Dossier en vue de l'habilitation à diriger des recherches : inédit

---

Dynamique lexicale des langues : éléments théoriques, méthodes automatiques, expérimentations en français contemporain

Emmanuel Cartier, LIPN équipe RCLN UMR7030 CNRS

Soutenance le 13 décembre 2018

Jury:
Jean-François Sablayrolles, Professeur émérite, Sorbonne Paris Cité (garant)
Richard Huyghe, Professeur, Université de Fribourg (rapporteur)
Dominique Legallois, Professeur, Université de Paris 3 (rapporteur)
Alessandro Lenci, Professeur, Université de Pise (rapporteur)
Esme Winter-Froemel, Professeure, Université de Trier (rapporteure)
Viviane Arigne, Professeure, Université de Paris 13
Thierry Charnois, Professeur, Université de Paris 13

Table des matières

Table des figures

Liste des tableaux

Introduction

Le présent travail propose une modélisation du phénomène de *changement lexical*, en vue d'une détection automatique des néologismes formels et sémantiques, et du suivi du cycle de vie des lexies en corpus. Il débouche sur un travail de description des néologismes en français contemporain, en s'appuyant sur les données fournies par la plateforme Néoveille.

Il s'agit d'un travail appartenant au champ du *changement linguistique*, plus spécifiquement du *changement lexical*, et je m'intéresserai précisèment à l'*innovation lexicale* ou *néologie*. Ce phénomène est l'une des manifestations de la vie des mots, et il est probable que la création lexicale, la diffusion, l'adoption, la préservation, la dégénérescence d'un usage et la disparition des unités lexicales font toutes appel aux mêmes ressorts. Même si nous concentrerons notre effort sur l'émergence et la diffusion de nouvelles lexies et de nouveaux usages, dans le présent document, ce qui nous intéressera au premier chef concerne les mécanismes sous-jacents au changement lexical, en considérant le dynamisme comme une propriété *intrinsèque* des langues.

Importance du changement lexical

L'importance du changement lexical est très souvent minimisé : il est vrai que dans les situations de communication auxquelles nous sommes confrontés quotidiennement, l'impression de nouveauté est assez rare : nous parlons sans y penser, nous comprenons sans avoir à nous arrêter pour interpréter des mots ou des tournures que nous entendons. Il semblerait que le changement linguistique dans son ensemble soit un épiphénomène dans l'économie générale des langues. Pourtant, les états de langue antérieurs sont bien différents de ceux que nous utilisons aujourd'hui. Il nous arrive également de nous arrêter pour ré-analyser un mot entendu ou rire d'une tournure inattendue. Ce paradoxe entre l'apparente stabilité des langues et l'évidence de la créativité lexicale est généralement expliqué par le fait que les changements linguistiques se dérouleraient de manière insensible et continue, dans une durée qui n'est pas celle de la vie humaine individuelle.

Allons plus loin, car le nombre de situations de communication où l'impression de nouveauté nous saisit est plutôt vaste : lorsque nous rencontrons ou entendons des personnes qui prononcent les mots de manière différente, parce qu'ils ne sont pas natifs ; lorsque nous rencontrons un ami ou une connaissance qui aime à détourner le sens des mots ; lorsque nous entendons des personnes marquées sociologiquement, et qui em-

ploient un vocabulaire et des tournures spécifiques inattendus pour nous, lorsque nous nous trouvons dans un contexte professionnel nouveau, et qu'un vocabulaire particulier est utilisé ; et plus généralement quand un nouvel objet ou une nouvelle pratique sont dénommés, ou que nous devons inventer un mot ou une expression pour désigner une situation, un objet que nous ne savons pas nommer sans avoir à y penser. Ces situations sont plus fréquentes qu'il n'y parait pour chacun d'entre nous, mais nous les minimisons généralement, en les attribuant à des particularismes de tel ou tel groupe social, à un trait de caractère personnel, à une situation exceptionnelle.

Le vocabulaire est dans doute le lieu des plus grandes variations et évolutions linguistiques. Cela n'est pas étonnant, puisque ces nouveautés n'ont un effet que très limité sur le système dans son entier : un mot forgé trace les frontières d'une réalité conceptuelle limitée, et son impact sur le système entier est faible. Mais pourtant, ces situations se produisent de façon continue et avec une fréquence non négligeable, et au final, l'ensemble de ces événements a des chances de modifier, à un moment ou à un autre, des pans entiers du système et de recomposer son équilibre général. On peut parler d'un système dynamique, d'une créativité continue articulée sur un système d'une nécessaire stabilité, pour la bonne communication entre les membres de la communauté.

Le changement lexical est une chose naturelle : notre environnement est lui-même à la fois stable et instable, fait de la répétition des mêmes événements (le soleil se lève, se couche, nous marchons sur nos deux jambes, les saisons se succèdent, nous voyons par les yeux, etc.) et en même temps cet environnement est toujours nouveau (les journées sont toutes différentes, des événements surviennent, nous avons de nouveaux sentiments et de nouvelles idées). Les langues sont là pour nous permettre de dire tout cela, qui est fait de répétition et de nouveauté. Le changement lexical est également nécessaire à la vie des langues : une langue morte est une langue qui n'a plus de membres qui l'utilisent et la transmettent et l'adaptent au monde extérieur et aux évolutions qui s'y produisent.

Mais quelle est exactement la place de l'innovation lexicale dans l'économie générale des langues ? Quels sont les procédés disponibles pour créer de nouveaux mots ? Y a-t-il des situations ou des contextes plus favorables que d'autres pour leur émergence ? Comment les nouveautés de l'un se diffusent-elles éventuellement à toute la communauté ? Quels sont les procédés d'innovation lexicale les plus productifs aujourd'hui ?

Objectifs du présent travail

L'objectif de ce travail sera triple : d'abord, mieux comprendre le phénomène de changement lexical, d'un point de vue linguistique, et proposer un modèle général permettant de l'expliquer et de le décrire ; ensuite, proposer un modèle opérationnel pour l'automatisation d'un certain nombre de tâches : détection automatique en corpus de la néologie formelle et de la néologie sémantique d'une part ; suivi du cycle de vie des lexies, des points de vue de leur émergence, de leur diffusion et de leur éventuelle lexicalisation, d'autre part. De ce point de vue automatique, nous nous appuierons largement sur le travail que nous avons mené dans le projet Néoveille (Cartier, 2016b; Cartier, 2017a), même si nous adopterons une posture prospective dans biens des cas. Enfin, il s'agira

aussi de rendre compte des tendances néologiques du français contemporain, en nous appuyant sur les données collectées par un groupe de linguistes sur cette même plateforme pendant près de trois ans maintenant.

Plan du présent travail

Le plan du présent travail découle de ce triple objectif : modélisation du changement lexical ; modélisation opérationnelle, automatisable des phénomènes d'émergence et de diffusion des innovations lexicales, description linguistique du changement lexical en français contemporain.

Nous aborderons dans une **première partie, *Modèles théoriques***, la problématique du changement lexical en essayant de situer ce phénomène dans l'économie générale des langues.

