NÉOVEILLE, PLATEFORME DE REPÉRAGE ET DE SUIVI DES NÉOLOGISMES EN CORPUS DYNAMIQUE

Emmanuel CARTIER
Université Paris 13 Sorbonne Paris Cité, LIPN – RCLN UMR 7030 CNRS – Labex EFL
emmanuel.cartier@lipn.univ-paris13.fr

Résumé

Nous présentons la plateforme Néoveille qui vise à détecter automatiquement, décrire linguistiquement et suivre l'évolution des innovations lexicales en corpus dynamique. Nous détaillons l'architecture générale du système, puis ses modules : gestionnaire de corpus, détection automatique et validation des néologismes formels, description linguistique, outils de suivi des néologismes. La plateforme est accessible à neoveille.org. Le code source à : https://github.com/ecartierlipn/neoveille2016.

Mots-clés : innovation lexicale, outillage informatique, détection automatique, émergence, diffusion et adoption, description linguistique

Abstract

We present a web platform devoted to the automatic detection, the linguistic description and the tracking of the life-cycle of lexical innovations in monitor corpora. We detail the general architecture of the system and its main modules : corpora manager, formal neology automatic detection, neologism linguistic description, tools to track the life-cycle of lexical innovations. The platform is available at : www.neoveille.org. The source code at : https://github.com/ecartierlipn/neoveille2016.

Keywords : lexical innovation, corpus tools, formal neology extraction, life-cycle tracking, linguistic description

Introduction

Nous présentons dans cet article la plateforme Néoveille, issue d'un projet financé par l'IDEX de la COMUE Sorbonne Paris Cité de 2015 à 2018 et qui a depuis continué à évoluer. Nous détaillons l'architecture générale du système, puis ses principaux modules : gestionnaire de corpus, module de détection automatique des néologismes formels, module de validation des néologismes, module de description linguistique, module de suivi des néologismes. Ce travail décrit une mise à jour du système présenté dans (Cartier 2016). Pour une présentation des hypothèses théoriques sous-jacentes, voir (Cartier 2018c : chapitres 1 à 3). Pour une présentation détaillée des tendances néologiques du français contemporain, étude menée à partir de Néoveille, voir (Cartier et al. 2018).

La plateforme est le résultat d'un projet collaboratif entre trois partenaires français (LIPN, équipe RCLN, UMR 7030 CNRS, CLILLAC-ARP EA 3967, HTL UMR 7597 CNRS) et plusieurs groupes de recherche internationaux. Le projet visait à :

- mettre en place une plateforme multilingue de veille et de suivi des néologismes à partir de corpus contemporains dynamiques de très grande taille dans sept langues (français, grec, polonais, tchèque, portugais du Brésil, chinois et russe) ;
- 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 ;
- utiliser cette plateforme pour mener une étude des emprunts (notamment mais pas exclusivement anglicismes) dans les différentes langues.

La plateforme est aujourd'hui accessible, comprenant une partie publique et une partie privée pour l'édition des données : www.neoveille.org. Depuis mi-2017, quatre autres langues ont été ajoutées au projet : l'allemand, l'espagnol, l'italien et le néerlandais.

Détecter automatiquement les changements lexicaux implique plusieurs sous-tâches :
- mettre en place une architecture reproduisant un modèle articulant langue et discours dans un flux continu, et permettant de caractériser les néologismes des points de vue linguistique, socio-pragmatique et cognitif (Schmid 2015) ;
- pour ce qui concerne les néologismes formels, développer des algorithmes pour la détection automatique ou semi-automatique des nouvelles formes (orthographiques/morphologiques) qui apparaissent dans les discours ;
- pour ce qui concerne les néologismes sémantiques, développer des algorithmes pour la détection automatique des modifications des propriétés fréquentielles, linguistiques et/ou socio-pragmatiques des formes lexicales dans les discours.

Enfin, étant donné qu'il s'agit pour nous, certes d'automatiser les différentes tâches, mais de permettre également la collaboration entre des processus automatiques et l'expertise humaine (correction de résultats, ajout d'informations, etc.), l'architecture générale du système doit prévoir une interaction entre les processus automatiques et les interventions humaines. Dans cette présentation, nous détaillons l'architecture de la plateforme et ses différents modules.

1. **Architecture générale : reproduction du flux langue-discours**

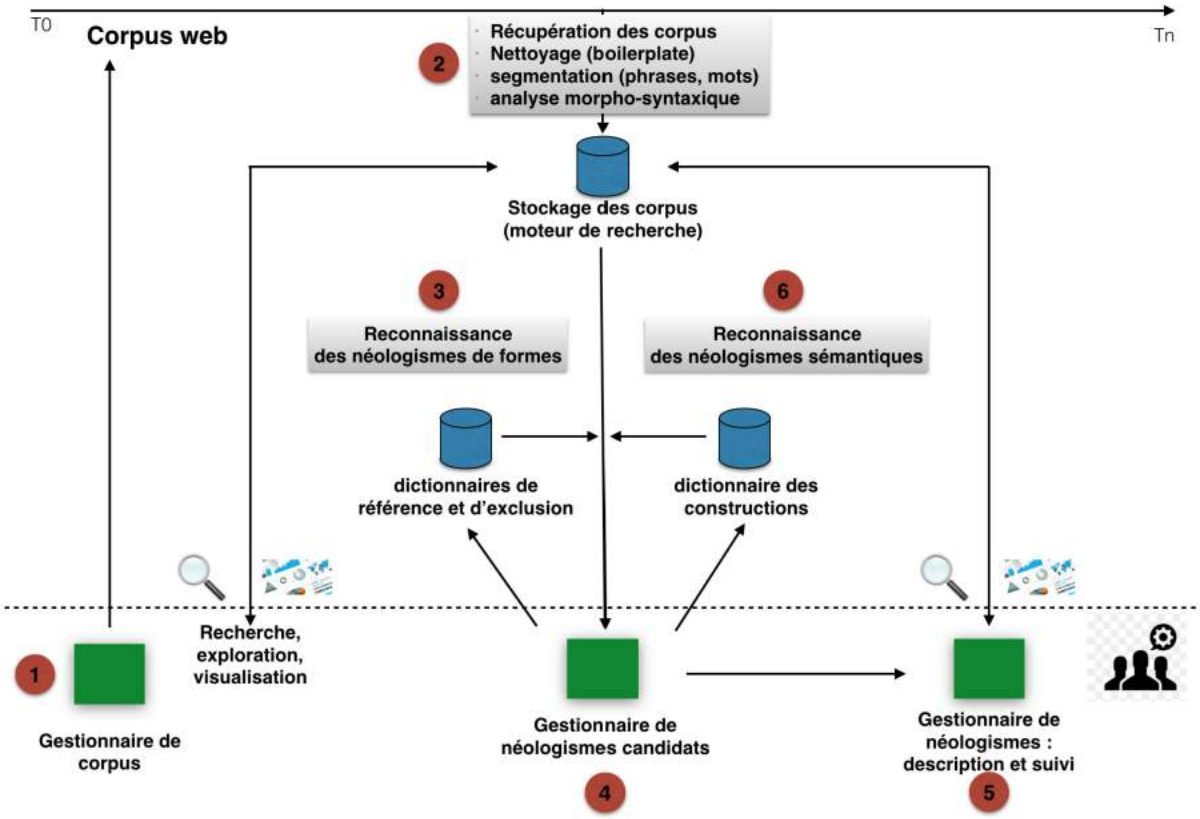

Figure 1. Architecture générale de la plateforme.

Dans cette architecture, le trait horizontal en pointillé sépare les composants où l'expert linguiste intervient (partie basse) des composants auxquels il n'a pas accès (partie haute : processus automatiques). Cette architecture reproduit au plus près le flux langue-discours, en prévoyant une alimentation continue du système en discours (les « corpus web »), ainsi que des processus de traitements (automatiques et manuels, ces derniers étant à la suite des premiers, mais en retour informant les traitements automatiques suivants) aboutissant à créer une *mémoire linguistique active*, sous forme de dictionnaires (carrés verts) et d'un espace de stockage des corpus bruts et annotés (moteur de recherche).

Le système combine également l'analyse automatique et la validation manuelle, les deux s'informant mutuellement : l'analyse automatique propose des candidats néologismes, sur la base de différents algorithmes ; l'expertise manuelle permet de corriger les erreurs des processus automatiques, qui sont automatiquement reversés dans les processus automatiques.

Nous présentons ci-après, dans cette architecture, les cinq premiers grands modules[1] (voir pastilles numérotées dans l'architecture).

2. **Gestionnaire de corpus : caractérisation socio-pragmatique des documents sources**

[1] Le module de détection automatique des néologismes sémantiques ne sera pas présenté ici, par manque de place, et du fait qu'il est encore en phase de tests. Nous renvoyons à (Cartier 2018c : chapitres 4 et 5) pour une présentation des méthodes utilisables pour détecter ce type d'innovation.

Le gestionnaire de corpus permet tout d'abord à l'expert linguiste de déterminer (ajouter, supprimer, modifier) les corpus qu'il souhaite faire analyser par le système, sous la forme d'une interface web spécifique. Actuellement, le système permet de récupérer des fils RSS[2]. À chaque source d'information sont associées des méta-informations, permettant de caractériser socio-pragmatiquement les discours : nom du journal, URL d'entrée, public visé (presse générale ou féminine à l'heure actuelle), domaine (informatique, santé, économie, mode, etc.), langue (parmi les onze langues du projet), pays du journal, type de la ressource (site web ou fil RSS actuellement), fréquence de parution. Nous présentons dans la figure 2 la répartition actuelle, pour le français, des sources d'informations.

