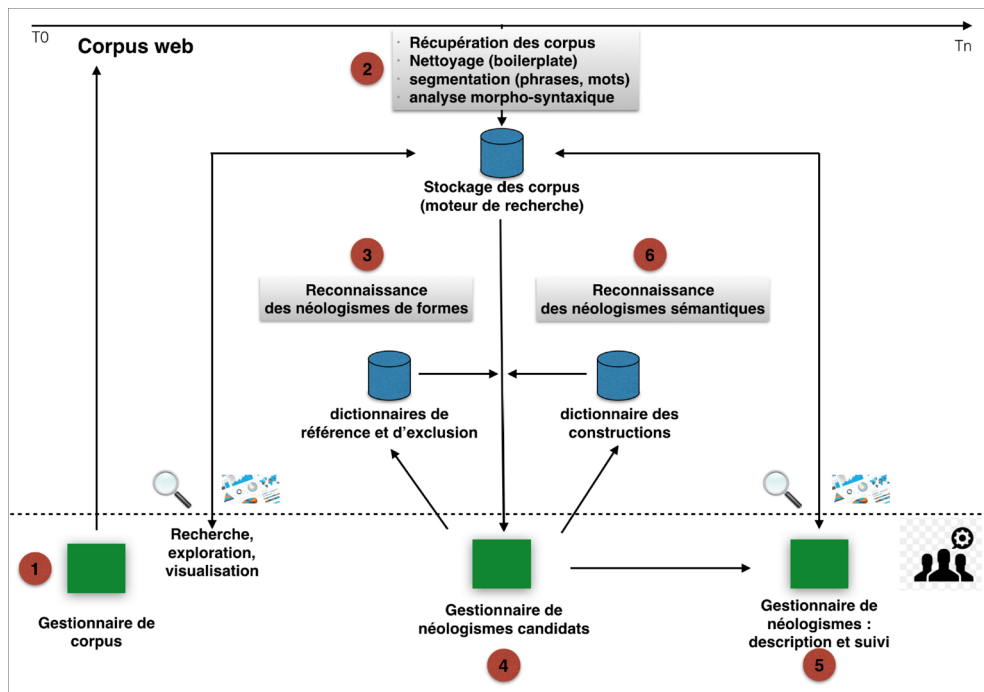


Figure 1 - Architecture générale de la plateforme

Dans cette architecture, le trait horizontal en pointillé sépare les composants où l'expert linguiste intervient (partie basse) des composants liés à un processus automatique. Cette architecture reproduit le flux langue-discours, en proposant d'une part l'alimentation continue du système en discours (corpus web), et des processus de traitements (automatiques et manuels) aboutissant à créer une mémoire linguistique, sous forme de dictionnaires (carrés verts) et d'un espace de stockage des corpus (moteur de recherche). On distingue six modules (pastilles numérotées) :

Le gestionnaire de corpus (1) : l'expert linguiste peut ajouter, supprimer, modifier les sources d'information qu'il souhaite étudier, soit des fils RSS, soit des sites web. Pour chaque source d'information, il explicite des méta-informations : nom du journal, url d'entrée, catégorie des informations fournies (presse générale, de vulgarisation ou spécialisée à l'heure actuelle), domaine (informatique, santé, économie, mode, etc.), langue, pays ou région du journal, fréquence de parution. Ces informations sont associées à chaque unité d'information (« article ») qui sera récupérée et permettent de filtrer les résultats néologiques, notamment pour *approcher*³ les notions de diastrie et de diatopie.

La récupération régulière des fils RSS, des articles liés, leur analyse linguistique et leur stockage dans un moteur de recherche (2) : ce module récupère deux fois par jour de nouveaux articles publiés dans les fils RSS et les pages web et effectue différents traitements : segmentation en mots puis analyse morphosyntaxique⁴. Ce module permet d'ajouter pour chaque article titre, contenu textuel, contenu étiqueté morphosyntaxiquement, lemmes (noms, verbes et adjectifs) et noms propres du document. Le moteur de recherche sert de base de stockage des corpus et peut être interrogé de manière riche, avec plusieurs modules de visualisation interactive liés à l'évolution de fréquence et aux paramètres diatopiques et diastratiques disponibles.

Le repérage automatique des néologismes de forme et leur stockage dans la base des néologismes candidats (3, 4): ce module permet de détecter dans les articles de presse stockés dans le moteur de recherche des candidats néologismes après application de la méthode dite « dictionnaire de référence ou d'exclusion » (MDE)⁵ : on détecte des formes nouvelles sur la base d'un dictionnaire des formes attestées dans une période antérieure. L'approche suivie ici est une version étendue de cette méthode : plusieurs filtres sont tout d'abord appliqués : dictionnaires de référence et d'exclusion, noms propres, erreurs typographiques. Les candidats néologismes sont ensuite présentés aux experts, qui valident ou invalident la reconnaissance automatique. Ce module permet un *apprentissage itératif*, puisque les décisions des experts sont ensuite réutilisées par le système automatique par réinjection dans les dictionnaires des néologismes non validés.

Le gestionnaire de néologismes (5) est une base de données permettant de décrire linguistiquement les néologismes validés, et d'obtenir un suivi de leur cycle de vie par suivi des occurrences dans le corpus dynamique. (il s'agit d'une version simplifiée de la base de données Neologia, voir Cartier et Sablayrolles, 2008).

Le repérage des néologismes sémantiques (6) comprend deux étapes : une étape de reconnaissance automatique des néologismes sémantiques, et une étape de validation par les experts. Il est ensuite possible de décrire linguistiquement les néologismes validés, et de suivre leur cycle de vie. Cette fonctionnalité est encore expérimentale.

Nous présentons maintenant le corpus constitué, ainsi que la méthodologie utilisée pour valider les néologismes.

2.2. Présentation du corpus

Le corpus dynamique du français contemporain est composé de 242 sources de presse récupérées deux fois par jour sur internet. Chaque source d'information comprend des informations liées à la diatopie (pays, région pour l'espace hexagonal) et à la diastratie (domaine). Un aperçu est présenté dans le tableau 2.

Tableau 2. Extraits sur la répartition par pays et par domaine (pour la presse hexagonale).