Dans le **chapitre 1, *La langue comme Energéia***, nous défendrons une conception dynamique des langues, en partant des concepts fondateurs de la linguistique moderne établis par la linguistique saussurienne : langue / parole, synchronie /diachronie, linguistique interne / linguistique externe, ainsi que la notion de signe linguistique. Il s'agira ici pour nous, étant donné la place congrue laissée au dynamisme des langues par Saussure - en tout cas du Saussure du Cours de Linguistique Générale (CLG) -, de revisiter ces notions en nous appuyant sur des travaux linguistiques plus récents, et notamment les travaux fondateurs de Coseriu et Weinreich, les travaux de la linguistique cognitive, des grammaires de construction et de la sociolinguistique. Le modèle que nous proposerons s'appuiera sur quelques hypothèses qui ont été notamment énoncées par ces courants linguistiques. Ce modèle aboutira notamment à la conclusion que l'étude des changements lexicaux (et linguistiques en général) nécessite minimalement une collaboration entre trois disciplines : la linguistique proprement dite, la sociolinguistique et la linguistique cognitive.

le **chapitre 2, *Unité lexicale et innovation lexicale*** est consacré aux notions d'unité lexicale et d'innovation lexicale. En partant des définitions proposées par les grammaires classiques, nous étudions les points communs et les différences entre l'unité lexicale (le morphème libre) au sens classique et les formants plus petits (morphèmes liés), d'une part, et entre l'unité lexicale et les unités plus grandes (unités polylexicales et constructions syntactico-sémantiques et syntaxiques), d'autre part. Nous arrivons à la conclusion que la notion de construction proposée par les grammaires du même nom, est la plus à même de rendre compte du continuum qui existe entre ces différents unités, qui ne peuvent être distinguées que sur la base de propriétés typiques. Nous évoquons rapidement les propriétés essentielles de ces unités constructionnelles. Puis nous évoquons les caractéristiques principales des innovations lexicales : paire forme-sens, déviation par rapport à une norme linguistique, périmètre des innovations du point de vue des notions d'émergence, de diffusion et d'adoption, typologie des procédés néologiques, et cycle de vie.

Le **chapitre 3, *Langue, variations, variétés***, évoquera les notions de variations

et de variétés, qui constituent la facette synchronique du changement linguistique. Il s'agira tout d'abord de montrer l'existence des variations et des variétés, puis de définir ces notions et la re-conceptualisation de la notion de langue qu'elles impliquent. En nous appuyant sur les travaux fondateurs de Weinreich et surtout de Coseriu, nous détaillerons les dimensions diatopiques, diastratiques et diaphasiques de la variation, en présentant les principaux travaux produits par la sociolinguistique pour préciser ces notions et les appliquer à des données réelles. Nous présenterons ensuite le travail de Koch et Oesterreicher (Koch et Oesterreicher, 1985; Koch et Oesterreicher, 2001; Koch et Oesterreicher, 1990) sur la proximité/distance communicative, une dimension complémentaire pour comprendre les variations. Nous invoquerons différents arguments pour la positionner différemment de ce qui est proposé par les auteurs eux-mêmes. Nous reviendrons alors sur une approche de la variation qui se base sur l'étude des flux de communication entre individus et les réseaux sociaux qu'ils construisent, et permet d'étudier finement le cheminement des variations et des changements linguistiques. Une conclusion tentera de faire une synthèse de ces différentes approches.

Avec ces différents éléments, nous abordons la **seconde partie, Modèles opérationnels**, consacrée à la construction d'une plateforme pour la détection, l'analyse linguistique et le suivi diachronique des innovations lexicales en corpus dynamique (**chapitre 4**) et à la détection automatique des néologismes formels (**chapitre 5**).

Le **chapitre 4, *Plateforme pour l'étude des néologismes en corpus*** détaille tout d'abord les caractéristiques idéales d'une plateforme permettant la détection, l'analyse linguistique et le suivi diachronique des innovations lexicales en corpus dynamique, en partant de l'analyse des outils existants, d'une collecte des besoins linguistiques pour l'analyse et la compréhension du phénomène d'innovation lexicale et en considérant la nécessaire collaboration entre les processus automatiques et l'intervention experte des opérateurs humains. Puis nous présentons la plateforme Néoveille, l'état actuel des développements et les modules disponibles. Nous terminons par les perspectives de développement pour améliorer l'existant.

Le **chapitre 5, *Repérage et suivi des changements lexicaux : néologismes formels*** comprend deux sections : la première section délimite le champ de la néologie formelle, en traçant d'abord ses frontières avec la néologie sémantique et en explicitant ses principaux procédés : emprunts lexicaux, dérivation, composition et réductions/-transformations. Puis nous tentons de définir plus précisèment chacun de ces procédés, en commençant par indiquer les frontières poreuses entre flexion et dérivation d'une part, et composition, formations polylexicales d'autre part. Nous présentons la typologie générale des néologismes proposée par Jean-François Sablayrolles (Sablayrolles et Pruvost, 2016), en mettant l'accent sur ses avantages par rapport aux typologies classiques proposées par les morphologues et en proposant quelques aménagements. La seconde section évoque la détection automatique des néologismes formels, les méthodes existantes, la méthode actuelle implémentée dans Néoveille, ses limites et une piste de travail pour mettre en place un système plus efficace.

Avec ces développements en traitement automatique des langues, nous arrivons à la

dernière partie, Application, consacrée à la description des néologismes du français contemporain. Le travail présenté est lié aux travaux menés autour de la plateforme Néoveille par un groupe de linguistes travaillant sur le français et d'autres langues.

Le **chapitre 6, *Tendances néologiques du français contemporain (2015-2018)*** détaille les travaux menés sur le français qui ont abouti à identifier environ 20 000 néologismes formels sur la période dans la presse généraliste en ligne. Dans une première section, nous présentons le corpus, la méthodologie de validation des néologismes détectés automatiquement puis la grille descriptive. Dans la seconde section, nous présentons alors les grandes tendances du français contemporain, la répartition par mécanisme, la distribution entre hapax néologique et les néologismes à faible ou forte diffusion, la distribution par pays, par domaine, par journaux et par parties du discours.

Dans le **chapitre 7, *Dérivation en français contemporain (2015-2018)***, nous nous intéressons spécifiquement aux phénomènes de préfixation et de suffixation. Après une section définitoire, nous détaillons les enseignements tirés par la collecte et l'analyse d'environ 13 000 créations affixales,

Dans le **chapitre 8, *Composition en français contemporain (2015-2018)***, nous nous intéressons spécifiquement au phénomène de composition.

Dans le **chapitre 9, *Emprunts en français contemporain (2015-2018)***, nous nous intéressons spécifiquement au phénomène d'emprunt lexical.

Enfin, une conclusion prospective tente de faire une synthèse de l'ensemble du travail, en proposant quelques éléments théoriques sur la modélisation des langues, et quelques éléments méthodologiques permettant une automatisation de la détection et du suivi des changements lexicaux, d'une part, et une description linguistique plus systématique, d'autre part.