Les sources d'informations stockées sont ensuite récupérées automatiquement deux fois par jour. Des (méta-)informations complémentaires sont alors récupérées pour chaque item d'information, dans le flux RSS : titre du document, auteur(s), date de publication et, le cas échéant, mots-clés et domaines. Les fils RSS comprennent également un lien vers l'article web complet : nous récupérons la page HTML, effectuons un zonage de la page pour ne conserver que le contenu textuel utile, segmentons en phrases et en mots le texte et analysons morpho-syntaxiquement chaque unité lexicale[3]. Enfin, toutes ces informations sont stockées dans un moteur de recherche (Apache Solr[4]), pour exploitation ultérieure. Au final, chaque source d'information dispose ainsi de plusieurs informations complémentaires, résumées dans le tableau 1.

Type général d'information	Type d'information	Exemples et/ou informations complémentaires

[2] Nous renvoyons par exemple à (http://www.rssboard.org/rss-specification) pour une présentation détaillée du format XML des flux RSS. Un fil RSS es tun fichier XML contenant une liste d'items d'information (ici des articles de presse), contenant minimalement un titre, la date de publication et un lien vers l'article complet.

[3] Nous renvoyons à (Cartier 2016) pour une présentation détaillée de ces traitements : actuellement, nous utilisons JusText pour effectuer le zonage (https://pypi.org/project/jusText/), et Treetagger pour l'analyse morphosyntaxique (http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/).

[4] http://lucene.apache.org/solr/

Méta-informations associées à la source d'information	Nom du journal	*Le Monde, Valor Economico, etc.*
	Public visé	Presse généraliste, de vulgarisation, spécialisée, presse féminine, etc.
	Type de texte	Dans notre cas, exclusivement « article de presse ».
	Domaine(s)	Général, économie, industrie, etc[5].
	Aire géographique	National, régional, pays, international
Méta-informations associées à chaque item d'information (texte)	Auteur(s)	Auteur(s) explicités dans le flux RSS pour l'item d'information
	Date de publication	Date explicitée dans le flux RSS pour l'item d'information
	Mots-clés	Mots-clés spécifiés dans le flux RSS pour l'item d'information et/ou dans les méta-informations de la page web.
	Domaine(s) et/ou thématique(s)	Informations thématiques spécifiées dans le flux RSS pour l'item d'information et/ou dans les méta-informations de la page web.
Informations liées au contenu textuel	Titre	Titre tel que spécifié dans le flux RSS pour l'item d'information
	Contenu textuel brut	Contenu textuel résultat de l'application du programme de zonage.
	Contenu textuel annoté morphosyntaxiquement	Contenu textuel annoté suite à la segmentation du texte brut en phrases et token. Pour chaque token, on obtient la forme brute, la partie du discours[6] et le lemme.

Tableau 1. Liste des informations disponibles pour chaque item d'information textuelle récupéré.

Sur l'interface web il est possible d'obtenir une vision synthétique des informations textuelles récupérées, de l'évolution diachronique, ainsi que les distributions pour chacun des paramètres (voir figure 2).

2.1. ***Perspectives***

L'inconvénient majeur des fils RSS ressortit à l'utilisation non systématique de ce format par les sites web, ce qui nécessite d'autres procédures pour accéder aux contenus non couverts. Actuellement, nous étudions une autre piste, permettant d'accéder à des sites internet bien plus diversifiés : il s'agit des corpus stockés par commoncrawl (commoncrawl.org), donnant accès à des pages web depuis 2013 sur l'ensemble des langues couvertes par la plateforme.

[5] Actuellement, nous utilisons une nomneclature inspirée de la typologie proposée par le consortium international IIPTC (https://iptc.org/standards/media-topics/), simplifiée à une quinzaine de catégories.

[6] Etant donné que nous utilisons actuellement Treetagger, les jeux d'étiquettes varient d'une langue à l'autre, et nous avons unifié les valeurs en utilisant le jeu d'étiquettes universelles proposé par « Universal Dependencies », très largement utilisé dans la communauté du Traitement automatique des langues. (voir http://universaldependencies.org/u/pos/).

D'autre part, un module complémentaire de détection des thématiques de chaque texte est en cours de développement, combinant les approches à base de ressources linguistiques et les méthodes non-supervisées (Cartier, Galand, Stirling et Aubry 2018).

Enfin, une étude plus précise des méta-informations à associer à chaque texte, ainsi que les procédures pour le faire (approche manuelle et/ou automatique) est en cours. L'objectif est de caractériser finement les situations de communication et d'ainsi spécifier les lieux d'occurrences des innovations lexicales[7].

3. **Repérage automatique des néologismes formels et validation manuelle**

Ce module permet, dans Néoveille, de détecter, dans les articles de presse stockés dans le moteur de recherche, des candidats néologismes après application de plusieurs filtres : dictionnaires de référence et d'exclusion, noms propres, erreurs typographiques. Les candidats néologismes sont ensuite présentés aux experts, qui vont valider ou invalider 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. Pour une présentation de l'état de l'art en détection automatique des néologismes formels, nous renvoyons à (Cartier 2018c, chapitre 5).

L'algorithme combine trois approches : d'une part, nous utilisons les sorties du Treetagger, qui détecte les formes inconnues, et donc potentiellement néologiques. Ensuite, nous éliminons les formes débutant par une majuscule[8], puis filtrons la liste des candidats avec un correcteur orthographique (*Hunspell*), permettant d'éliminer la majorité des erreurs typographiques et autres coquilles[9]. Enfin, nous utilisons des listes d'exclusion construites dynamiquement par le biais de l'étape de validation manuelle. Cette amélioration continue du système permet de passer d'une précision initiale (sans dictionnaire d'exclusion, mais avec le dictionnaire sous-jacent du Treetagger) d'environ 45 % à un taux supérieur à 60 %, après une phase de construction itérative du dictionnaire d'exclusion. Cette méthode itérative permet en outre de construire et de mettre à jour la ressource lexicographique de référence, pour d'autres usages. Les évaluations menées montrent que pour les langues romanes, cette approche permet, à partir de deux semaines de travail et le traitement d'environ 5 000 néologismes candidats, d'éliminer les erreurs de détection les plus fréquentes. Cependant, pour d'autres langues, le système d'apprentissage itératif est moins efficace.

3.1. ***Perspectives***

Le système actuel donne satisfaction, repérant, pour le français, pour 250 sources d'informations et entre 3000 et 7000 documents par jour, entre 100 et 200 néologismes

[7] Nous ne détaillons pas ici cet aspect, mais il s'agit de l'un des deux paramètres essentiels permettant de caractériser les néologismes, avec le paramètre linguistique. L'évolution des néologismes, de leur émergence à leur éventuelle diffusion, revient à étudier les modifications des propriétés linguistiques de la lexie, et les modifications des contextes d'énonciation.

[8] Certains systèmes (par exemple Le Logoscope) visent également à repérer les néologismes désignant des entités rigides, c'est-à-dire des noms propres. Les noms propres n'échappent effectivement pas à la dynamique créative des langues. Dans le cadre de Néoveille, nous excluons les noms propres néologiques, pour deux raisons principales : la détection automatique nécessiterait une liste d'exclusion impossible à construire et cela occasionnerait beaucoup trop de bruit dans les résultats. Nous repérons, par contre, les néologismes à base noms propres, dès lors que les bases sont orthographiées avec la casse minuscule (exemples : *macronien, trumpiste*, etc.) (voir Cartier, 2018b, pour une étude).

[9] Afin d'éviter la sur-correction, un seuil d'édition de 3 pas est fixé.

candidats. Nous avons cependant expérimenté des méthodes d'apprentissage automatique (Lejeune et Cartier 2017) montrant qu'à partir d'un corpus d'apprentissage (10 000 néologismes validés, 20 000 lexies appartenant à un dictionnaire de référence et 10 000 lexies non-néologiques), il est possible d'atteindre un taux de précision d'environ 70 %. Nous avons mené une étude, non publiée, avec un réseau de neurones *feed-forward*, montrant qu'il est même possible d'atteindre une précision proche de 90 %, en ne tenant compte que de la forme interne des lexies. Nous allons mener une étude plus approfondie sur ce point, étant donné d'une part qu'un tel système est beaucoup plus léger que le système existant, et qu'il peut facilement être étendu à de nouvelles langues, à partir d'un jeu de référence.

3.2. *Validation manuelle des néologismes candidats*

Dans Néoveille, les néologismes formels sont d'abord détectés automatiquement, puis les experts humains doivent valider ou non les propositions du système. Une interface spécifique permet d'effectuer cette opération, en catégorisant les lexies présentées selon deux catégories : non-néologismes et néologismes.

3.3. **Présentation des interfaces**

Le module « néologismes de forme » donne accès à une liste de néologismes candidats (NC) sous forme d'un tableau (figure 3).