Pays	Nbre titres de presse	Nbre d'articles
France	156	912504
Algérie	53	75291
Canada	4	72693
Belgique	4	39503
Sénégal	25	39296
Total	242	1 139 287

Domaine	Nbre d'articles
Général	730386
Sport	99802
Presse féminine	27108
Politique	13945
Informatique	11451
Industrie	9527
Sciences	4516

Les différences dans la distribution par pays tiennent au nombre de titres de presse, d'une part, et au démarrage de la récupération, plus tardive (début 2016) pour les pays francophones hors France. Concernant les domaines, la liste est liée, en la simplifiant, à l'ontologie des thèmes proposés par l'IPTC⁶.

2.3. Processus de validation des néologismes détectés automatiquement

Comme indiqué plus haut, les néologismes sont d'abord détectés automatiquement par la méthode MDE. En moyenne, pour le français, entre 100 et 200 néologismes candidats (NC) sont repérés chaque jour. Puis les experts linguistes, sur la plateforme, doivent assigner à chacun des NC, soit la catégorie « néologisme », soit la catégorie « non-néologisme ». Le

système prévoit, pour la première catégorie, d'assigner directement l'un des mécanismes détaillés en section 1.1, soit, pour la seconde, un type particulier de non-néologisme (erreurs typographiques, autres erreurs, généralement des mots étrangers dans une citation ; gentilé, particularisme⁷, xénisme⁸, dictionnaire de référence (mot simple ou composé), dictionnaire terminologique).

Le processus de validation des néologismes suit le protocole suivant : chaque membre du groupe de travail⁹ annoté sur la plateforme une partie des néologismes candidats, sur la base d'une fiche d'instructions détaillant les catégories de néologismes et de non-néologismes. Puis, lors de réunions collectives mensuelles, une validation est effectuée, les cas litigieux étant tranchés sur la base d'un vote majoritaire. Ces discussions collectives ont permis un certain nombre d'aménagements des catégories existantes (notamment parmi les non-néologismes : xénismes, particularismes, gentilés). Ce processus de validation a également permis de vérifier le taux de précision du repérage automatique, qui est proche de 60 % pour le français. Au final, pour le français, nous avons ainsi pu valider, sur deux ans et six mois, un peu plus de 21 000 néologismes.

3. Tendances néologiques du français contemporain (2015-2017)

Nous analysons maintenant les résultats du travail collaboratif de validation et de description des néologismes, de 2015 à fin 2017. Nous présentons tout d'abord les tendances générales identifiées, puis détaillons les résultats par type de mécanismes néologiques. Un rapport détaillé comprenant l'ensemble des données, sera disponible sur la plateforme du projet.

3.1. Tendances générales du français contemporain

De juillet 2015 à décembre 2017, à partir d'environ 250 sources d'informations, 1 143 912 articles pour un total de plus de 92 millions de mots (1 037 876 formes différentes) ont été récupérés. Parmi environ 35 000 néologismes formels candidats, 22 475 néologismes¹⁰ ont été validés, correspondant à 726 222 occurrences. Les néologismes représentent donc 2,16 % des formes rencontrées, et, au niveau du nombre d'occurrences, 0,78 %.

3.1.1. Répartition par mécanisme

Dans le tableau 3, qui donne la répartition par matrices, les colonnes 2 et 3 indiquent le nombre de néologismes différents, les colonnes 4 et 5 le nombre d'occurrences, et la colonne 6 le nombre moyen d'occurrences par matrice.

Tableau 3. Synthèse sur les mécanismes néologiques.

Mécanisme néologique principal	Nombre de néologismes (formes uniques)		Nombre d'occurrences de néologismes		Moyenne d'occ. par forme néologique
	Nb	%	Nb	%	
préfixation	17 051	75,87%	485 566	66,86%	28
composition	1 646	7,32%	31 173	4,29%	19
emprunt	1 429	6,36%	132 104	18,19%	92
suffixation	1 245	5,54%	65 262	8,99%	52

fracto-composition	791	3,52%	7 039	0,97%	9
onomatopée	92	0,41%	665	0,09%	7
troncation	73	0,32%	2 678	0,37%	37
composition savante	68	0,30%	479	0,07%	7
compoaction	47	0,21%	1 043	0,14%	22
composition hybride	33	0,15%	213	0,03%	6
mot-valise	9	0,04%	100	0,01%	11
Totaux	22 475	100,00%	726 222	100,00%	

On constate que :

- le procédé le plus utilisé est la préfixation (75 % des formes néologiques). La composition, les emprunts, la suffixation et la fracto-composition représentent entre 3 et 7 % du contingent. Les autres mécanismes sont quantité négligeable.
- le nombre d'occurrences révèle un classement légèrement différent : en dehors des emprunts (de 6 % à 18 %, 92 occurrences en moyenne), de la suffixation (de 5 % à 9 %, 52 occurrences en moyenne), les autres procédés ont un nombre d'occurrences faible, en moyenne. Nous tenterons d'analyser ces faits dans la section 3.1.4. et les sections spécifiques ;
- en l'état actuel, nous n'avons analysé que le mécanisme principal, il serait pertinent d'obtenir des statistiques en tenant compte également des néologismes issus de plusieurs mécanismes (par exemple, *démacroniser* est classifié comme préfixation, mais la base est elle-même le résultat très récent d'une suffixation).

3.1.2. Répartition par journaux, domaine, pays

Les chiffres précédents doivent être ramenés aux paramètres diatopiques et diastratiques disponibles sur la plateforme. On constate alors que :

- Les **journaux** les plus productifs de néologismes sont : *L'Express* (46 889 occurrences, 6,45 %), *Libération* (28 551, 3,93 %), *France Soir* (27 828), *Le Huffington Post* (26 237) et *le Monde* (25 701). Du point de vue des innovateurs, en se basant sur les premières occurrences, les répartitions sont quasiment les mêmes, avec cependant l'apparition dans les premiers rangs du *Parisien*, de la *Voix du Nord* et la *Nouvelle République*.
- La **répartition par domaine** révèle la prédominance de trois domaines¹¹ : le sport (10 %), la presse féminine (10 %) et l'informatique (5 %). Étant donné la distribution de notre corpus, correspondant grosso modo à cette répartition, il est difficile d'en tirer une quelconque conclusion. Cependant, les domaines informatique et de la presse féminine ont une autre particularité, la grande productivité néologique en termes d'emprunts, qui représentent près de 50 % des innovations, dans l'un comme dans l'autre cas.
- La **répartition géographique** révèle une distribution des néologismes équivalente à la distribution des corpus, avec une prédominance du français métropolitain (83 % des occurrences de néologismes), sans modification de la distribution des mécanismes selon les pays d'origine.