Première partie

Modèles théoriques

Résumé

Dans cette partie nous abordons l'innovation lexicale d'un point de vue théorique, dans le but de délimiter et modéliser notre champ d'étude.
Pour ce faire, nous examinons dans le **chapitre 1** la place de l'innovation lexicale dans l'économie générale des langues : considérée comme accessoire et accidentelle dans la linguistique structuraliste (Saussure, 1916), elle devient l'un des composants essentiels des langues à partir des travaux de (Coseriu, 1952) et de (Weinreich *et al.*, 1968). Dans cette conception dynamique des langues, les articulations langue/discours et synchronie/diachronie doivent être revisitées, le discours étant le lieu même de la préservation et du changement continu des langues, et les langues étant elles-mêmes un potentiel de réalisations plutôt qu'une structure figée d'éléments et de règles de leur combinaison. Nous insistons également sur les trois perspectives complémentaires qui doivent être adoptées pour rendre compte du dynamisme intrinsèque des langues : le point de vue linguistique, le point de vue sociolinguistique et le point de vue cognitif.
Le **chapitre 2** détaille une conception des unités lexicales qui rend compte de la multiplicité des paires forme-sens et de la continuité qui existent entre elles : l'unité lexicale se manifeste non seulement au niveau de l'unité lexicale au sens « classique », mais également, en deçà du mot simple, dans les morphèmes liés, et, au-delà du mot, dans les unités polylexicales, les constructions lexico-syntaxiques et jusqu'aux constructions syntaxiques. Nous établissons certaines propriétés prototypiques de ces variétés de paires forme-sens, montrant qu'il existe un continuum indéniable entre chacune des catégories. Puis, dans ce cadre, nous évoquons les propriétés distinctives des innovations lexicales, notamment la variation par rapport à l'usage d'une communauté linguistique, et fixons le périmètre qui sera le nôtre dans la suite de l'étude.
Le **chapitre 3** aborde les relations entre langue et société, en focalisant sur les notions de variation, le pendant synchronique du changement lexical, et de variété et leur articulation avec la notion de langue. Considérée dans la tradition structurale comme unique pour une communauté linguistique donnée, la réalité démontre non pas l'existence d'une langue, mais l'existence de variations et de variétés au sein d'un même communauté linguistique. A l'aide des travaux développés notamment par la sociolinguistique, nous détaillons les différentes manières de caractériser ces variations et ces variétés, qui sont autant de moyens de caractériser l'émergence et la diffusion des innovations.

Chapitre 1

Dynamique des langues

Sommaire

Pour approcher l'innovation lexicale, il est tout d'abord nécessaire de caractériser le changement linguistique dans l'économie générale des langues. La conception structuraliste rejette le changement linguistique dans la diachronie et dans les discours, comme une série d'événements accidentels hors du champ de la linguistique. Cette position n'est pas satisfaisante, et nous reprendrons à notre compte une vision de la langue comme *energéia*, énergie, qui considère les systèmes linguistiques comme des systèmes dynamiques, qui redonne au changement linguistique toute sa place. Une fois posé ce modèle général, il nous faudra détailler la notion d'unité lexicale et ses caractéristiques, puisque les innovations lexicales sont d'abord des unités lexicales et héritent de leurs propriétés. Ensuite, nous délimiterons la notion d'innovation lexicale, en insistant sur ses propriétés distinctives.

Le propos sera donc organisé en trois sections : la première détaille les hypothèses générales concernant le fonctionnement des langues, en partant des grandes oppositions proposées par Saussure et en aboutissant à une conception des langues comme systèmes dynamiques et adaptatifs. La seconde partie traitera des unités lexicales, et explicitera une conception proche de celle des grammaires de construction. Puis nous aborderons la notion d'innovation lexicale : après une définition générale, nous présenterons un certain nombre de paramètres permettant de les décrire.

Une conclusion fera la synthèse de ce parcours.

1.1 Conception structuraliste

La linguistique structurale [1] considère que l'objet véritable de la science *linguistique* est la langue. Le chapitre 3 du CLG explicite l'essentiel des arguments en faveur de cette position.

D'une part, il y a le constat que le phénomène linguistique présente toujours deux aspects indissociables, l'un individuel, l'autre social : « le langage implique à la fois un système établi et une évolution ; à chaque moment il est une institution actuelle et un produit du passé »(Saussure, 1916, p.15). le linguiste suisse insiste également sur la multiplicité des points de vue qu'on peut adopter sur le langage : psychologique, anthropologique, sociologique etc.

Afin de mieux délimiter l'objet de la linguistique, l'auteur identifie la *langue* et la *parole* : « C'est à la fois un produit social de la faculté du langage et un ensemble de conventions nécessaires, adoptées par le corps social pour permettre l'exercice de cette faculté chez les individus »(Saussure, 1916, p.15). Cette langue, *institution sociale*, s'oppose à l'exercice de cette faculté chez les individus, représenté par la parole. L'opposition social/individuel est encore renforcée par une autre opposition : la langue constitue un *système de signes distincts correspondant à des idées distinctes* (Saussure, 1916, p.26), c'est-à-dire une structure stable, essentielle, qui s'oppose aux événements de la parole, qui sont accidentels et intrinsèquement soumis aux hasards des situations de communications. La langue est ailleurs comparée à un jeu d'échecs avec ses pièces et les règles de leur déplacement : un système linguistique est composé de signes linguistiques (association arbitraire - c'est-à-dire normée -, entre signifiant et signifié) et des règles de leur combinaison (phonologie, morphologie, grammaire). L'articulation langue / parole est ainsi dégagée de l'individu et rattachée à la « masse parlante » : « C'est un trésor déposé par la pratique de la parole dans les sujets appartenant à une même communauté, un système grammatical existant virtuellement dans chaque cerveau, ou plus exactement dans les cerveaux d'un ensemble d'individus ; car la langue n'est complète dans aucun, elle n'existe parfaitement que dans la masse. »(ibid, p.26)

Il insiste pourtant sur l'articulation intrinsèque entre langue et parole :

> « Sans doute, ces deux objets sont étroitement liés et se supposent l'un l'autre : la langue est nécessaire pour que la parole soit intelligible et produise tous ses effets ; mais celle-ci est nécessaire pour que la langue s'établisse ; historiquement, le fait de parole précède toujours. Comment s'aviserait-on d'associer une idée à une image verbale, si l'on ne surprenait pas d'abord cette association dans un acte de parole ? D'autre part, c'est en entendant les autres que nous apprenons notre langue maternelle ; elle n'arrive à se déposer dans notre cerveau qu'à la suite d'innombrables expériences. Enfin,

1. Nous parlons ici de la conception initiale saussurienne telle qu'elle a été transmise par (Saussure, 1916), correspondant à ce qu'on peut appeler le structuralisme issu du CLG. D'autres manuscrits retrouvés ultérieurement montrent un Saussure bien plus nuancé, et qui rétablit l'importance de la parole - plutôt sous le terme de discours (voir (Turpin, 1995; Puech, 2005; Testenoire, 2015; Saussure, 2002). Notons également que Bally indique, en marge du texte de 1916, que la linguistique de la parole devait faire l'objet du troisième cours de Saussure, qui n'aura pas lieu suite à son décès.)