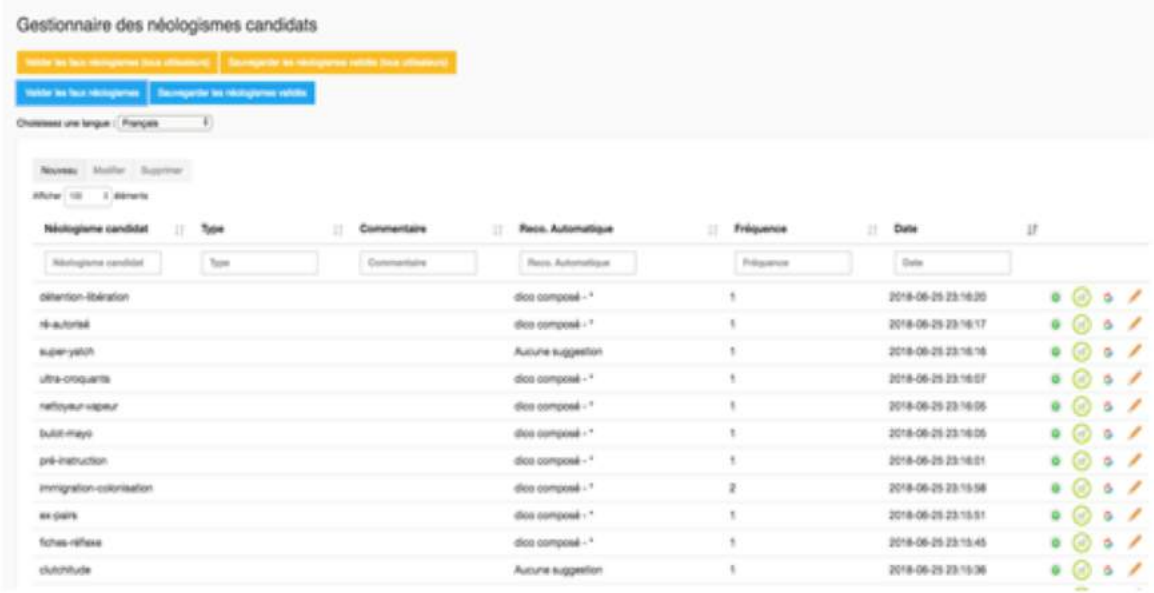

Figure 3. Interface de validation-invalidation des néologismes de forme automatiquement détectés.

Cette interface présente un certain nombre d'informations pour chaque NC, en dehors de la forme exacte reconnue : informations sur la reconnaissance automatique, suggérant à l'utilisateur la catégorie du NC ; une information sur la fréquence constatée dans le corpus du NC ; une information sur la date de la première occurrence rencontrée. Deux champs (type et commentaire) sont à remplir par les experts linguistiques, d'une part pour indiquer le type du

NC (voir plus loin), d'autre part, pour saisir tout commentaire souhaité. Pour décider si un NC est un néologisme ou non, l'expert dispose de plusieurs informations complémentaires :
- la visualisation du ou des contextes d'apparition de la forme exacte (en cliquant sur l'icône vert rond, à droite de chaque ligne, voir figure 4).
- la visualisation des contextes enrichie d'informations sur les caractéristiques socio-pragmatiques des discours dans lesquels apparaît le NC (actuellement : pays d'origine du journal source, domaine, nom du journal) et l'évolution temporelle des occurrences (en cliquant sur l'icône vert représentant un graphe, à droite de chaque ligne, voir figure 5).
- la visualisation des occurrences éventuelles de cette forme dans Google Ngrams, donnant accès à une information sur l'existence ou non de cette forme dans un corpus couvrant la période 1800-2010 (en cliquant sur l'icône Google à droite de chaque ligne).

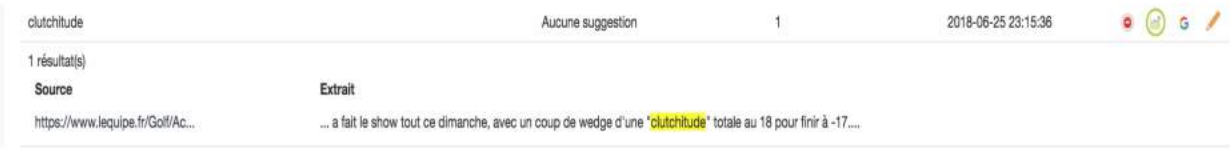

Figure 4. Exemple de visualisation de contextes pour le candidat néologisme *clutchitude*

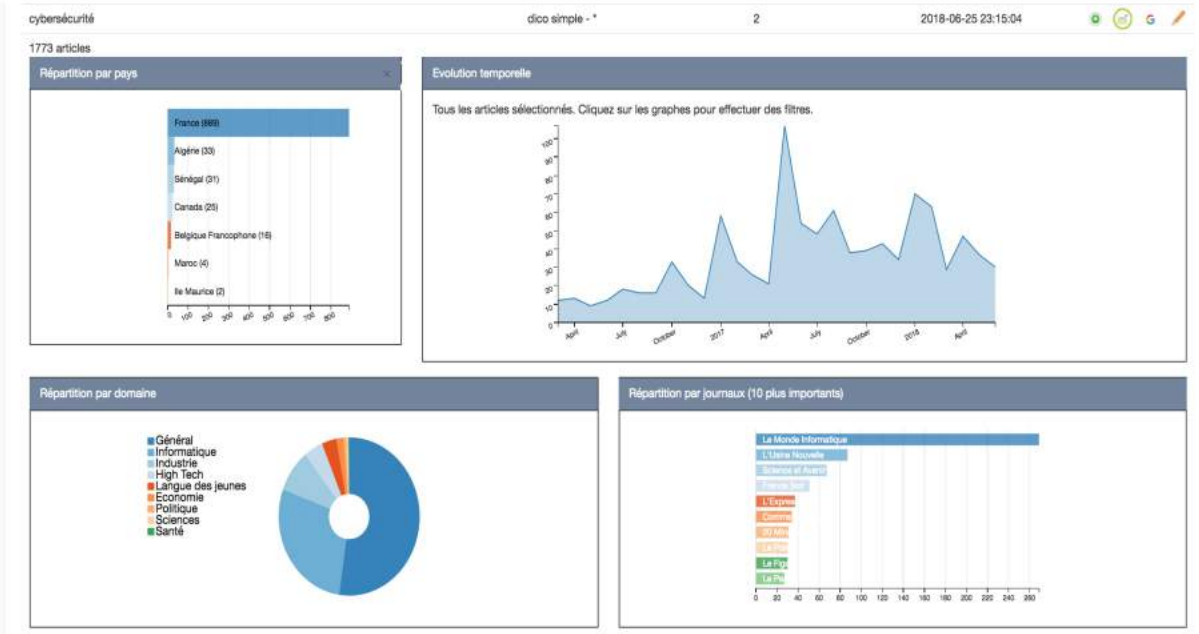

Figure 5. Exemple de visualisation enrichie (les contextes sont omis) pour le candidat néologisme *cybersécurité*.

3.4. *Critères de validation et d'invalidation des néologismes candidats*

Le travail effectué durant la période 2015-aujourd'hui a permis également d'affiner la catégorisation des non-néologismes. En effet, un dictionnaire de référence et d'exclusion ne peut jamais contenir l'ensemble des non-néologismes, car plusieurs types de fomes-lexies se présentent, résumés dans le tableau 2.

Catégorie	Descriptif rapide	Exemples

Dictionnaire mot simple	Lexie non présente dans le dictionnaire de référence et qui devrait y figurer	*Courriel, événementiel, blog...*
Dictionnaire mot composé	Lexie à trait d'union non présente dans le dictionnaire de référence et qui devrait y figurer	*Pontier-cabine,plongeur-démineur, ultra-simple, primo-arrivant, etc.*
Dictionnaire terminologique	Lexie appartenant à un domaine terminologique	*Nucifera, polykystose, micromoteur, etc.*
xénisme	Lexie empruntée à une autre langue, mais dans le cadre du *code-switching* et dénotant une réalité locale	*Lujo, furoshiki, rojigualda, tawakkul, etc.*
gentilé	Lexie désignant un individu ou une caractéristique lié à un lieu, une zone géographique, une culture spécifique	*Amuesha, cubano-mexicaine, sino-russe, etc.*
particularisme	Lexie entrée dans l'usage pour une aire socio-géographique spécifique	*Xessal, tcha-tcho*[10]
Erreur typographique et autres erreurs	Erreurs diverses liées à l'orthographe	Spect, terroristea, berbatov, jijadiste, acceuille, traditionel, endless, etc.

Tableau 2. Catégories de non-néologismes.

Parmi ces catégories, les deux premières comprennent des lexies entrées dans l'usage, mais qui ne sont pas contenues dans le dictionnaire de référence (initial ou constitué via la plateforme). Le classement de ces unités permet de mettre à jour une ressource lexicographique de référence[11]. Les lexies terminologiques représentent une autre catégorie qui montre la porosité entre le vocabulaire spécialisé et le vocabulaire général, avec des passages de l'un vers l'autre par le biais des articles de vulgarisation scientifique de la presse généraliste. Là encore, on peut y voir l'impossibilité pour un dictionnaire de référence de couvrir l'ensemble des lexies attestées. Les gentilés sont une autre catégorie de lexies qui sont exclues : il s'agit de formations à partir de noms propres géographiques ou socio-ethniques (notamment par suffixation : les *nzebis*, et par composition : *sino-russe, anglo-néo-zélandais*, etc.). Les particularismes, unités employées dans une aire socio-géographique déterminée, sont plus rares.