3.1.3. Répartition par parties du discours

La distribution par parties du discours est la suivante : Nom (79,61 %), Adjectifs (9,76 %), Verbe (8,34 %) et Adverbe (2,29 %). Une comparaison avec la distribution constatée sur le

corpus complet¹² montre que la néologie s'applique principalement sur les noms et les adjectifs, les verbes et les adverbes étant sous-représentés.

3.1.4. Hapax et cycle de vie des néologismes

Comme nous l'avons vu, la moyenne d'occurrences est relativement faible par néologisme. La déviation standard¹³ est importante (111 par forme néologique, 237 par occurrence totale), montrant qu'il existe quelques néologismes qui sont employés de façon massive dès leur apparition (notamment toutes les innovations liées à une actualité : *loi-travail*, *nuit-debout*, *penelopegate*, *cyberattaquant*, *street(-)wear*, *street(-)art*, etc.). Si nous utilisons la médiane, le nombre d'occurrences tombe à 4 : une très grande majorité de néologismes sont donc principalement des *hapax* ou des *quasi-hapax*. Le tableau 3 détaille ces répartitions (colonne 1 : nombre d'occurrences entre deux seuils, seuil supérieur non inclus ; colonnes 2 et 3 : total d'occurrences des néologismes ; colonnes 4 et 5 : total de documents couverts).

Tableau 3. Synthèse sur les répétitions des néologismes.

Nb d'occurrences	Nbre d'occurrences total	%	Nbre de documents différents	%
[1-5]	12382	55,22%	13736	61,26%
[5, 10]	2759	12,30%	2532	11,29%
[10, 50]	4302	19,19%	3794	16,92%
[50, 100]	1099	4,90%	906	4,04%
[100, 1000]	1612	7,19%	1454	6,48%
[1000,]	268	1,20%	0	0,00%

On constate que :

- près de 85 % des néologismes sont des hapax ou quasi-hapax, en considérant qu'une répétition jusqu'à 50 occurrences fait encore partie de l'hapax¹⁴ ;
- la distribution est encore plus nette en considérant une répétition basée sur le document, et non sur le nombre d'occurrences totales, qui recouvre des répétitions au sein du même document ;
- pour les occurrences supérieures à 50 occurrences, il est difficile de se prononcer sur le sort à venir des néologismes en cause, étant donné la faible amplitude temporelle du corpus.

Le cycle de vie des néologismes ne se résume pas à l'évolution fréquentielle des occurrences. Au moins deux autres éléments permettent d'identifier une diffusion accrue : d'une part, la diffusion de l'innovation d'un domaine à l'autre ; d'autre part, notamment pour ce qui concerne les emprunts, l'intégration morphologique flexionnelle et dans le système de morphologie productive. Dans le cadre du présent projet, nous n'avons pas encore suffisamment d'empan temporel pour effectuer cette étude.

3.2 Dérivation / Affixation

3.2.1 Définitions

La dérivation regroupe tous les phénomènes d'affixation : préfixation, suffixation et parasyntèse, résultat de l'adjonction simultanée d'un préfixe et d'un suffixe à une base¹⁵. La préfixation a lieu lorsqu'un affixe est ajouté devant une base, simple ou non. Pour catégoriser un néologisme comme créé par préfixation, nous sommes partis d'une liste de 59 formants¹⁶.

La suffixation est le second procédé de dérivation. Elle consiste à ajouter un affixe à une base. Pour étudier la suffixation, nous sommes partis de la liste de suffixes établie par *Le Petit Robert*¹⁷. Ce procédé induit généralement un processus de transcatégorisation (Sablayrolles, 2000 : 264-265).

3.2.2. Préfixation

Le tableau 4 explicite les 41 préfixes productifs¹⁸ et le nombre de néologismes constatés, dans notre corpus.

Tableau 4. Liste des préfixes productifs, par ordre décroissant.

1	<i>anti</i>	1222	16	<i>re/ré</i>	108	31	<i>archi</i>	12
2	<i>ex</i>	1008	17	<i>super</i>	97	32	<i>méga</i>	11
3	<i>non</i>	696	18	<i>co</i>	88	33	<i>pluri</i>	11
4	<i>mini</i>	611	19	<i>pré</i>	72	34	<i>maxi</i>	10
5	<i>ultra</i>	482	20	<i>extra</i>	71	35	<i>hors</i>	9
6	<i>mi</i>	377	21	<i>tout</i>	68	36	<i>in</i>	8
7	<i>post</i>	343	22	<i>micro</i>	65	37	<i>après</i>	6
8	<i>hyper</i>	284	23	<i>sur</i>	63	38	<i>intra</i>	6
9	<i>auto</i>	258	24	<i>contre</i>	51	39	<i>avant</i>	6
10	<i>demi</i>	255	25	<i>inter</i>	46	40	<i>sans</i>	5
11	<i>sous</i>	209	26	<i>pseudo</i>	29	41	<i>infra</i>	5
12	<i>semi</i>	198	27	<i>mono</i>	21	42	<i>poly</i>	5
13	<i>quasi</i>	177	28	<i>bi</i>	18			
14	<i>pro</i>	127	29	<i>néo</i>	14			
15	<i>multi</i>	119	30	<i>dé</i>	13			

Dans ce tableau, 18 préfixes sont absents (*a/-an-*, *ab-abs-*, *a(d)-*, *ambi-*, *ana-*, *apo-*, *cata-*, *circo-/circum-*, *dia-*, *dis-*, *dys-*, *ecto-*, *endo-*, *épi-*, *eu-/ev-*, *juxta-*, *pén(é)-*, *per-*), devenus non-productifs ou en tout cas non attestés dans les corpus. Les autres préfixes sont déjà mentionnés par les études antérieures (Dubois, 1962 ; Corbin, 1987). On peut rapporter l'ordre des préfixes à leur *productivité potentielle*, calculée selon leur capacité à s'appliquer à plusieurs catégories. Les préfixes s'appliquant aux verbes sont bien moins représentés dans cette liste (*post*, *auto*, *sous*, *re*, *co*, *pré*, *sur*, *contre*, etc.) que les préfixes produisant des noms, adjectifs et adverbes mais ce phénomène est lié à l'innovation lexicale produisant principalement noms et adjectifs.