> c'est la parole qui fait évoluer la langue : ce sont les impressions reçues en entendant les autres qui modifient nos habitudes linguistiques. Il y a donc interdépendance de la langue et de la parole ; celle-là est à la fois l'instrument et le produit de celle-ci. Mais tout cela ne les empêche pas d'être deux choses absolument distinctes. »(ibid, p.26)

Et Saussure conclut le chapitre, en ajoutant deux concepts, la *synchronie* et la *diachronie* :

> « Ainsi la linguistique se trouve ici devant sa seconde bifurcation. Il a fallu d'abord choisir entre la langue et la parole ; nous voici maintenant à la croisée des routes qui conduisent l'une, à la diachronie, l'autre à la synchronie.
>
> Une fois en possession de ce double principe de classification, on peut ajouter que *tout ce qui est diachronique dans la langue ne l'est que par la parole*. C'est dans la parole que se trouve le germe de tous les changements : chacun d'eux est lancé d'abord par un certain nombre d'individus avant d'entrer dans l'usage. L'allemand moderne dit : *ich war, wir waren*, tandis que l'ancien allemand, jusqu'au XVIe siècle, conjuguait : *ich was, wir waren* (l'anglais dit encore : *I was. we were*). Comment s'est effectuée cette substitution de *war* à *was* ? Quelques personnes, influencées par *waren*, ont créé *war* par analogie ; c'était un fait de parole ; cette forme, souvent répétée, et acceptée par la communauté, est devenue un fait de langue. Mais toutes les innovations de la parole n'ont pas le même succès, et tant qu'elles demeurent individuelles, il n'y a pas à en tenir compte, puisque nous étudions la langue ; elles ne rentrent dans notre champ d'observation qu'au moment où la collectivité les a accueillies.
>
> Un fait d'évolution est toujours précédé d'un fait, ou plutôt d'une multitude de faits similaires dans la sphère de la parole ; cela n'infirme en rien la distinction établie ci-dessus, elle s'en trouve même confirmée, puisque dans l'histoire de toute innovation on rencontre toujours deux moments distincts : 1/ celui où elle surgit chez les individus ; 2/ celui où elle est devenue un fait de langue, identique extérieurement, mais adopté par la collectivité. »(ibid, p. 26)

On voit bien ici que Saussure distingue fermement les événements linguistiques (les paroles) et les états linguistiques (la langue), en rejetant la parole hors du champ de la linguistique.

La dichotomie langue-parole détermine également deux autres dichotomies liées aux perspectives pouvant être prises pour étudier les langues :

— **linguistique diachronique / linguistique synchronique** : soit on prend le point de vue historique, et alors on s'intéressera aux états de langue successifs et aux changements linguistiques qui sont survenus dans le temps, soit on s'intéressera à un état spécifique de langue, et on étudiera le système que constitue cet état. « Sur ce point, il est évident que l'aspect synchronique prime l'autre, puisque pour la masse parlante il est la vraie et la seule réalité. Il en est de même

pour le linguiste : s'il se place dans la perspective diachronique, ce n'est plus la langue qu'il aperçoit, mais une série d'événements qui la modifient. »(ibid, p.95) En complément à ces deux arguments, Saussure considère également que l'étude synchronique est la seule qui nous permette une généralisation, une linguistique générale, puisqu'au fond, chaque langue particulière matérialise le langage en tant que faculté humaine générale : « La linguistique synchronique s'occupera des rapports logiques et psychologiques reliant des termes coexistants et formant système, tels qu'ils sont aperçus par la même conscience collective. La linguistique diachronique étudiera au contraire les rapports reliant des termes successifs non aperçus par une même conscience collective, et qui se substituent les uns aux autres sans former système entre eux. »(ibid, p.95)

— **linguistique interne / linguistique externe** : par ailleurs, on peut étudier les langues de plusieurs points de vue : un point de vue psychologique, en s'intéressant aux mécanismes cognitifis sous-jacents à l'utilisation des langues par les individus ; un point de vue sociologique, puisque chaque langue est une *norme sociale*, et il est donc possible de rendre compte de ses caractéristiques socio-linguistiques ; un point de vue purement linguistique, en ne considérant que la langue en elle-même, en dehors des circonstances de son exploitation dans les paroles. Là encore, Saussure considère que, afin de se constituer comme science indépendante, la linguistique doit uniquement prendre un point de vue interne.

1.2 Critique de la conception structuraliste

La dichotomie *langue / parole* a rapidement été remis en cause par les successeurs de Saussure : Benveniste avec la notion d'énonciation (Benveniste, 1970), Bahktine avec la linguistique du texte (Bakhtine, 1978), la philosophie analytique américaine (Austin, 1970; Searle, 1972) donnant naissance à la pragmatique, enfin les premiers travaux de (Coseriu, 1962; Coseriu, 1964; Coseriu, 1958; Coseriu et Polo, 1986; Coseriu, 1980) puis (Weinreich *et al.*, 1968).

Nous reprenons ici la critique de la conception structuraliste et la conception alternative proposée dans (Coseriu, 1980).

Selon Saussure, le primat de la synchronie est soutenu par deux arguments principaux : d'une part, le fait « que c'est uniquement en synchronie que la langue est envisageable comme un tout, comme un système, et d'autre part, que la synchronie, l'état de langue, est la seule réalité pour le locuteur. »(Coseriu, 1980, p.3) Le linguiste roumain affirme au contraire que l'individu qui parle ne saisit jamais qu'un système linguistique partiel, celui lié à ce qu'il dit, et qui ne saurait correspondre au système linguistique dans son entier. Saussure parle aussi d'une "conscience collective", ce que récuse Coseriu, considérant que celle-ci n'est qu'une dimension de la conscience individuelle. Il affirme ainsi :

> « il n'y a guère de locuteur qui se trouvé confronté à un système linguistique homogène et unique. (...) le locuteur (...) se trouve confronté, dans son expérience réelle, à l'état d'une langue historique, dont la synchronie est

> différenciée des points de vue diatopique, diastratique et diaphasique. Tout locuteur, s'il ne connaît pas la langue historique dans son ensemble, connaît, au moins jusqu'à un certain degré, plus d'un dialecte et plus d'un niveau de langue ; et tout locuteur maîtrise plusieurs styles de langue. »(ibid, p.4-5)

Coseriu remet également en cause la langue comme un état, mais considère au contraire qu'il s'agit d'un « savoir orienté vers l'avenir, et par là, quelque chose de potentiellement dynamique. »(ibid, p.5) Cela provient du fait que les langues, au final, sont des systèmes de règles plutôt que des réalisations concrètes, un potentiel plutôt que des réalisations, qui, elles, sont du ressort de la parole.

Deux autres arguments de Saussure à l'encontre de la diachronie sont liés d'une part à son caractère événementiel, qui ne peut donc de ce fait constituer un système, et, d'autre part, l'idée que le changement linguistique se produit fondamentalement en dehors du système de la langue.

A cela, Coseriu répond par un premier constat, celui de la continuité de la langue dans le temps : « une caractéristique saillante des langues par rapport à d'autres traditions collectives, est plutôt que ce sont des traditions tellement figées, c'est-à-dire qu'elles se transmettent en principe sans mutations profondes (ce qui est précisément une condition nécessaire aux thèses de F. de Saussure), et qu'une accélération du changement ne se produit que dans des circonstances historiques particulières. »(ibid, p.6). Coseriu pointe également sur la conception erronée du changement linguistique comme événement extérieur à la langue, dans la parole. À l'évidence, « toutes les étapes du changement linguistique (adoption, sélection, mutation), y compris la première (innovation), ont lieu dans la langue en tant que telle, parce qu'il s'agit précisément du changement linguistique, et non pas simplement du changement dans la parole (la parole en tant que telle ne saurait d'ailleurs changer, car elle ne possède pas de continuité historique). »(ibid, p.7) Le changement linguistique intervient ainsi dans la langue elle-même et en constitue l'un des principes.