Les catégories de néologismes sont issues de la typologie proposée par (Sablayrolles, 2016). Parmi les procédés, l'affixation, parfois assimilée à la *morphologie productive* et donc partie intégrante des procédés internes disponibles en langue (Dal 2003), se situe à la frontière des non-néologismes. On peut en effet considérer que ces formations font partie du *vocabulaire potentiel* d'une langue. Cependant, la situation est plus complexe, puisque, à côté de formants extrêmement disponibles et dont les formations sont d'ailleurs généralement ressenties comme non-néologiques (*non, anti, ex,* etc.), malgré l'absence d'attestations dans le passé (par exemple *anti-austérité, anti-cybercriminalité, anti-héros, anti-raciste*), d'autres éléments forment des lexies ressenties comme néologiques (par exemple *plurifilière*).

[10] En français du Sénégal, ces deux termes désignent l'action de se dépigmenter volontairement la peau.
[11] Nous allons débuter une action pour mettre à jour le dictionnaire Morfetik en utilisant les données Néoveille (voir Grezka et al. 2015)

Comment faire le départ entre ces cas ? Dans Néoveille, nous avons considéré que le procédé d'affixation, *dans toutes ses réalisations*, fait partie de la néologie.

En complément, nous considérons qu'une lexie est néologique dès lors qu'elle n'est pas attestée avant 2010, en essayant de réduire au maximum le recours au *sentiment néologique*. Un ensemble de ressources permet de vérifier cette non-attestation : corpus (*Google Ngram*[12], qui permet d'obtenir les courbes de fréquence des lexies sur un corpus représentant actuellement près de 7 % des livres imprimés depuis les années 1600 jusqu'à 2008 (Michel et al. 2010). *Europresse*[13], qui permet de chercher dans la presse française depuis 1945 des attestations. *FrWac*[14], qui permet de vérifier la non-attestation sur un corpus web de 2010 à 2013 (Baroni et al. 2010)) ; nous utilisons également des dictionnaires de référence (le TLFI[15], le Larousse, le Robert, Wiktionnaire) ainsi que des dictionnaires historiques (comme le Dictionnaire historique de la langue française).

On constate un continuum entre les non-néologismes et les néologismes : si certains cas sont clairs (lexies attestées non incluses dans le dictionnaire de référence, erreurs typographiques), d'autres cas sont limitrophes, du côté de ce qui ne fait pas partie de la mémoire linguistique *générale* (technolectes et variétés diatopiques), et du côté de ce qui constitue un procédé disponible dans la langue pour la création lexicale (affixation, gentilés). Le continuum entre xénisme et emprunt est également indéniable, les xénismes constituant l' « antichambre » des emprunts lexicaux (voir (Cartier 2018b) pour une étude)

3.5. ***Protocole de validation des néologismes***

Dans Néoveille, les validations ont été effectuées en suivant le protocole suivant : chaque membre du groupe de travail pour le français annote individuellement 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 (voir 3.4.). Puis, lors de réunions collectives mensuelles, une validation est effectuée, les cas litigieux étant tranchés sur la base d'un vote majoritaire. Ce protocole a permis de valider, sur deux ans et six mois (juin 2015- fin 2017), pour le français, un peu plus de 21 000 néologismes à composante formelle. En moyenne, cela représente environ 60 % des candidats néologismes, pour les évaluations que nous avons menées en français, en espagnol et en russe.

4. ***Description des néologismes***

Une fois validés, les néologismes sont ensuite décrits dans le gestionnaire des néologismes. En suivant le modèle de (Schmid 2015), trois types de propriétés permettent de caractériser les néologismes : les propriétés linguistiques, les propriétés socio-pragmatiques et les propriétés cognitives. Pour chacune de ces perspectives, nous pouvons décrire les trois phases saillantes de la vie d'un néologisme : son émergence, son éventuelle diffusion et son éventuelle lexicalisation (ou adoption). Dans Néoveille, nous nous intéressons aux descriptions linguistiques et socio-pragmatiques puis à l'évolution de ces propriétés.

[12] https://books.google.com/ngrams
[13] Malheureusement, cette base n'est pas accessible gratuitement. La base *Factiva* propose un corpus équivalent.
[14] https://corpora.dipintra.it/public/run.cgi/first_form
[15] http://www.cnrtl.fr/definition/

4.1. ***Description linguistique des néologismes***

La description linguistique des néologismes consiste tout d'abord à caractériser les mécanismes de formation. En utilisant une version simplifiée du modèle de (Cartier et Sablayrolles 2009). Néoveille propose une microstructure générique pour l'ensemble des néologismes comprenant les champs détaillés dans le tableau 3.

Informations	**Définition succincte**	**Exemple pour *food truck***
Partie du discours	Catégorie morphosyntaxique parmi : nom, verbe, adjectif, etc.	Nom commun masculin
Classe sémantique	Classe sémantique générique. Inspirée de (Le Pesant et Mathieu-Colas 1998)	Inanimé concret
Définition		Véhicule utilitaire ambulant délivrant de la nourriture. La dénomination empruntée est utilisée pour désigner un mouvement en cours lié à une mode de vente ambulante de nourriture ethnique, née aux États-Unis.
Procédé(s) néologique(s) impliqués	Le ou les mécanismes néologiques impliqués dans l'innovation lexicale, en partant de la typologie de (Sablayrolles 2016)	emprunt
Configuration syllabique	Description générique et détaillée de la configuration syllabique de la lexie, au moyen des notions de syllabe ouverte (O) et fermée (F).	F F (food-truck)
Configuration morphologique	Décomposition morphologique de l'innovation, au moyen des notions de radical, d'affixe et de formant.	RAD RAD (food-truck)
Lexie base	Identification de la ou des lexies ayant servi de base au néologisme	Food, truck
Partie du discours lexie base	Identification de la partie du discours de la lexie base, ou de la racine.	Nom

Tableau 3. Microstructure utilisée dans Néoveille

À ces informations, permettant notamment par la suite d'effectuer des statistiques sur les procédés les plus couramment utilisés, il faut ajouter d'autres informations concernant des néologismes spécifiques, par exemple, l'influence éventuelle d'une autre langue (exemple : *réaliser* dans le sens 'comprendre', est influencé par l'anglais *to realize*) et le mode de cette influence.

Il faut encore y ajouter trois informations linguistiques qui sont disponibles de manière automatique sur la plateforme depuis 2018 : la *famille morphologique* associée à l'innovation étudiée ; le *profil combinatoire* des occurrences dans le corpus, permettant de détecter les collocations (Firth 1957), les collostructions (Stefanovitsch et Gries 2003) et les constructions lexico-syntaxiques les plus fréquentes, aboutissant à la notion de profil combinatoire (Gries,

2010) ; le *profil distributionnel*[16], permettant d'accéder aux lexies sémantiquement similaires (et donc notamment (quasi-)synonymes, hyperonymes et hyponymes). Nous illustrons les deux premières informations dans le tableau 4, la troisième nécessitant un corpus plus étendu pour obtenir des résultats fiables[17].

Type d'information	**Description sommaire**	**Exemples pour *food***
Famille morphologique	Ensemble des lexies formées sur la même base (y compris mot composé à trait d'union)	*foodies, fooding, foods, food-biz,food-market(s), food-truck(s), food-deco, foodeur(s), foodflock, foodista(s)...* liste complémentaire (noms propres) : *Food4Good, FoodChéri, FoodOrganic, FoodStocks, FoodTech, FoodTemple, FoodWatch, Foodora ...*
Profil combinatoire	Ensemble des collocations, des collostructions et des constructions lexico-syntaxiques représentatives[18]	Collocations : *fast food* (16), s*low food (16), street food (11), raw food (9), junk food (7), food market (7)* Collostructions : *tendance food* (10) => N food(ADJ ?) phénomène food (9) => N food(ADJ ?) *projet food* (5) = > N food(ADJ) *Det (masc) food* (10) => food (NOM) Constructions lexico-syntaxiques : food + verbe : *aller, débarquer, arriver, consister, cartonner…*

Tableau 4. Informations linguistiques automatiquement détectées dans Néoveille, pour food

4.2. *Caractérisation socio-pragmatique des occurrences de néologismes*

La description interne – linguistique – doit être complétée par une description socio-pragmatique, qui permet de caractériser les propriétés des contextes d'emploi : caractéristiques des énonciateurs, des audiences, des supports matériels de diffusion, des types de discours, domaines. Le courant de l'analyse du discours fournit dans ce cadre plusieurs pistes pour caractériser les discours dans lesquels s'insèrent les innovations lexicales (voir Charaudeau 1995 et Paveau 2017 par exemple), ainsi que les travaux sur l'ethnographie de la communication (Hymes 1974). Les métadonnées proposées dans

[16] La notion de profil distributionnel provient des intuitions du distributionnalisme. Harris évoque très tôt cette 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.

[17] Pour illustrer le profil distributionnel, citons par exemple *tsunami*, qui, des années 1900 à nos jours, a pour lexies sémantiquement similaires « raz-de-marée, phénomène violent, etc. ». À partir des années 1970, on voit apparaître « phénomène social, tendance, etc. » montrant une extension de sens (*tsunami politique, social,* etc.). Enfin, depuis 2004, apparaissent de nouveaux synonymes « grand nombre de, plusieurs, etc. », liés à l'apparition de l'emploi comme déterminant complexe au sens de « grand nombre d'éléments, avec connotations de grande puissance et de soudaineté » (*un tsunami d'applaudissements*). (Hamilton et al. 2016) proposera une première application de ce principe, sur le corpus Google Ngrams. Il montre par exemple 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 1970 (à noter que le *Dictionnaire historique de la langue française* date les premières apparitions en anglais d'Amérique du Nord de ce sens au début des années 50 comme adjectif et en 1971 comme nom). Cette technique, *modulo* les limitations dues au corpus source, est l'une des voies les plus prometteuses pour détecter des changements sémantiques.