Concernant la préfixation par superlatif intensif, les données recueillies dans le projet confirment l'implantation (Corbin, 1987) de l'intensifieur *ultra-* (482 néologismes), avec une application majoritairement aux adjectifs (468) mais également aux noms (14 occurrences : *ultra-bike*, *ultra-centre*, *ultra-communication*, *ultra-connexion*, etc.). Les autres intensifieurs (*hyper*, *super*, *méga*, *extra*, et dans une moindre mesure *archi* et *méga*, qui ne s'appliquent respectivement qu'à des adjectifs et des noms, et qui sont plus connotés), se partagent le reste du champ.

On notera enfin la forte productivité de *mini*, dont l'emploi a explosé à partir de la création de *mini(-jupe)*, dans les années 1970 (Corbin, 1987), avec une application aux noms et aux

adjectifs, *micro-* étant dorénavant d'un emploi plus restreint (*micro-déchet, micro-entrepreneur, etc.*).

3.2.3. Suffixation

Nous présentons dans le tableau 5 les 20 suffixes les plus productifs sur la période.

Tableau 5. Liste des suffixes productifs, par ordre décroissant.

Suffixe	Nbre de réalisations	Type transcatégorisation	Exemples
is(er)	243	N => V	instagramiser, tiersmondiser, facebookiser, googleliser...
> isation	26	N => N	routinisation, gangsterisation, premiumisation, dronisation...
> isateur/trice	22	N => ADJ/N	vampirisateur, socialisateur, commercialisateur...
isme	213	N => N	clintonisme, validisme, montebourgeoisme
ien(ne)(s)	109	N => N/ADJ	daeshien, gorafien, macronien, trumpien, facebookiennes
iste	94	N => N/ADJ	lemairiste, juppeistes, ségoléniste, laisser-fairiste...
eur(euse)(s)	54	N => N	snapchateur, zeuzeuteurs, shoppeuse,
itude	47	N => N	basiquitude, modernitude, djeunitude, cancritude...
esque	46	N => ADJ	hanounesque, internetesque, googlesque, uluberluesque...
able	34	N => ADJ	costumisable, pocketables, twittable, shoppable
ade	17	N => N	macronade, estofinade, pétrolade, ruquiade
ette	13	N => N	berlinette, sarkozette, balladurette, trumpinette
erie	11	N/ADJ => N	cheaperie, kitscherie, leperseries, hollanderie, merguezerie
age	11	N => N	spoilage, squelettage, décolettage, youtubage...
issime	9	ADJ => ADJ	glamourissime, horriblissime, punkissime, macronissime
ité	8	N => N	auctorialité, macronité, guerriérité, catalinité, odieusité
ite	7	N => N	ibrahimovite, comparaisonnite, fillonite, luddite, coudinite
ique(s)	7	N => ADJ	bla-bla-tiques, rugbistiques, guitaristiques, autonomiques
ie	7	N => N	fillonie, trumpie, numératie, mummyrexie
el(le)(s)	6	N => ADJ	interconvictionnel, réputationnel, mictionnels

La triple suffixation en *-iser, -ation et -isateur/trice* constitue la famille suffixale la plus productive dans le corpus journalistique, dénotant diverses transformations sociétales (*twitterisation, macronisation...*).

Plusieurs autres suffixes (*-isme, -itude, -ité, -ie*) permettent également de créer des noms abstraits généralement à partir d'une base nom propre (*macronisme, macronitude, macronité*).

Plus classiques sont les formations en *-ien, -iste, -eur* pour former les agents à partir d'une base dénotant une activité ou un ensemble de positionnements liés à une personnalité publique (*macroneur/iste/ien*).

A noter enfin qu'en dehors de *-iser*, aucun suffixe ne permet de créer des verbes, *-er* étant évidemment le seul morphème productif.

3.3. Composition

3.3.1. Définitions

Nous considérons (voir section 1.1.) quatre sous-classes de composition entre lexies : la composition « simple », la synapsie (ou locution figée), la composition savante (composée

de formants savants : *batracianophile*) et hybride (composée d'un formant savant et d'une lexie : *e-commerce*, *aquacinéaste*). Parmi les compositions par amalgame, nous distinguons la compocation (troncation + concaténation, terme forgé par (Cusin-Berche, 1999), *hélicoptère* : *hélicoptère*>*héli* et *aéroport*>*port*), la fracto-composition¹⁹ (combinaison de deux lexies dont la première est tronquée, exemple *téléspectateur*), la mot-valisation (fusion de deux lexies simples sur la base d'une homophonie à la frontière des deux lexies, *gangsterrorisme*) et la factorisation (factorisation d'un élément phonique commun mais sans superposition syllabique : *optimessimisme*).

3.3.2. Composition simple

1 646 néologismes ont été validés, dont 642 avec un seul emploi, ce qui révèle que ces néologismes répondent souvent à un besoin ponctuel. Les néologismes résultants sont des noms (87%) ou des adjectifs (13%).

Le schéma syntaxique le plus fréquent reste Nom-Nom, dénotant des combinaisons variées (*scientifique-justicier*, *journaliste-confesseur*, *poisson-sanglier*, *veste-électrocardiogramme*, *bus-cuisine*, etc.). D'autres schémas, plus rares, sont utilisés : adjectif-adjectif (*footballistico-temporelle*), adverbe-nom (*déjà-amoureux*), nom-adjectif (*budget-compatible*), verbe-nom (*savoir-fashion*), verbe-verbe (*savoir-être*), verbe-adverbe (*vivre-ensemble*).