1.3 Conception Cosérienne des langues

Ces critiques nécessitent donc une autre conception du langage qui soit dynamique. Coseriu reprend à son compte la conception de Humboldt :

> « Le langage est précisément dans son essence *energéia*, c'est-à-dire, une activité libre et créatrice : non pas seulement utilisation de ce qui a été créé linguistiquement, mais, originellement et en premier lieu, création linguistique. Et une langue particulière (un 'système linguistique') est une tradition technique du langage : une technique historiquement donnée en vue de la réalisation de cette activité qui est en elle-même créatrice." (ibid, p.8)

La langue n'est donc pas un "produit", un état réalisé, mais une production, une énergie, une technique ouverte et potentiellement dynamique, « dans laquelle la possibilité de son propre dépassement (changement) et les lignes de force de son développement

ultérieur sont données d'avance »(ibid, p.8). Le changement linguistique est donc inhérent aux langues, qui se combine avec la continuité dont Coseriu reconnaît d'ailleurs le primat. Comme l'indique Coseriu, « une langue (quand elle n'est pas une 'langue morte') n'est à aucun moment un produit entièrement fini : elle se produit de plus en plus à travers le soi-disant changement linguistique. »(ibid, p.8)

Les systèmes linguistiques sont donc en constante évolution, tout simplement parce que l'activité de parler et les discours qui réalisent cette activité assurent une double fonction : la préservation du système, en reproduisant et transmettant la norme linguistique et les règles mémorisées, mais également une fonction intrinsèquement créatrice, puisque chaque discours est par définition un événement qui produit un sens nouveau, et qui peut le faire parce que le système comprend en lui-même un potentiel de réalisations, par toutes les règles qu'il prévoit. Cela conduit à inverser les oppositions saussuriennes langue (système) / discours (acte) et synchronie / diachronie : les systèmes linguistiques sont constamment renouvelés par les discours, sont dynamiquement construits par les paroles et leur diffusion dans la communauté. La synchronie, comme état, n'est qu'une abstraction pratique, qui oublie le potentiel créatif des langues. C'est en diachronie, par les événements linguistiques, que les langues restent vivantes en se réalisant, et se modifient continuellement par les expressions créatrices des locuteurs.

Coseriu, pour décrire précisèment cette langue, distingue "trois couches techniques" de la langue : norme, système et type. « La norme comprend tout ce qui a été créé concrètement de par l'utilisation d'une technique linguistique, et qui, dès lors, 'existe' en tant que langue déjà produite : elle est l'ensemble des réalisations traditionnelles au sein d'une langue (y compris les règles de réalisation) et comporte dès lors aussi des caractéristiques non fonctionnelles, mais nécessaires à la réalisation ou tout simplement 'usuelles'. Le système linguistique comprend ce qu'il y a de fonctionnel dans la technique linguistique, c'est-à-dire, les oppositions et les procédés fonctionnels de la langue concernée, et son organisation correspond ainsi à ce qu'on appelle précisément aussi en linguistique structurale un 'système linguistique' (ou 'structure linguistique'). Un même système linguistique admet toutefois différentes réalisations, et peut dès lors correspondre aussi à plusieurs normes linguistiques. Le type linguistique, de son côté, comprend les catégories de fonctions et de procédés, les principes fonctionnels d'une technique linguistique ; il peut se réaliser dans différents systèmes de différentes façons et à différents degrés, et peut dès lors correspondre en principe à plusieurs systèmes linguistiques. »(ibid, p.9-10), voir aussi (Coseriu, 1962).

L'apport de Coseriu, ici, concerne notamment la norme, qui est une idéalisation des réalisations passées de la langue ayant pour les locuteurs une valeur normative. Il y a pour tout système une ou plusieurs normes, qui sont à mettre en relation avec les variétés. On peut également mettre en relation cette notion avec celle d'usage qui est très proche. Le type, quant à lui, est à mettre en relation avec la typologie des langues, du point de vue de leur fonctionnement.

Il peut se produire des changements linguistiques aux trois niveaux. Au niveau de la norme, il s'agit de la simple réalisation d'un potentiel du système linguistique (fonctions oppositives et procédés de constructions), par exemple la réalisation des exemples de

Saussure (*interventionnaire, répressionnaire, firmamental*). De la même façon, au niveau du système, il peut se produire un changement dans les fonctions et les procédés, non-existants à un moment donné du temps, mais néanmoins potentiellement déjà présent dans le type linguistique. Par exemple, le passage d'un système casuel à un système flexionnel est une potentialité de type qui s'est réalisée diachroniquement entre le latin et les langues romanes. Il n'est donc pas possible de détacher l'état et l'événement, synchronie et diachronie, car ils vont de pair et sont indissociables à la fois pour la préservation et l'évolution des langues.

Cette ambivalence des discours, à la fois condition de préservation de la *mémoire linguistique collective*, et condition de son renouvellement continu par son application à des situations nouvelles, est synthétisée dans la figure 1.1.

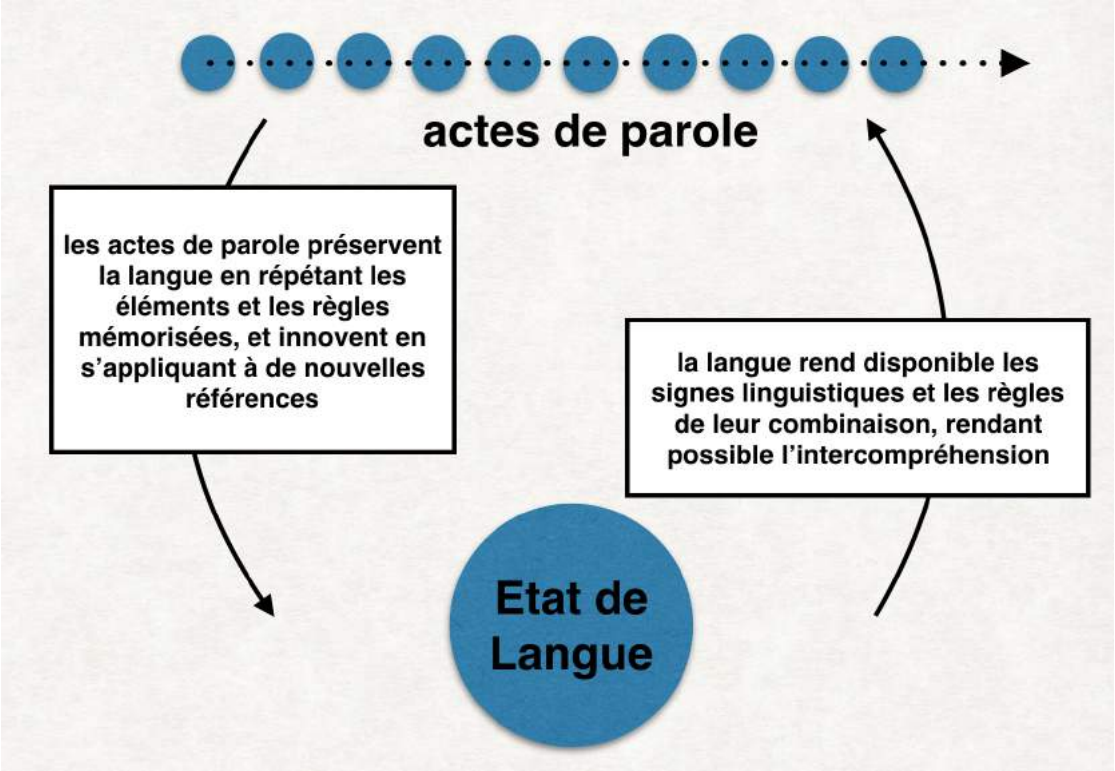


FIGURE 1.1 – Modélisation de l'interaction langue - discours sur l'axe diachronique

1.4 Perspectives pour l'analyse des langues

La conception de la langue comme *energéia* impose non seulement une révision des articulations langue/parole et synchronie/diachronie, mais également une révision de la limitation du champ de la linguistique à la linguistique interne. En effet, la langue peut être abordée pour elle-même, mais peut aussi être abordée d'au moins deux autres points de vue complémentaires : d'abord, il faut considérer les langues comme des systèmes qui

vivent et sont façonnés par des communautés humaines, et nous devons prendre un point de vue sociolinguistique ; ensuite, la langue est manipulée par les individus et plus exactement leur appareil cognitif, et nous devons prendre un point de vue psycholinguistique.