[18] Nous indiquons ici la fréquence constatée dans le corpus Néoveille.

(Siepmann *et al.* 2016) pour caractériser les textes du Corpus de référence du Français Contemporain peuvent également servir de guide. Dans le cadre de Néoveille, actuellement, seulement quatre propriétés sont disponibles : le public visé, le type de texte (exclusivement article de presse à ce jour), le domaine, lié à l'information thématique fournie par les producteurs eux-mêmes pour chaque document ; le pays ou la région d'origine des journaux. Par exemple, pour l'emprunt *confort food*, apparu en français mi-2017 dans la presse couverte par Néoveille, en utilisant ces paramètres, on peut obtenir la répartition par journaux, et la distribution temporelle de cette répartition (figure 6) montrant que le terme est très lié à la presse féminine, et aux francophonies belge et québécoise.

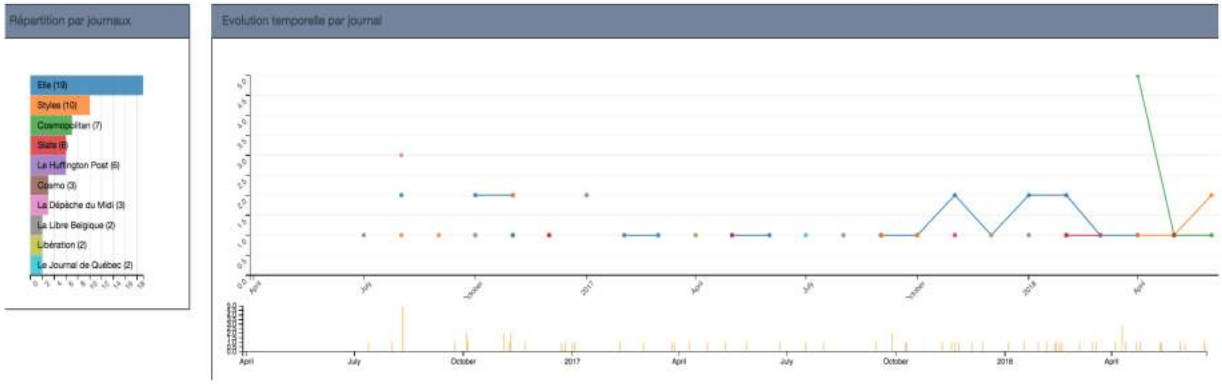

Figure 6. Répartition par journal de l'emprunt *confort food*, et distribution temporelle par journal.

4.3. ***Synthèse globale sur les néologismes validés***

Ces différentes caractérisations permettent non seulement de décrire les néologismes particuliers, mais permettent ensuite des constats globaux. Ainsi, pour le français, on constate que la presse destinée aux femmes (et dans cette catégorie principalement pour certains domaines comme la mode et les réseaux sociaux) utilise de manière plus massive l'emprunt (à l'anglo-américain international), par contraste avec les autres diffuseurs. Du point de vue des matrices de création néologique, pour le français, pour les 20 000 néologismes validés, on constate une prédominance des préfixations (plus de 75 % des formations), suivies des formations composées (7,32 %) et des emprunts (6,36 %). (voir tableau 5 pour le détail sur la répartition par mécanisme néologique).

	Nombre de néologismes (formes uniques)		Nombre d'occurrences de néologismes		
Mécanisme néologique principal	Nombre	%	Nombre	%	Moyenne d'occ. par forme néologique
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
compocation	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 %	

Tableau 5. Synthèse sur les mécanismes néologiques pour le français (2015-2017). 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.

Ces synthèses peuvent être générées automatiquement via un onglet spécifique dans le module de description des néologismes.

5. **Suivi du cycle de vie des néologismes : émergence, diffusion, lexicalisation**

Nous avons vu que trois phases saillantes sont identifiables pour caractériser le cycle de vie des néologismes : l'émergence, la diffusion et la lexicalisation. La plateforme Néoveille a développé un certain nombre d'outils, et établi une première série de critères pour caractériser ces différentes phases.

5.1. ***Émergence : définition et critères***

On définit l'émergence comme le moment d'apparition d'une nouvelle forme (ou d'un nouvel usage) : le critère le plus évident est donc l'existence d'une première attestation.

Si l'on étudie la fréquence des néologismes détectés dans Néoveille, on constate que la moyenne d'occurrences est relativement faible, d'autant plus si nous limitons le comptage aux occurrences survenant dans des documents différents. Pour les néologismes du français, par exemple, la fréquence moyenne est de 26 par néologisme. La déviation standard[19] est cependant importante (111 par forme néologique, 237 par nombre total d'occurrences), montrant qu'il existe un petit nombre de 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 large majorité de néologismes sont donc principalement des hapax ou des quasi-hapax. La figure 7 présente les données de manière plus détaillée : pour les six types de néologismes principaux (en ordonnée), on identifie la distribution des néologismes par nombre d'occurrences (en abscisse).

[19] En statistique, on appelle *déviation 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.

fréquence	1	2	3	4	5	6-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100	100 et +	total
nbre de néol.	363	166	84	85	58	167	148	69	51	35	43	17	15	16	16	244	1262
%	28,76%	13,15%	6,66%	6,74%	4,60%	13,23%	11,73%	5,47%	4,04%	2,77%	3,41%	1,35%	1,19%	1,27%	1,27%	19,33%	100,00%

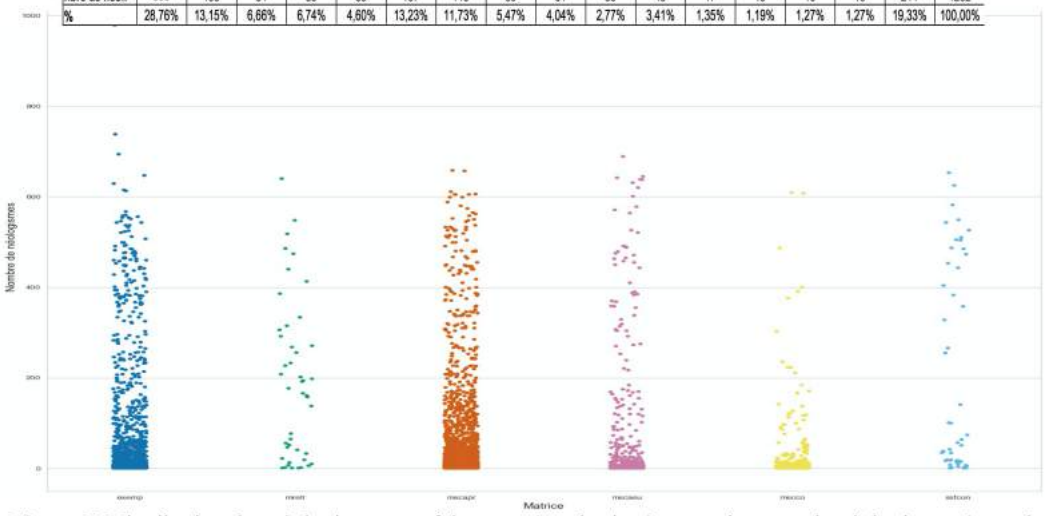

Figure 7. Distribution des néologismes par fréquence (en abscisse), pour six types de néologismes (en ordonnée)

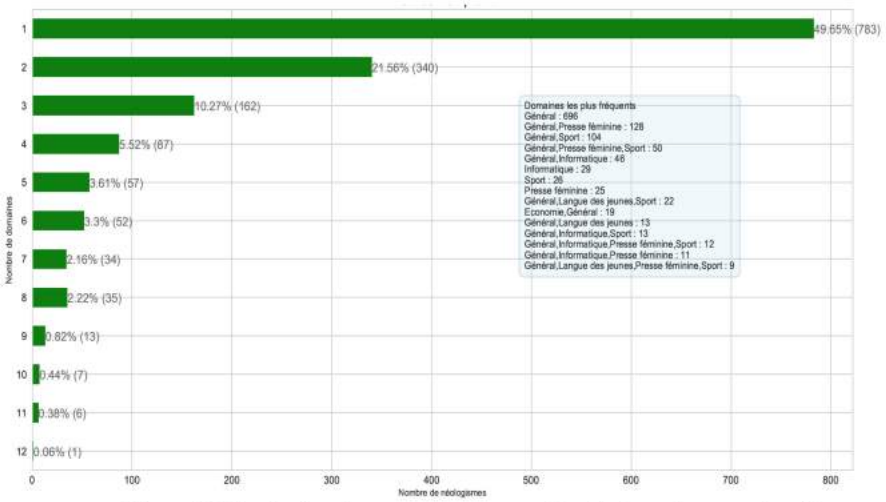

Figure 8. Distribution des emprunts par nombre de domaines représentés.