Les données recueillies confirment la productivité des schémas de construction anciens : *Nom-clé* (141 occurrences, *réforme-clé*, *scrutin-clé*), *Nom-phare* (91, *smartphone-phare*), *Nom-surprise* (68, *limogeage-surprise*), *Nom-choc* (56, *accessoire-choc*), *Nom-culte* (*réclame-culte*), *Nom-éclair* (*casse-éclair*). De nouveaux patrons apparaissent : *robot-N* (56 occurrences : *robot-coiffeur*, *robot-voiturier*, *robot-pompier*, *robot-vendeur*, *robot-livreur*, *robot-cuisinier*) ; *N-compatible* (*jihad-compatible*) et *N-réalité* (*youtube-réalité*).

Parmi les rares synapsies, de nouveaux schémas apparaissent. Notamment *prêt-à-Nom*, dû à l'ancien *prêt-à-porter*, qui est à l'origine d'un paradigme : *prêt-à-pousser*, *prêt-à-consommer*, *prêt-à-cuire*, *prêt-à-nager*, *prêt-à-gober*, *prêt-à-licker*, *prêt-à-agir*, etc.

3.3.3. Fracto-composition

Cyber-, *bio-*, *éco-* et *e-* sont les formants les plus productifs. Ils représentent respectivement les lexies *cybernétique*, *biologique*, *écologique* et *électronique*. La forte productivité de ces formants reflète les changements sociétaux actuels.

Tableau 6. Synthèse sur *cyber-*, *e-*, *bio-* et *éco-*

	cyber-	e-	bio-	éco-
Nb	92	60	51	19
Exemples	<i>cybercondriaque</i> , <i>cyberathlète</i> , <i>cyberattaquer</i>	<i>e-citoyenneté</i> , <i>e-enseignant</i> , <i>e-recruter</i>	<i>bio-exorciste</i> , <i>affinité</i> , <i>bio-diversifier</i>	<i>éco-jardin</i> , <i>éco-touristique</i>

Le fracto-lexème entre en composition principalement avec des noms et des adjectifs, et beaucoup plus rarement avec un verbe (*cyber-menacer*, *e-recruter*, *bio-diversifier*, etc.).

3.3.4. Compocation et mot-valise

Très peu de compocations et de mots-valises ont été détectés. Notons la compocation *trotscoot*, où la première syllabe de *trottinette* et celle de *scooter* ont été gardées et

assemblées. Mais pour *blogistador* (construit à partir de *blog* ou *blogueur* et *conquistador*) ou *mammouphant* (de *mammouth* et *éléphant*) la fin du mot correspond à la fin de la seconde lexie. Parmi les mots-valises, notons *macronpatible*, où le chevauchement des lexies est réalisé au niveau du segment phonique qu'elles partagent, le son [ɔ]. Ce même mécanisme opère dans *hollandouille* (sur *Hollande* et *andouille*) et *peopopulaire* (sur *people* et *populaire*).

3.4. Emprunts

Les emprunts représentent 1 430 formes (6,36 % du total) pour 132 104 occurrences (18,19 %). Il faut y ajouter près de 1 000 xénismes non comptabilisés comme néologismes. La langue source la plus représentée est l'anglais²⁰ à environ 91 %, suivi de l'espagnol, de l'arabe et de l'italien. Les xénismes ont des langues-sources beaucoup plus diversifiées. On retrouve ici une différence fondamentale entre les emprunts anglais et les emprunts à d'autres langues : tandis que les seconds dénotent pour une très grande majorité des concepts culturels spécifiques et sont majoritairement monosémiques, les premiers sont plus enclins à la polysémie et renvoient à des aires culturelles variées (Chesley et Baayen, 2010). On peut y voir une différence dans la perception collective de l'anglais international, perçu comme une langue de prestige (et de nécessité, dans le contexte économique), tandis que les autres langues sont perçues comme des marqueurs d'identités. Ces résultats consolident les chiffres proposés par (Martinez, 2009) sur les provenances des emprunts enregistrés dans *Le Petit Robert* de 1997 à 2011.

Concernant la répartition par partie du discours, nous obtenons la répartition suivante : 83,8 % de noms, 9,7 % d'adjectifs et 6,5 % de verbes. De très nombreux emprunts nominaux sont également attestés en tant que verbes par conversion (*twitter*, *facebooker*, *shopper*...)

Trois domaines innovateurs sont particulièrement productifs : presse féminine (*Elle*, *Grazia*, *Cosmo*, *Styles*), informatique (*01Net*, *Le monde informatique*) et sport (*L'équipe*). Les emprunts à l'anglais ne se limitent plus au transfert de lexies. Du point de vue phonologique et orthographique, l'influence de l'anglo-américain est perceptible depuis longtemps (prononciation de *-ing*, *-ee-*, etc.). La pénétration est également visible par l'implantation d'affixes, notamment le fracto-lexème *e-* et le suffixe *-ing*. Dans notre corpus, ces formants ont une productivité importante : 86 lexies pour le premier (soit emprunts directs : *e-voting*, *e-shopping*, etc. soit hybrides : *e-défilé*, *e-vendeur*, *e-marché*, *e-citoyenneté*, etc.), 303 pour le second. Le morphème *-ing* a pénétré le français depuis plus d'un siècle (*parking*, *camping*, *pressing*, *meeting*, *dancing*, etc.), formant essentiellement des noms de lieux où une action se déroule, par métonymie du sens anglais. A partir des années 50, le morphème obtient le statut de quasi-suffixe, exprimant « une action, son résultat ou le lieu où se déroule cette action » (Dubois, 1962 : 14). (Mudrochova, 2017) étudie une trentaine de formes attestées dans *Le Petit Robert*, entre 1996 et 2002. Dans le présent projet, nous avons relevé 303 formes. La concurrence avec *-age* fait qu'il reste limité à l'expression de pratiques sportives (*running*, *beatboxing*, *snorkeling*, *cardiotraining*, etc.) professionnelles (*networking*, *packaging*, *branding*, *fact-checking*, *coworking*, *crowdfunding*,...) ou socio-culturelles (*bashing*, *ghosting*, *pet-sitting*) spécifiques, sans équivalents synthétiques en français.