Enfin, de par l'importance du discours dans un modèle dynamique des langues, il faut approfondir cette notion.

1.4.1 Perspective sociolinguistique

L'importance du lien *langue / société* avait déjà été souligné par (Meillet, 1904) et même par Saussure puisqu'il parlait de la langue comme d'un « fait social ». Mais ni l'un ni l'autre n'en tireront de conséquence pour l'étude des langues. Dans le structuralisme, la langue est un système constituant une norme partagée par la communauté linguistique, et Saussure ne considère pas la variation au sein des langues comme un aspect essentiel [2]. Pourtant, à l'évidence, par exemple en français, il existe deux normes correspondant *grosso modo* à la langue parlée et à la langue écrite (voir (Gadet, 2007; Guerin, 2008). Pour rendre compte de ces phénomènes de variété, Coseriu fera une distinction entre la langue considérée comme une *faculté universelle de l'être humain*, la *langue d'un point de vue historique*, qui est liée à la création d'une tradition discursive (la langue française par exemple), et la *langue comme une entité fonctionnelle*, qui est celle qui est effectivement utilisée par les locuteurs : il existe donc pour le français au moins deux langues fonctionnelles, la langue écrite et la langue parlée, qui sont des réalisations concrètes et perceptibles de la langue historique 'français'. Une langue historique est l'architecture générale au sein de laquelle plusieurs langues fonctionnelles peuvent coexister. Coseriu proposera de caractériser les langues fonctionnelles selon trois dimensions : une dimension géographique, la diatopie, une dimension sociale, la diastratie, et une dimension liée aux individus et aux situations de communication, la diaphasie. Nous reviendrons plus en détail sur les conceptions proposées par la sociolinguistique dans le chapitre suivant, car les innovations lexicales doivent d'abord être considérées comme des variations qui émergent dans un contexte sociolinguistique spécifique, et peuvent éventuellement diffuser au-delà de la communauté d'émergence.

1.4.2 Perspective cognitive

Le troisième point de vue que nous devons prendre pour analyser les langues permet de les replacer dans le cadre du système cognitif humain, lui-même en interaction constante avec notre environnement. Après tout, les locuteurs individuels sont les seuls porteurs de la langue, les langues sont des systèmes symboliques qui associent des formes à des états mentaux, et une bonne compréhension des langues nécessite aussi d'articuler ces systèmes avec les différents processus cognitifs : perception, attention, interprétation, mémorisation, catégorisation, prototypie, inférence, planification, action, etc.

Les travaux et les approches de la psycholinguistique et de la linguistique cognitive mériteraient (au moins) un chapitre distinct. Dans le présent travail, nous nous limi-

2. Chez Saussure, comme souvent, l'analyse est cependant plus nuancée, car il reconnaît l'existence de manières différentes de parler au sein des langues nationales, mais il ne s'attarde pas sur la question.

terons à quelques notions qui paraissent essentielles pour comprendre les innovations lexicales et le changement linguistique en général. Rappelons d'abord que la grande majorité des linguistes cognitivistes adoptent une méthode basée sur l'usage linguistique, et considèrent le changement linguistique comme une propriété essentielle des langues.

1.4.2.1 Représentations mentales

La linguistique cognitive considérera, en réaction aux apories des travaux des grammaires génératives, que le langage est un phénomène mental. L'objectif est donc d'analyser les langues en relation avec les états mentaux et les processus cognitifs qui les génèrent. Il en a découlé une vision des unités lexicales - et singulièrement des représentations mentales attachées aux formes linguistiques - et de leur organisation sémantique que nous pouvons résumer ainsi :

— **une conception radiale de la catégorisation humaine organisée autour de prototypes** : à partir d'expériences réalisées sur la catégorisation naturelle des couleurs et d'objets de la vie quotidienne, Eleanor Rosch (Rosch, 1978) énoncera une théorie de la catégorisation selon laquelle les catégories sémantiques (par exemple *oiseau*) sont organisées autour de prototypes (par exemple *moineau*), membres exemplaires et/ou propriétés typiques (par exemple *voler, avoir des plumes, un bec etc.*), et l'appartenance d'une entité à une catégorie est dés lors liée au degré de partage de ses propriétés avec le prototype ; il en découle d'une part qu'une grande partie des catégories sémantiques sont organisées selon le principe des ressemblances de famille (par exemple un pingouin ne vole pas mais a un bec et des plumes), et que les frontières entre catégories sont généralement floues (la tomate est-elle un fruit ou un légume ?). Il existe également un niveau privilégié de catégorisation, le niveau de base (par exemple *chien*, par rapport à *labrador* ou *Médor*, ou encore *canidé*), qui est le plus approprié dans nos pratiques courantes, car il permet de discriminer les objets tout en restant suffisamment concret. (nous renvoyons à (Geeraerts, 2010, chap.5) et (Kleiber, 1990) pour une discussion) ; cette théorie du prototype fournit un moyen d'expliquer par exemple les évolutions de sens en considérant un sens initial prototypique ; surtout, elle permet de rendre compte d'une catégorisation sémantique non déterminée par des conditions nécessaires et suffisantes, mais par des propriétés typiques qui en définissent le noyau, laissant une liberté d'évolution et d'application à des objets ne répondant qu'imparfaitement au prototype ; Les nombreuses relations qui lient les polysèmes à partir du prototype seront formalisées par (Lakoff, 1987) dans sa théorie du réseau radial de relations.

— **une organisation cognitive globale des catégories autour des notions de cadre et de scenarii** : une hypothèse forte de la linguistique cognitive est de considérer que nos états mentaux sont fondamentalement le résultat de nos interactions avec l'environnement extérieur, et que notre esprit génère, par schématisation, une représentation mentale abstraite des catégories ; mais la représentation mentale reste ancrée dans les situations et aux savoirs que nous avons accumulés. La sémantique lexicale doit donc rendre compte d'une connaissance

holistique et encyclopédique : le sens de *pain* n'est pas limité à la description de l'objet concret, mais doit également rendre compte de l'ensemble des pratiques, des *scenarii*, dans lesquels il peut se trouver impliqué. Pour approcher ce sens, (Fillmore, 1977; Fillmore, 1985) proposera le modèle des frames sémantiques qui donnera lieu ensuite aux travaux de FrameNet. Le lien entre nos représentations mentales et les interactions que nous avons avec le monde extérieur sont la source même du changement de ces représentations ;

— **le sens linguistique comme une perspective sur les choses** : la linguistique cognitive prendra enfin une optique perspectiviste, considérant que nos catégorisations ne sont rien d'autre que des conceptualisations de la réalité qui nous sont utiles dans nos interactions avec l'environnement, mais qui sont tout à fait indépendantes de la réalité elle-même, qui peut être considérée de différents points de vue ; on peut trouver l'inspiration de ce relativisme linguistique dans l'hypothèse Sapie-Whorf et même dans la vision de Humboldt de la *WeltAnschaung*. Surtout, il s'agit là encore d'un principe qui permet l'innovation lexicale, qui peut être considérée comme un nouveau point de vue sur les objets ;

Parmi les opérations fondatrices de la cognition, les linguistes cognitivistes insisteront particulièrement sur la métaphore et dans une moindre mesure sur la métonymie, comme processus cognitifs des évolutions de sens. Nous y reviendrons dans la section consacrée à la néologie sémantique.