On constate que les hapax proprement dits (une seule occurrence) représentent *seulement* 25 % du total des néologismes, et il existe un continuum entre les fréquences très basses et les fréquences plus hautes, sans qu'il soit possible de tracer une frontière *linguistique* pertinente entre les hapax et les autres néologismes (notamment ceux ayant une fréquence « faible »). Si nous étendons le moment d'émergence à deux semaines à partir de la première apparition, nous constatons que plus de 70 % des néologismes sont représentés, ce qui tend à montrer que la non-diffusion serait mieux définie comme une répétition « faible » sur une période courte, plutôt que par la notion d'hapax. On peut encore affiner cette définition : *l'émergence est une période courte (de l'ordre de quelques jours ou quelques semaines) durant laquelle une innovation lexicale apparaît et peut se répéter, très généralement dans le même domaine que celui de la première apparition.* Par exemple, *zebracake* et *tigercake*, malgré un *buzz* (7 et 9 occurrences) en avril 2016, sont restés cantonnés à la presse féminine dans les rubriques

culinaires. De même pour *street-girl*, qui est resté cantonné à la presse féminine. La figure 8 illustre la validité de ce critère, en montrant la répartition des emprunts selon le nombre de domaines représentés : on constate que près de 50 % des innovations sont cantonnées à un seul domaine (principalement le domaine général : 696, l'informatique : 29, le sport : 26 et la presse féminine : 25). Par contre, dès qu'une innovation est représentée dans plus d'un domaine, il s'agit d'un signe de diffusion (par exemple, ici la combinaison domaine général-presse féminine comprend 128 lexies, la combinaison sport-général 104).

Ces critères additionnels à la notion d'hapax pour caractériser l'émergence sont liés à la structuration contemporaine de la communication : d'une part, la communication numérique accélère la diffusion des autres informations publiées, favorisant la répétition ou les *mèmes* ; d'autre part, de nombreux groupes de presse détiennent plusieurs journaux, et il est fréquent de voir des répétitions survenant quasiment au même moment, dans des journaux différents, avec reprise intégrale de la phrase complète (voire de l'article...). Le texte suivant, par exemple, se retrouve, à quelques heures d'intervalle, dans trois titres différents, le 18 mars 2016 (nous soulignons les innovations, partiellement basées sur trois emprunts anciens (*cookie, doughnut* et *brownie*, début du XIX^{e}), sur le mode de la compocation) :

> Les croisements sont à la mode. Ils ont déjà accouché du crookie (croissant mélangé avec un cookie Oreo), le duffin (mariage entre doughnut et muffin anglais), le bronut (brioche feuilletée sucrée), et surtout le cronut, union entre le croissant et le doughnut, du chef français Dominique Ansel, basé à New York. (18 mars 2016, *L'Express, Libération, Le Parisien*).

Enfin, un dernier critère est particulièrement utile pour détecter une phase d'émergence : la mise en exergue de la lexie, d'une part, et l'existence d'une glose dans l'environnement immédiat du néologisme. Cette glose est évidemment requise pour la très grande majorité des innovations, dans un souci de compréhension par les destinataires : « … nombre d'applications identifiées par Symantec comme étant des "madwares". À la croisée des malwares et des adwares, cette génération d'apps... » (*Le Monde Informatique*, 13 avril 2016). Cependant, là encore, ce critère n'est pas suffisant, notamment lorsque l'énonciateur considère que le sens est inférable par composition : « ...Réponse jeudi où se tiendra juste après le défilé, une grande after-party... » (*Elle*, 21/02/2017). Nous renvoyons à (Jacquet-Pfau 2018) pour une étude détaillée de ces marqueurs de glose pour les emprunts.

Une dernière caractéristique des quasi-hapax est liée au périmètre sémantique des unités lexicales : elles désignent dans leur très grande majorité, soit des réalités locales – à la limite du xénisme – (*dibbuk, wapeningen, escrache*), soit des concepts circonscrits à un domaine spécifique (*pika-don, nanotrading, cutlet, fadeaway*) ou à des pratiques sociales confidentielles (*selfie-whore-stick, denki-buro, nightswapping*).

5.2. **Diffusion des néologismes : exemples et enseignements**

Parmi les néologismes repérés, plus de 85 % sont des hapax ou des quasi-hapax (fréquence inférieure à 50 occurrences sur une période inférieure à deux semaines). Mais qu'en est-il des 15 % restants ? Suivons les chemins de leur diffusion.

5.2.1. *Adaptations phonologique, orthographique et morphosyntaxique*

Une première étape de diffusion concerne les emprunts, puisqu'ils proviennent d'un autre système linguistique. Cette adaptation au système du français concerne les niveaux phonologique, orthographique et morphosyntaxique. Concernant l'adaptation orthographique, le corpus Néoveille ne présente pas de particularité concernant les emprunts à des langues alphabétiques, dont les lexèmes sont généralement rendus tels quels[20]. Pour l'arabe, la translittération donne parfois lieu à des hésitations (*jihadiste, djihadiste,* avec cependant une prédominance depuis environ 2010 de la version *dj-*).

Concernant l'adaptation phonologique, elle est plus difficile à déterminer étant donné le corpus écrit. Notons cependant le cas de *check*, qui dans tous les cas conserve la prononciation anglaise (/ʧ/).

L'adaptation morphologique, pour les anglicismes, ne présente pas, généralement, de difficultés pour les noms (sans ajout de morphème) et les verbes (par ajout du morphème *-er*), mais une particularité déjà étudiée par (Saugera 2017 : 123-138) concerne les adjectifs empruntés à l'anglais, qui, au pluriel, s'accordent ou non, selon le cas. Dans notre corpus, on retrouve ainsi un grand nombre d'adjectifs en *-y*, invariables (*arty, sketchy, glowy, skinny, girly, creepy, healthy, edgy, catchy, flashy, bluesy*, etc.), au point qu'il est probable que ce formant soit devenu un formant suffixal productif en français.

5.2.2. *Intégration à la morphologie productive*

Un autre signe de diffusion concerne l'intégration à la morphologie productive. La grande majorité des néologismes à fréquence élevée, relevés dans notre corpus, subissent ce « passage » à la dérivation. Les exemples les plus typiques sont liés aux bases de *noms propres* issus des réseaux sociaux (tableau 6). Dans ce tableau, nous distinguons les dérivations par morphèmes grammaticaux (nom et verbe) des affixations proprement dites. Nous constatons que l'étendue des dérivations est liée à la popularité du réseau, *Twitter* étant largement en tête. Toutes ces dérivations appliquent les procédés affixaux les plus productifs du français contemporain (Cartier *et al.* 2018). On notera la particularité de *twitto(s)*, pour désigner le ou les émetteur(s) d'un *tweet*, directement emprunté de l'anglais et sans concurrence du suffixe français pourtant très productif en *-eur*(*euse*) et utilisé pour les autres termes. De même, on notera que seuls *Twitter* et *Snapchat* ont une lexie (directement empruntée) pour désigner le type de message (*tweet* et *snapchat,* parfois tronqué en *snap*). Cependant, *Facebook* a aussi permis l'emprunt de *like* (N) et *liker* (V).

	facebook	twitter	instagram	snapchat	youtube
Intégration morphologique (morphème flexionnel)	facebooker (v)	twit(s) (n) tweeter (v) twitto(s) (n)	instagram(m)er	snapchater (v) snapchat(s) (n) > snap(s)	youtuber (v)
Intégration morphologie	facebookeur(euse)	twitteur(euse) tweeteur	instagram(m)eur(euse)	snapchat(t)eur(euse) snapchat(t)ien(ne)	youtubeur(euse) youtubing

[20] Notons toutefois le cas de *twitter / tweet,* qui, pour l'emprunt simple, est plus souvent non-adapté, mais qui l'est la plupart du temps dans ses versions dérivées (*twittos, twictée, twitonaute, twitcam,* etc.). Un autre exemple concerne les dérivés à base *instagram*, dont les dérivés connaissent deux graphies concurrentes : *instagrammeur, instagrameur.* Enfin, le pluriel des noms peut donner lieu à des variantes (*smartwatchs* ou *smartwatches*).

productive (affixes, fracto-lexèmes et formants savants)	facebookien(ne) facebooking anti-facebook facebookisme	(euse) re(-)tweeter tweeterisation tweeting anti-tweet demi-tweet non-tweet auto-tweet pseudo-tweet tweetesque tweetable tweetonade			

Tableau 6. Échantillon de dérivations attestées pour cinq bases nom propre issues des réseaux sociaux.

Si l'on s'intéresse aux emprunts à base *nom commun* (plus ou moins récents), on pourra faire les mêmes constats pour les emprunts les plus populaires (*blog, food, hashtag, check, shop, geek, market, game*, etc.). Par exemple, l'histoire de la pratique journalistique du *fact-checking* (terme dont les premières attestations en français datent de 1998, mais dont l'emploi connaîtra un premier pic avec les élections américaines de 2012 et un second, plus intense encore, avec celles de 2016 et en France en 2017) est éloquent : jusqu'aux élections américaines de 2012, les seuls emplois attestés sont *fact-checking,* avec recours quasi-systématique à la glose-traduction (par exemple : « Au printemps, le site va aussi s'allier avec d'autres médias afin de développer le "fact checking", une méthode répandue dans les pays anglo-saxons qui consiste à vérifier les chiffres et les affirmations des hommes politiques. », AFP, 18/02/2011). Puis des emplois non métalinguistiques apparaissent (« le fact-checking fait sa rentrée sur les ondes radio », *Libération*, 18/09/2012). Mais c'est seulement avec les élections américaines fin 2016 et les élections françaises en 2017 qu'une cohorte de dérivés fait son apparition (*fact-checker, factcheckeur, fact-checkings*) ainsi que, plus récemment encore, des composés sémantiquement liés (notamment le *fast-fact-checking* devenu *fast-checking*). Cependant, l'emploi verbal reste limité à l'infinitif, sauf timides exceptions (« les médias qui fact-checkeront les articles litigieux », *Libération*, 11/01/2017).