Une dernière caractéristique des emprunts à l'anglais concerne l'émergence de patrons lexico-syntaxiques productifs : formations en *-gate* (56 occurrences : *dieselgate*, *couscousgate*, *penaltygate*, *penelopegate*, etc.) ; *street-* (25 lexies : *streetstyle*, *street-artiste*, etc.) ; *food-* (23 lexies : *food-truck*, *foodosphère*, *foodocratie*, *foodivores*, *street-fooders*, etc.) ; *-bashing* (11 lexies : *agribashing*, *sucre-bashing*, *macronbashing*, etc.), *-shaming* (14 lexies : *fatshaming*, *name-shaming*, *skillshaming*), *it-* (8 lexies : *it-jean*, *it-bag*, etc.). Nous relevons également 144 occurrences du patron N/ADJ-Ving (*car-jacking*, *home-staging*,

speed-dating, *speed-watching*, *binge-viewing*, *ride-sharing*). Cette forte productivité indique une implantation du schéma emprunté (Viaux et Cartier, 2018).

4. Conclusion et Perspectives

Nous avons présenté dans ce travail une plateforme de détection, d'analyse et de suivi des néologismes et les premiers résultats qui en découlent, du point de vue des néologismes formels. Après l'exposé de quelques jalons théoriques, et une brève présentation de la plateforme, nous avons exposé les grandes tendances néologiques du français contemporain, en nous basant sur un corpus de presse conséquent.

Plusieurs conclusions en découlent :

D'abord, du point de vue de la linguistique descriptive, les données néologiques collectées et validées permettent de se faire une idée relativement précise des mécanismes néologiques formels : dérivation, composition, emprunts, réductions morphologiques sont à l'œuvre en langue française, avec une forte dominance de la préfixation. Les emprunts se font essentiellement à l'anglais ; la suffixation, principal mécanisme de transcatégorisation, est à l'œuvre avec une liste finie de formants ; la composition donne lieu à l'apparition de nouveaux schémas productifs, notamment issus de l'anglais. Les données collectées et analysées seront mises à disposition de la communauté, de façon régulière sur le site internet du projet.

Du point de vue théorique, le travail applicatif amène à réviser et compléter le modèle initial, avec quelques notions clés : le *continuum* qui existe entre la fracto-composition et la préfixation met au jour les mécanismes de formation des morphèmes liés, qui, par *réduction* puis *productivité*, transforment certaines lexies en fractolexèmes puis en préfixes ; de même, au sein de la composition, la réduction est à l'œuvre dans les différents procédés d'amalgamation. La notion de productivité s'applique également à tous les procédés : affixation, mais également composition, avec des schémas productifs. L'étude permet aussi d'affiner la notion d'*hapax*, qui concerne près de 75% des néologismes, et qui doit être redéfinie par une combinaison de paramètres : fréquence, empan temporel, signature diatopique et diastratique. Les emprunts permettent enfin d'étudier les mécanismes d'assimilation des matériaux étrangers, à la fois d'un point de vue phonologique, mais aussi flexionnel, et concernant l'intégration à la morphologie productive.

Enfin, d'un point de vue méthodologique, le travail théorique et pratique permet de consolider l'approche initiale, à la fois concernant la détection, l'analyse linguistique et le suivi en corpus : si l'émergence des néologismes de forme est maintenant consolidée sur la plateforme, un travail est en cours pour détecter les changements lexicaux sémantiques et plus généralement étudier le cycle de vie des lexies, en se basant sur une combinaison de critères : évolution fréquentielle, changement diastratique et diatopique, évolution combinatoire et distributionnelle.

Crédits

Le projet Néoveille a été subventionné de 2015 à 2017 par une subvention IDEX « Initiatives d'excellence » (ANR-11-IDEX-0005-02).

Remerciements

Nous remercions chaleureusement les deux évaluateurs de la proposition initiale, dont les critiques, remarques et suggestions ont grandement permis d'aboutir à la présente version.