1.4.2.2 Processus mentaux : mémorisation des représentations mentales

Les hypothèses précédentes, validées par de nombreuses expériences pour les processus de catégorisation et qui ressortissent d'hypothèses de travail pour les autres, sont complétées par des hypothèses sur le processus psychologique de mémorisation des unités linguistiques. Un concept central introduit par Ronald Langacker est celui d'*entrenchment*[3]:

> « There is a continuous scale of entrenchment in cognitive organization. Every use of a structure[4] has a positive impact on its degree of entrenchment, whereas extended periods of disuse have a negative impact. With repeated use, a novel structure becomes progressively entrenched, to the point of becoming a unit ; moreover, units are variably entrenched depending on the frequency of their occurrence. »(Langacker, 1987, p.59)

L'implantation cognitive des unités linguistiques a trois propriétés :

— elle dépend principalement du nombre d'*expositions* à l'unité linguistique en contexte et donc à la fréquence d'usage : une plus grande exposition, et donc une plus grande fréquence d'occurrences, impliquera mécaniquement une plus grande implantation cognitive ; de nombreuses expérimentations ont montré l'importance

3. (Legallois et François, 2011) proposent de traduire le terme par « enracinement cognitif. »Afin de tenir compte du fait que le processus couvre également le déracinement cognitif, nous préfèrerons le terme d' *implantation cognitive*.

4. Ce terme chez Langacker correspond à toute unité linguistique, lexicale ou lexico-syntaxique, le terme de construction n'étant pas présent en 1987.

de la répétition dans la mémorisation des unités linguistiques (Blumenthal-Dramé, 2012) ;

— Elle concerne non seulement les unités linguistiques proprement dites, mais également les unités constructionnelles (structure syntaxique, lexico-syntaxique, schémas de dérivation), ainsi que les routines discursives ; il s'agit d'un processus général de mémorisation non seulement d'unités individuelles (les lexies), mais également de procédures de création lexicale (l'affixation et la composition), et de constructions syntaxiques ; il s'agit donc d'un processus de schématisation à partir d'exemplaires : plus on sera exposé à des occurrences du schéma de formation cyber-NOM, plus ce schéma de construction sera disponible et l'accès automatisé ;

— le processus d'implantation aboutit à une routinisation/automatisation de l'accès au sens des unités ou constructions, qui sont alors accessibles de manière automatique, avec un minimum ou une absence d'effort cognitif ; pour illustrer ce processus, Langacker le compare à certaines routines comportementales, comme lacer ses souliers ou réciter l'alphabet ; les lexies qui sont entrées dans l'usage d'une communauté ne sont rien moins que des associations inconsciemment accessibles, tandis que les innovations lexicales sont généralement d'abord des unités qui nécessitent un effort cognitif pour leur compréhension ;

Hans-Jörg Schmid propose une définition synthétique de l'entrenchment, dans laquelle il insiste sur son caractère dynamique :

> « Entrenchment refers to the ongoing reorganization and adaptation of individual communicative knowledge, which is subject to exposure to language and language use and to the exigencies of domain-general cognitive processes and of the social environment. Specifically, entrenchment subsumes processes related to :
>
> a) different strengths of the representations of simple and complex linguistic elements and structures,
>
> b) degrees of chunking resulting in the availability of more or less holistically processed units,
>
> c) the emergence and reorganization of variable schemas providing the means required for generative linguistic competence. »(Schmid, 2017, p.12)

Prenons un exemple pour illustrer les degrés d'implantation cognitive : si je dis (ou j'entends) *voilà un ciel menaçant!*, je n'ai plus besoin de faire l'analyse compositionnelle de 'ciel menaçant' (qui ne m'amènerait d'ailleurs par directement au sens, puisqu'une menace est prototypiquement une action agentive) , j'accède directement au sens de l'ensemble *ciel menaçant* comme associé à l'idée d'un 'ciel annonciateur d'orage'. À l'inverse, *ciel annonciateur d'orage* n'est pas accessible comme une unité, mais comme un sens compositionnel de type procédural. On peut dès lors émettre l'hypothèse que la langue est le résultat des implantations cognitives continuelles (mémorisation et oubli), induites par les expositions auxquels chacun est confronté : cela constitue un système, car ce ne sont pas seulement des unités linguistiques qui sont mémorisées, mais également des règles de construction (formations des mots, constructions lexico-syntaxiques, syntaxiques, routines discursives), par schématisation.

la fréquence n'est pas le seul déterminant de l'implantation cognitive. Nous renvoyons à (Schmid, 2007) pour une étude plus approfondie sur ce point.

La notion d'implantation cognitive doit être mise en regard d'une notion connexe en lexicologie, celle de lexicalisation, qui désigne le processus par lequel une unité linguistique est adoptée par l'ensemble de la communauté linguistique. Certes, la notion d'implantation cognitive ne se situe pas au niveau de la collectivité, mais au niveau de l'individu. Mais on retrouve cependant ici une vision non plus binaire lexicalisation/non-lexicalisation, mais graduelle et fondée sur un critère psychologique (évident). Le lexique mental de chacun est lié aux interactions linguistiques auxquelles il a été confronté, et il n'est donc de ce fait jamais *exactement* le même d'un individu à l'autre. Mais surtout le phénomène d'implantation cognitive est un processus continument évolutif, qui couvre bien plus que la seule lexicalisation (pour tel individu), il couvre tous les états possibles : de la première exposition, l'émergence, à une exposition plus accrue, à une éventuelle automatisation de l'accès, et à sa dégénérescence et à son oubli.

(Schmid, 2016; Schmid, 2015a) proposera d'inclure cette notion d'*entrenchment* dans un modèle plus large, qui tient compte des trois perspectives que nous devons adopter pour analyser les langues :

- **une perspective linguistique**, cherchant à décrire les propriétés phonologiques, morphologiques, syntaxiques et sémantiques des unités lexicales, et les mécanismes permettant d'expliquer les modifications de l'une ou de plusieurs de ces propriétés, aboutissant à l'innovation lexicale ; la lexicalisation (anglais : lexicalization) représente dans ce cadre le processus d'intégration ou non du néologisme dans le vocabulaire de la langue considérée ;
- **une perspective cognitive**, qui cherche à modéliser et décrire les mécanismes de formation et d'*entrenchment* des unités lexicales dans l'esprit des locuteurs d'une communauté linguistique ; la notion de formation de concept (anglais : concept-formation) couvre dans cette perspective le processus d'implantation cognitive ou non du symbole linguistique dans l'esprit des individus (voir (Schmid, 2016) pour un état de l'art) ;
- **une perspective socio-pragmatique**, qui cherche à modéliser les paramètres sociaux et pragmatiques permettant de décrire comment les innovations lexicales deviennent ou non graduellement intégrées à la mémoire collective. La notion d'institutionnalisation (anglais : institutionalization) subsume dans cette perspective le cycle de vie des néologismes pour la communauté linguistique.