Enfin, dans le même ordre d'idée, la popularité d'une base lexicale empruntée se traduit également par la génération de composés et de fractocomposés. Toujours sur l'exemple *twitter*, nous relevons : *tweet-boomerang, tweet-choc, tweetosphère, tweet-série, feu-tweet, tweeteur-en-chef* ainsi que plusieurs autres formations directement empruntées : *tweetwall, acrostweet, live-tweet, tweetdeck, tweetstorm, fake-tweet, tweetbot, commander-in-tweet,* etc.

5.2.3. *Processus de mise en place d'un profil combinatoire*

La mise en place d'un *usage* des néologismes passe par l'abandon progressif (sauf visée didactique spécifique) des marques métalinguistiques qui accompagnent l'émergence des lexies. De ce point de vue, les données permettent d'observer le passage de l'un à l'autre. Prenons l'exemple de *ghosting*, cette pratique consistant, dans une relation amoureuse, à disparaître brusquement, sans plus répondre aux sollicitations du/de la partenaire. Le terme apparaît dans la presse américaine en 2014[21]. Très rapidement le terme se répand aussi en

[21] Nous prenons appui sur l'analyse faite sur Wikipédia et reprise par le Collins, qui a introduit le terme en 2015, et date l'émergence « écrite » du terme de l'article ci-après : https://jezebel.com/charlize-theron-broke-up-with-sean-penn-by-

français, avec dérivation morphologique (*ghoster qn, ghosteur (euse), ghosté(e)*, et intégration à la morphologie productive (*anti-ghosting*, *néo-ghosting*). Si l'on scrute les emplois de *ghosting*, dans le corpus Néoveille, sur 17 attestations (voir extraits tableau 7), on constate que les emplois métalinguistiques (guillemets, glose) tendent à disparaître, en tout cas dans la presse féminine et la presse magazine parisienne. Cette tendance est encore plus forte avec le verbe *ghoster*, apparu dans un second temps, et dont l'emploi transitif, également emprunté (to ghost someone > *ghoster + Nom ;* également emploi passif *être ghosté par Nom,* et factitif : *se faire ghoster par Nom*) et les contraintes argumentales sur l'objet (*personne, mec, type, Pauline,* etc.) dessinent très rapidement l'usage syntaxico-sémantique.

Date	Journal	Domaine	Extrait
24/05/18	*Elle*	Presse féminine	...Et dans le registre -ing de nos comportements amoureux, le ***ghosting*** demeure le plus célèbre : on disparaît sans un mot...
30/03/18	*Slate*	Général	...La pratique du ***ghosting*** – la rupture sans explication – en est le signe...
14/03/18	*Slate*	Général	...Le no-show, équivalent du ***ghosting*** mais version Guide Michelin C'est tellement simple que la...
26/12/17	*Nouvel Observateur*	Général	...Elle avait choisi à un moment le " ***ghosting*** ", c'est-à-dire de disparaître totalement dans une forme de...
25/10/17	*Le Progrès*	Général	...Inutile de préciser que ces cas de ***ghosting*** se produisent en grande partie suite à des relations...
15/10/17	*Nouvel Observateur*	Général	...me dissimule pas, je ne fais pas du ' ***ghosting*** ' (l'art de disparaître en pleine séduction)...
05/06/17	*Nouvel Observateur*	Général	Faut-il encore expliquer ce qu'est le ***ghosting***...
07/02/17	*Libération*	Général	...On avait déjà recensé le ghosting, qui consiste à disparaître sans donner de nouvelles...
13/10/16	*Cosmo*	Presse féminine	...Le ***ghosting*** est plus violent qu'une rupture amoureuse normale...
13/10/16	*Cosmo*	Presse féminine	...% des filles ont déjà vécu l'expérience charmante du ***ghosting***...
13/10/16	*Cosmo*	Presse féminine	...Dans le cas du ***ghosting***, le drapeau blanc persiste à flotter au vent...
01/09/16	*Cosmo*	Presse féminine	...Le benching, pourquoi est-ce pire que le ***ghosting***...

Tableau 7. Échantillon de contextes avec *ghosting* depuis début 2016.

Nous renvoyons également au tableau 3, qui détaille les collocations, collostructions et constructions syntaxiques que nous avons pu constater à partir de la lexie *food*.

5.2.4. *Évolution des contextes socio-pragmatiques des innovations*

ghosting-him-1712760688

L'exemple précédent illustre un autre phénomène : l'importance de l'inscription socio-pragmatique des innovations lexicales : *ghosting, ghoster* ne sont plus glosés dans la presse féminine, mais le sont encore dans la presse généraliste, montrant qu'ils restent, pour le grand public, en phase d'émergence. L'attestation d'un néologisme hors de son domaine d'émergence est donc un autre signe de sa diffusion. Dans Néoveille, les paramètres pour décrire ces propriétés sont encore trop grossiers (domaine, journal, pays de la source d'information), mais permettent néanmoins de constater des diffusions différenciées. Dans les figures 9 et 10, à titre d'illustration, nous pouvons visualiser la distribution temporelle (2016-2018) des domaines pour deux innovations lexicales, l'une datant des années 2000 (*smartphone*) et une autre bien plus récente (*smartwatch*) : on constate que la première a maintenant pénétré beaucoup de domaines, montrant un emploi non limité à un groupe socio-économique, tandis que la seconde, dont on peut constater l'émergence dans les domaines *hightech* et *informatique*, se diffuse maintenant dans la presse généraliste.

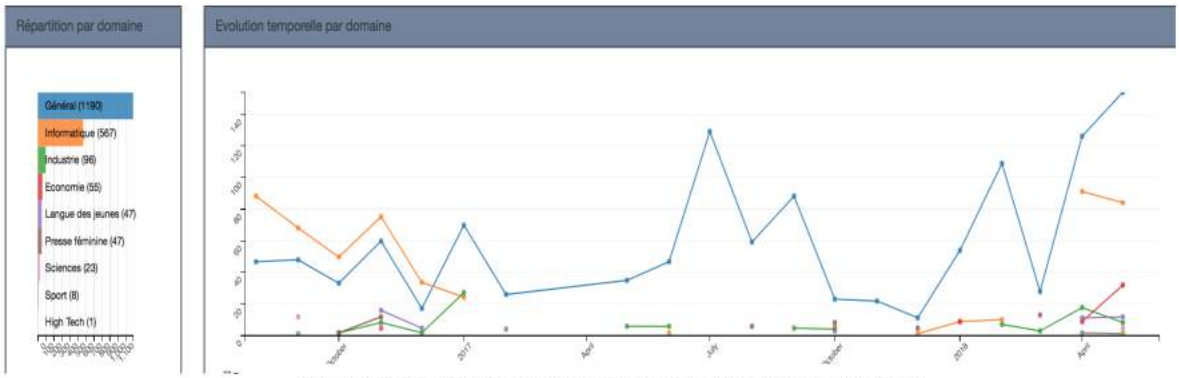

Figure 9. Distribution temporelle par domaine pour *smartphone*.

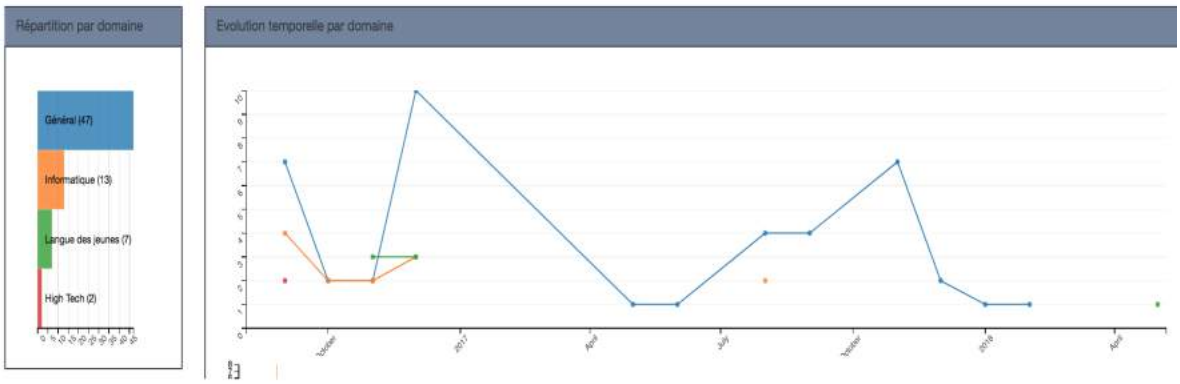

Figure 10. Distribution temporelle par domaine pour *smartwatch*.

L'observation des modifications de domaine fournit au moins deux autres informations : d'une part, elle permet de connaître les restrictions éventuelles de domaine d'application de la lexie (par exemple *phablet(te)*, apparu dans le domaine informatique mais pénétrant peu les autres domaines ; les termes des réseaux sociaux n'ont pas de telles limitations). D'autre part, elle permet d'identifier, dans le cadre de la théorie de l'innovation (Rogers 2010 [1962]), les groupes sociaux *innovateurs* et *diffuseurs* (adopteurs), grâce au suivi temporel de la distribution domaniale.