Références

- Baayen, R.H. (2009), « Corpus linguistics in morphology: morphological productivity », dans Lüdeling, Anke / Kytö, Merja, *Corpus linguistics. An international handbook*, p. 900-919.
- Baroni, M. et Lenci, A. (2010). « Distributional memory: A general framework for corpus-based semantics ». *Computational Linguistics*, 36(4), p. 673–721.
- Blumenthal, P. (2005). *Profil combinatoire des mots: analyse contrastive. La phraséologie dans tous ses états*, p. 131–148.
- Bréal, M. (1899). *Essai de sémantique* (2e éd.). Paris, Librairie Hachette et Cie.
- Bybee, J. (2010). *Language, usage and cognition*. Cambridge University Press.
- Bybee, J. (2016). *Language Change*. Cambridge University Press.
- Carnoy, A. J. (1927). *La science du mot. Traité de sémantique*. Louvain, Éditions « Universitas », VIII, 426 S.
- Cartier E. et Sablayrolles J.-F. (2008) « Néologismes, Dictionnaires et Informatique », *Les Cahiers de Lexicologie*, n° 93, 2008- 2, p. 175-192.
- Cartier E. (2016). « Neoveille, système de repérage et de suivi des néologismes en sept langues », *Neologica 10*, p. 101-131.
- Cartier, E. et Viaux, J. (2018). « Étude de la pénétration des anglicismes de type N ou ADJ(-)Ving à partir d'un corpus contemporain journalistique : les exemples de *bashing* et *shaming* en français », dans Jacquet-Pfau C., Napieralski A. et Sablayrolles J.-F., *Emprunts et équivalents autochtones : études interlangues, Folia Litteraria Romanica*, Presses Universitaires de Łódź, p. 11-34.
- Chesley P. et Baayen R.H. (2010), « Predicting New Words From Newer Words: Lexical Borrowings In French », *Linguistics*, 48(6), p. 1343-1374.
- Constant, M., Eryigit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., et Todirascu, A. (2017). « Multiword expression processing: A survey ». *Computational Linguistics*, 0(ja), p. 1–92.
- Corbin, D. (1987) *Morphologie dérivationnelle et structuration du lexique*, 2 vol., Tübingen, Max Niemeyer Verlag.
- Coseriu, E. (1964). *Pour une sémantique diachronique structurale*. p. 9–186.
- Cusin-Berche, F. (1999) « Le lexique en mouvement : création lexicale et production sémantique », *Langages* 136 p. 5-26.
- Darmesteter, A. (1887). *La vie des mots étudiée dans leurs significations*. C. Delagrave.
- Deroy L. (2003[1956]), *L'emprunt linguistique*. Presses Universitaires de Liège.
- Dubois J. (1962), *Étude sur la dérivation suffixale en français moderne et contemporain*. Librairie Larousse.
- Dubois J. (2007) [1994], *Linguistique & sciences du langage*, Larousse.
- Firth, J. R. (1957). *Papers in Linguistics 1934–1951*. Oxford University Press.
- Gévaudan, P. et Koch, P. (2010). « Sémantique cognitive et changement lexical. Grandes voies et chemins de traverse de la sémantique cognitive ». *Journée scientifique de la Société de Linguistique de Paris*, p. 103–146.
- Goldberg, A. E. (2013). « Constructionist approaches ». *The Oxford handbook of construction grammar*, Oxford University Press, p. 15–31.
- Gries, S. T. (2010). « Behavioral profiles: A fine-grained and quantitative approach in corpus-based lexical semantics ». *The Mental Lexicon*, 5(3), p. 323–346.
- Hamilton, W. L., Leskovec, J., et Jurafsky, D. (2016). « Cultural shift or linguistic drift ? Comparing two computational measures of semantic change », *ACL 2016*.
- Harris, Z. S. (1954). « Distributional structure ». *Word*, 10(2-3), p. 146–162.
- Hermann, P. (1886). *Principien der Sprachgeschichte*. 5th ed. 1920, Halle:Max Niemeyer Verlag.
- Koch, P. et Oesterreicher W. (1985). « Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit » dans *Spannungsfeld von Sprachtheorie und Sprachgeschichte. Romanistisches Jahrbuch* 36/85, p. 15-43.
- Koch, P. (2000). « Pour une approche cognitive du changement sémantique lexical: aspect onomasiologique ». *Mémoires de la société linguistique de Paris*, p. 10–75.
- Labov W. (1994), *Principles of linguistic change, tome 1 : internal factors*, Oxford, Blackwell.
- Labov W. (2001), *Principles of linguistic change, tome 2 : social factors*, Oxford, Blackwell.
- Langacker, R. W. (1987). *Foundations of cognitive grammar. Volume 1. Theoretical prerequisites*. Stanford University Press Stanford.
- Langacker, R. W. (1991). *Foundations of cognitive grammar: Volume II :Descriptive Application*. Stanford university press.

- Martinez C. (2009), *L'évolution de l'orthographe dans les Petit Larousse et les Petit Robert 1997-2008 : une approche généalogique du texte lexicographique*. Thèse de doctorat soutenue à l'université Cergy-Pontoise.
- McMahon A. (1994), *Understanding language change*. Cambridge : Cambridge University Press.
- Meillet, A. (1904). Comment les mots changent de sens. *L'Année sociologique* (1896/1897-1924/1925), 9, p.1–38.
- Mikolov, T., Yih, W.-t., et Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of HLT-NAACL*, volume 13, p. 746–751.
- Mudrochová R. (2017), « À propos des mots en *-ing* d'origine anglaise issus du dictionnaire Le Petit Robert ». *Linguistica Pragensia*, 27(1), p. 7-19.
- Rastier, F. et Valette, M. (2009). « De la polysémie à la néosémie ». *Texto! Textes et Cultures*, 14(1), p. 97–116.
- Renner V. (2015) « Panorama rétro-prospectif des études amalgamatives ». *Neologica : revue internationale de la néologie*, Paris : Garnier, 2015, 9, p. 97-112.
- 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. et Pruvost, J. (2016). *Les néologismes*. Presses Universitaires de France-PUF, collection Que sais-je ?
- Schmid, H.-J. (2015). « A blueprint of the entrenchment-and-conventionalization model ». *Yearbook of the German Cognitive Linguistics Association*, 3(1), p. 1–27.
- Sinclair, J. (1991) *Corpus, Concordance, Collocation*. Oxford University Press
- Stefanowitsch, A. et Gries, S. T. (2003). « Collocations: Investigating the interaction of words and constructions ». *International Journal of Corpus Linguistics*, 8(2), p. 209–243.
- Stern, G. (1931). *Meaning and change of meaning; with special reference to the english language*. Bloomington: Indiana University Press.
- Tournier, J. (1985). *Introduction descriptive à la lexicogénétique de l'anglais contemporain*. Champion Books.
- Tournier, J. (1991). *Structures lexicales de l'anglais: guide alphabétique*. Nathan
- Traugott, E. C. et Trousdale, G. (2013). *Constructionalization and Constructional Changes*. Oxford University Press, Oxford Studies in Diachronic and Historical Linguistics.
- Weinreich, U., Labov, W., et Herzog, M. (1968). *Empirical foundations for a theory of language change*. Austin : University of Texas Press.

¹ L'ensemble des données est accessible sur la plateforme web de Néoville à l'adresse suivante : <http://www.neoville.org>. Un certain nombre des fonctionnalités décrites sont également accessibles dans la partie publique du site internet.

² Cette étude a été faite à partir des relevés de fréquence disponibles dans *Google Ngrams*, et dans *Europresse* et *Gallica* pour les contextes.

³ Approcher seulement, car les paramètres actuels doivent être affinés.

⁴ Ces traitements sont actuellement effectués avec Treetagger (<http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>), mais seront prochainement connectés avec l'outil Spacy (<https://spacy.io/>), plus générique et multilingue.

⁵ Voir (Cartier, 2016) pour une revue des méthodes utilisées pour la détection automatique des néologismes formels et le détail de la méthode MDE.

⁶ <https://iptc.org/standards/media-topics/>

⁷ Nous entendons par *particularisme* un emploi spécifique à une zone socio-géographique définie. Par exemple, *amender* dans le sens de 'donner une amende' est un particularisme du français parlé au Sénégal, non attesté hors de cette zone. On parle très souvent de canadianismes, de québecismes, de belgismes etc. pour dénoter des particularismes de telle ou telle zone de la francophonie.