5.2.5. *De l'innovation lexicale à la productivité affixale : lexie, formant savant, fractolexème, affixe*

Les innovations lexicales ne se matérialisent pas seulement par des lexies, mais également par de nouveaux formants permettant de construire des lexies.

Pour les préfixations, le tableau 8 explicite les 41 préfixes les plus productifs[22] et le nombre de néologismes constatés dans notre corpus, pour le français.

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

Tableau 8. Liste des préfixes productifs, par ordre décroissant.

Dans ce tableau, 18 préfixes classiquement classifiés comme tel, 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

[22] 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.

préfixes produisant des noms, adjectifs et adverbes mais ce phénomène est lié à l'innovation lexicale produisant principalement noms et adjectifs.

On notera également 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.).

Pour la composition simple, on remarque la productivité de 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*), ainsi que de nouveaux patrons : *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 schèmes apparaissent. Notamment le schème 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-liker, prêt-à-agir,* etc.

Parmi la fracto-composition, on remarque que c*yber-, bio-, éco- et e-* sont les formants les plus productifs. *Ils représentent respectivement les lexies cybernétique, biologique, écologique et électronique* (tableau 9).

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

Tableau 9. Synthèse sur cyber-, e-, bio- et eco-

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

Comme on le constate, ces trois catégories (auxquelles il faudrait ajouter la composition savante, dans laquelle un certain nombre de formants savants sont également très productifs) fournissent des formants (savant, fracto-lexème, affixe) dont la productivité est évolutive et non-nécessairement ordonnée, puisque par exemple certains fractolexèmes sont plus productifs que certains préfixes. Ces constats statistiques requièrent de préciser les critères définitoires pour chacune des catégories, d'une part, et sans doute de considérer que ces catégories sont dans un continuum. Le phénomène de troncation, de ce point de vue, pourrait être considéré comme le principe permettant le passage de l'un à l'autre. On le constate par exemple à partir de l'emprunt *instagram*, qui donne lieu, depuis quelque temps, à des formations composées sur la base de son troncat *insta* (*instagirl, instashop, instapreneur,* etc;) : Quoi qu'il en soit, les données statistiques qui peuvent être extraites des descriptions linguistiques fournissent un matériau de choix pour appuyer tel ou tel modèle.

Conclusion

Dans cet article, nous avons présenté la plateforme Néoveille (et ses différents modules) qui permet de détecter semi-automatiquement, de décrire linguistiquement et socio-pragmatiquement puis de suivre le cycle de vie des néologismes sur un corpus dynamique

contemporain. Nous avons présenté pas à pas ces différents modules, en illustrant les résultats auxquels nous sommes parvenus à ce jour.

Il nous semble essentiel que la linguistique, pour la qualité et l'objectivité de ses travaux, accorde aux données numériques actuelles, et aux traitements automatiques qui peuvent s'y appliquer, toute leur place : nous pensons avoir montré que les tendances néologiques globales d'une langue donnée dans une période donnée peuvent être établies à partir de ces données massives, d'une part, et que la description des propriétés linguistiques et socio-pragmatiques des innovations lexicales, et leur évolution, gagnait également à une formalisation dans un système automatique, qui, en retour, donne une matière à la nécessaire analyse linguistique. Il ne s'agit pas d'affirmer la mainmise autoritaire des approches de la linguistique de corpus ou du TAL, car les données doivent toujours être interprétées, et les modèles peuvent être révisés, comme nous l'avons évoqué notamment pour ce qui concerne la caractérisation socio-pragmatique des sources d'information. Mais il s'agit d'asseoir l'analyse linguistique sur les innombrables réalisations discursives et leurs différentes caractéristiques.

Références bibliographiques

BAAYEN, R. Harald (2009), « Corpus linguistics in morphology: morphological productivity », dans Lüdeling, Anke / Kytö, Merja, *Corpus linguistics. An international handbook*, p. 900-919.

BARONI Marco et LENCI Alessandro (2010), « Distributional memory: A general framework for corpus-based semantics », *Computational Linguistics*, 36(4), p. 673–721.

BARONI Marco. et BERNARDINI Silvia, FERRARESI Adriano et ZANCHETTA Eros (2009), « The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora », *Language Resources and Evaluation* 43(3): p.209-226

CARTIER Emmanuel et SABLAYROLLES Jean-François (2009), « Néologismes, Dictionnaires et Informatique », *Les Cahiers de Lexicologie*, n° 93, 2008- 2, p. 175-192.

CARTIER Emmanuel (2016), « Néoveille, système de repérage et de suivi des néologismes en sept langues », *Neologica 10*, p. 101-131.

CARTIER Emmanuel, SABLAYROLLES Jean-François, BOUTMGHARINE Najet, HUMBLEY John, BERTOCCI Massimo, JACQUET-PFAU Christine, KÜBLER Natalie et TALLARICO Giovanni (2018), « Détection automatique, description linguistique et suivi des néologismes en corpus : point d'étape sur les tendances du français contemporain », Actes du Congrès Mondial de Linguistique Française, Mons (Belgique), 9-13 juillet 2018, 20 p.

CARTIER Emmanuel et VIAUX Julie (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.

CARTIER Emmanuel (2018a), « Emprunts en français contemporain : étude linguistique et statistique à partir de la plateforme Néoveille », dans *Emprunts en question(s)*, Kacprzak, A. ; Mudrochová, R. ; Sablayrolles, J.-F. (éds), La Lexicothèque, Limoges, Lambert-Lucas, 27p.

CARTIER Emmanuel (2018b), « Noms propres et innovation lexicale : étude linguistique et statistique à partir de Néoveille », *Cahiers de Lexicologie*, n°113, 2018-2, Néologie et noms propres, p. 203-224

CARTIER Emmanuel (2018c), *Dynamique lexicale des langues : éléments théoriques, méthodes automatiques, expérimentations en français contemporain*, document inédit HDR, 207 p., url : https://lipn.univ-paris13.fr/neoveille/html/data/ecartier/ecartier_inedit_final_09122018.pdf

CARTIER Emmanuel, GALAND Loïc, STIRLING Peter, AUBRY Sara (2018), « Néonaute: mining web archives for linguistic analysis », International Internet Preservation Consortium Web Archiving Conference, Wellington, 12-15 nov. 2018.

CHARAUDEAU Patrick (1995), « Une analyse sémiolinguistique du discours ». *Langages*, p. 96–111.

CORBIN Danielle (1987), *Morphologie dérivationnelle et structuration du lexique*, 2 vol., Tübingen, Max Niemeyer Verlag.

DAL Georgette (2003), « Productivité morphologique: définitions et notions connexes », *Langue française*, p. 3–23.

DUBOIS Jacqueline (1962), *Étude sur la dérivation suffixale en français moderne et contemporain,* Librairie Larousse.

FIRTH John Ruppert (1957), *Papers in Linguistics 1934–1951,* Oxford University Press.

GREZKA Aude, CARTIER Emmanuel ET MATHIEU-COLAS Michel (2015), « Dictionnaires morphologiques du français contemporain : présentation de Morfetik, éléments d'un modèle pour le TAL », *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN)*, Caen, France.

GRIES Stefan Th. (2010), « Behavioral profiles : A fine-grained and quantitative approach in corpus-based lexical semantics », *The Mental Lexicon*, 5(3), p. 323–346.

HAMILTON William L., LESKOVEC Jure, et JURAFSKY Dan (2016), « Cultural shift or linguistic drift ? Comparing two computational measures of semantic change », *ACL 2016.*

HARRIS Zellig Sebbataï (1954), « Distributional structure », *Word*, 10(2-3), p. 146–162.

HYMES Dell (1974), *Foundations in Sociolinguistics: An Ethnographic Approach,* Philadelphia: University of Pennsylvania Press.

LEJEUNE Gaël, CARTIER Emmanuel (2017), « Character Based Pattern Mining for Neology Detection », *Proceedings of the First Workshop on Subword and Character Level Models in NLP , EMNLP 2017*, Copenhagen, p.25-30.

MICHEL Jean-Baptiste, SHEN Yuan Kui, AIDEN Aviva Presser, VERES Adrian, GRAY Matthew K., The Google Books Team (2010), « Quantitative Analysis of Culture Using Millions of Digitized Books », *Science*, 14, vol. 331-6014, p.176-182

MIKOLOV Tomas, YIH Wen-Tau, et ZWEIG Geofrrey (2013), Linguistic regularities in continuous space word representations, In *Proceedings of HLT-NAACL*, volume 13, p. 746–751.

ROGERS Everett M. (2010), *Diffusion of innovations*, fourth edition [1962], Simon and Schuster.

SABLAYROLLES Jean-François et PRUVOST Jean (2016), *Les néologismes,* Presses Universitaires de France-PUF, collection Que sais-je ?

SAUGERA Valérie (2017), *Remade in France : Anglicisms in the lexicon and morphology of French,* New York, NY : Oxford University Press, [2017]

SCHMID Hans-Jörg (2015), « A blueprint of the entrenchment and conventionalization model », *Yearbook of the German Cognitive Linguistics Association*, 3(1), p. 1–27.

SIEPMANN Dirk, BÜRGEL Christoph et DIWERSY Sascha (2016), « Le Corpus de référence du français contemporain (CRFC), un corpus massif du français largement diversifié par genres », *SHS Web of Conferences*, 27 (2016) 11002

STEFANOWITSCH Anatol et GRIES Stefan Th. (2003), « Collostructions: Investigating the interaction of words and constructions », *International Journal of Corpus Linguistics*, 8(2), p. 209–243.