⁸ Nous utilisons la définition de (Guilbert, 1975). La notion s'applique « à un terme étranger qui désigne une réalité inconnue ou très particulière et dont l'emploi s'accompagne, nécessairement, d'une marque métalinguistique qui peut être une paraphrase descriptive, soit une note explicative en bas de page quand il s'agit d'un texte écrit ». Nous faisons également nôtres les définitions de (MacMahon 1994 : 209) : « At first, loans are 'xénismes' foreign words normally italicised or enclosed in quotes in a text, and generally translated. These may be nonce forms, or may enter a second stage of 'pérégrinisme', or true adoption, in which they begin to be used more widely, partly by non-bilinguals; at this stage, loans are still seen as foreign ». Les xénismes peuvent donc être considérés comme un premier état vers l'emprunt. Dans notre corpus, des plats typiques, des pratiques sportives, des habitudes spécifiques sont ainsi classifiés comme xénisme.

⁹ L'ensemble des auteurs a participé à cette analyse.

¹⁰ Nous rappelons que dans ce travail nous considérons comme néologisme toute innovation lexicale (ici de forme) dès sa première apparition. De même, concernant la distinction entre néologie et morphologie productive, nous considérons comme néologisme toute forme n'ayant pas d'attestation avant les années 2010 (vérifiés sur Google Ngram).

¹¹ Nous excluons le domaine « général », qui représente plus de 75 % des néologismes, mais les articles correspondants mériteraient sans doute une caractérisation plus fine.

¹² La distribution constatée, restreinte aux mêmes catégories, est la suivante : 53,19 % de noms, 26,50 % de verbes, 11,33 % d'adjectifs, 8,98 % d'adverbes.

¹³ En statistique, on appelle *déviations standard* ou *écart-type (standard deviation)*, la distance entre la valeur minimale et la valeur maximale d'une série. Cette mesure permet d'approcher la dispersion d'une distribution. La médiane rend compte de la valeur moyenne, en additionnant toutes les valeurs individuelles. Elle rend donc mieux compte de la tendance générale d'une série.

¹⁴ L'étude des néologismes ayant une fréquence entre 10 et 50 occurrences montrent que plus de 90% des occurrences sont concentrées sur une période temporelle inférieure à deux semaines. Il conviendrait donc de redéfinir la notion d'hapax (qui, étymologiquement, dénote l'unicité) en combinant fréquence, empan temporel, paramètres diastriques et diatopiques. Il serait peut-être préférable de parler des *néologismes de faible diffusion*.

¹⁵ Cette catégorie est parfois considérée comme étant le résultat d'une double-affixation *non simultanée*, voir notamment (Corbin, 1987).

¹⁶ La liste a été établie sur la base des formants issus du grec et/ou du latin, décrits dans le Petit Larousse (édition 2016) dont on a soustrait les formants savants, qui sont traités dans les compositions savantes et hybrides. De même, nous avons exclu de cette liste ce que nous considérons comme des fracto-lexèmes, formés par troncation sur des lexies contemporaines. Nous avons ajouté à cette liste un petit nombre de formants français, ayant à la fois valeur de préposition (ou adverbe) et de préfixes (*entre, outre, sans, sous, sur*), ainsi que les formants numériques (*déca, déci, hepta, tétra*, etc.). Nous sommes conscients que cette catégorisation est sujette à caution, mais l'établissement de critères pour identifier les affixes, de manière générale, est toujours en débat. La plupart des travaux définissent les affixes sur la seule base de leur non-autonomie, ce qui aboutit à regrouper les différents cas ici

distingués. D'autres critères « synchroniques » ont été proposés, et notamment : (i) l'endocentricité sémantique des lexèmes dérivés, (ii) l'attribution du genre au dérivé, (iii) l'application du formant à plusieurs catégories, (iv) l'émergence d'un sens nouveau non-autonome, pour distinguer les emplois préfixaux d'autres emplois autonomes. Notre classification est plus « diachronique », distinguant les préfixes, issus du grec ancien et/ou en latin, les formants savants, liés à l'émergence du mouvement de dénomination scientifique à partir du XVII^eme, et les fractolexèmes plus récents et construits à partir de lexies modernes ou contemporaines.

¹⁷ <http://robert-illustre.lerobert.com/pdf/dictionnaire-des-suffixes.pdf>

¹⁸ Nous appréhendons la notion de productivité selon deux approches, quantitatives et qualitatives, en nous fondant sur les travaux de Baayen et de Corbin. Nous entendons ici la productivité au sens de la productivité en expansion (*expanded productivity*, Baayen, 2009) qui quantifie le nombre d'hapax nouveaux (de néologismes) créés par la catégorie (ici les préfixes). Cette productivité s'oppose à la *productivité réalisée*, c'est-à-dire passée (correspondant aux réalisations attestées et lexicalisées), et à la productivité potentielle (*potential productivity*) qui cherche à mesurer l'étendue maximale possible de cette productivité en relation avec les contraintes de la règle. Par exemple, *non-* a une productivité potentielle plus grande que *ex-* car il peut être adjoint à des noms et des adjectifs alors que *ex-* s'adjoint uniquement aux noms. Chez Corbin, « la productivité désigne à la fois la régularité des produits de la règle, la disponibilité de l'affixe, c'est-à-dire précisément la possibilité de construire des dérivés non attestés, de combler les lacunes du lexique attesté, et la rentabilité, c'est-à-dire la possibilité de s'appliquer à un grand nombre de bases et/ou de produire un grand nombre de dérivés attestés. » (Corbin, 1987 : 177). Nous entendons dans le tableau la productivité au sens de la productivité en expansion.

¹⁹ Ils sont parfois appelés quasi-lexèmes ou quasi-préfixes (voir Renner, 2015 pour une revue de ce concept), étant donné qu'ils partagent un certain nombre de propriétés avec les préfixes (notamment d'être non-autonomes et plus ou moins productifs)

²⁰ Ou plus exactement l'anglo-américain, ou même mieux l'anglo-américain international.