(Schmid, 2008, p.3) reprend alors à son compte le terme d'*establishment* (Bauer, 2001, p.46) pour subsumer les trois termes de lexicalisation, formation de concept et institutionnalisation. Il reprend par ailleurs les trois phases classiques pour décrire le cycle de vie des innovations lexicales : *creation, consolidation, establishing*, qui se déclinent selon les trois perspectives (voir figure 1.2).

La présentation faite ici est parcellaire, les relations entre langue et cognition mériteraient un chapitre distinct. Dans la présente étude, nous retiendrons les quelques éléments explicités ci-dessus, et renvoyons à quelques références : (Divjak *et al.*, 2016; Schmid, 2007). On pourra également consulter (Geeraerts et Cuyckens, 2007; Geeraerts,

Perspectives: / Stages:	Structural perspective	Socio-pragmatic perspective	Cognitive perspective
creation	(product of) nonce-formation	(process of) nonce-formation	pseudo-concept
consolidation	stabilization	spreading	(process of) hypostatization
establishing	lexicalized lexeme	institutionalized lexeme	hypostatized concept

FIGURE 1.2 – Trois perspectives et trois phases du cycle de vie des innovations lexicales (Schmid, 2008, p.3)

2010; Fortis, 2011; Fortis, 2012; Labelle, 2001; Francois et Cordier, 2006; Divjak *et al.*, 2016)

1.5 Modélisation du discours

Chez Coseriu, la conception dynamique de la langue ne change pas véritablement le focus de la linguistique, dont l'objet reste la langue comme système. Pourtant Coseriu considère le discours comme le lieu de réalisation de la langue fonctionnelle et par là met l'accent sur son caractère central. Il est donc essentiel, notamment dans l'objectif d'une meilleure compréhension du changement linguistique, de disposer d'une modélisation du discours et des situations de communication : les innovations lexicales émergent dans des circonstances communicatives particulières, et suivre la diffusion éventuelle de ces innovations nécessite une caractérisation fine des situations de communication par lesquelles elles passent, jusqu'à leur éventuelle lexicalisation.

Deux pistes principales ont été explorées pour tenter d'analyser les paroles et les situations de communication. Il s'agit d'abord de toute la tradition s'intéressant aux types de texte, qui remonte aux travaux d'Aristote et va jusqu'aux travaux de la linguistique textuelle, notamment représentée par les travaux de Jean-Michel Adam(Adam, 1990; Adam, 1993; Adam, 2005). Le courant de l'analyse du discours, en partant des premiers travaux sur l'énonciation (Benveniste, 1970; Benveniste, 1974), des acquis de la pragmatique américaine (Searle, 1972; Austin, 1970; Sperber et Wilson, 1989) et du schéma de communication de (Jakobson, 1963), aboutissent à des analyses des textes normés et moins normés.

Là encore, il s'agit d'une piste *nécessaire* à explorer, à la fois pour une meilleure compréhension des discours, et, pour ce qui nous concerne plus spécifiquement, pour mieux caractériser les conditions d'émergence et de diffusion des innovations lexicales. Nous renvoyons à quelques textes de référence et aux analyses initiales faites dans la synthèse : (Halliday et Hasan, 1976; Halliday, 2006; Dijk, 1985; Van Dijk, 2008; Van Dijk, 2013;

Charaudeau, 1995; Charaudeau et Maingueneau, 2002; Charaudeau, 2015; Charaudeau, 2017; Maingueneau, 2005; Maingueneau, 2016).

1.6 Conclusion

Nous avons dans ce chapitre exposé les grandes lignes d'une conception dynamique de la langue qui permet de reconnaître l'existence et l'importance du changement linguistique : les langues sont des systèmes qui à la fois sont extrêmement stables dans le temps, mais qui dans le même temps sont en continuelle évolution, de par les changements qui se produisent dans le monde extérieur, de par les mouvements de populations, et de manière globale par l'histoire globale et l'histoire multiforme de chacun des individus constituant une communauté linguistique. Le changement linguistique touche toutes les composantes de la langue : la phonologie, la morphologie, la syntaxe, le lexique et même au-delà car les traditions discursives évoluent également. Ce changement se manifeste non seulement par des créations ou innovations, mais également, pour certaines créations, par la diffusion puis l'adoption. Il faut également considérer que les lexies déjà adoptées sont elles-mêmes dans un état potentiel de changement, car elles peuvent être moins utilisées, voire disparaître ou encore évoluer. Cela concerne à la fois les unités lexicales mais également les procédés de formation de ces unités lexicales. Ces évolutions font partie intégrante de la vie des langues, avec la préservation qui reste le socle permettant la communication et la transmission des langues : les discours sont d'abord là pour que nous puissions communiquer, et donc transmettre la langue partagée par la communauté.

L'innovation lexicale, qui se manifeste par de nouvelles formes lexicales et par de nouveaux usages, est sans doute le phénomène le plus visible et peut-être le plus intense dans les discours. Ces innovations lexicales sont nécessaires pour adapter les langues aux évolutions technologiques, sociales, économiques, culturelles qui se produisent, pour nous permettre de décrire des objets, des événements, des émotions, des idées nouvelles qui se présentent à nous, pour jouer avec la langue également.

Pour situer le phénomène du changement linguistique dans l'économie générale des langues, nous avons, en nous appuyant notamment sur les travaux de Coseriu, revu les dichotomies saussuriennes langue/discours et synchronie/diachronie. Si le discours est le lieu de transmission des langues et le lieu de leur modification continue, il faut également considérer que le changement est lui-même contenu dans les langues, car elles sont non pas une nomenclature d'unités à sens fixe, mais une suite de lexies ayant un sens potentiel qui sera réalisé chaque fois de manière différente dans les discours. De même, les langues comprennent des procédés de création lexicale : dérivation, composition, mais également création polylexicale et potentiel créatif des combinaisons syntaxiques. Les contacts avec d'autres langues permettent également l'émergence de nouvelles formations lexicales et jusqu'à des procédés de formation.

Nous avons également insisté sur la nécessité de prendre une triple perspective pour étudier l'innovation lexicale et le changement linguistique en général : une *perspective linguistique interne*, car il est évidemment nécessaire de décrire les mécanismes linguistiques

permettant l'innovation lexicale ; *une perspective cognitive*, car ce sont bien les individus qui font vivre et évoluer les langues, et il faut pouvoir comprendre les processus cognitifs qui président à la préservation comme au changement linguistique. De ce point de vue, le processus de mémorisation et son résultat, l'entrenchment, est central. La fréquence d'exposition et d'usage est le principal déterminant de la mémorisation et fonde une étude statistique et probabiliste des langues. Enfin, *une perspective sociolinguistique*, car l'innovation lexicale est avant tout une variation linguistique, et les langues sont traversées de variations et comprennent des variétés qu'il faut mettre au jour pour comprendre l'inscription sociale de certaines innovations ; de même, les innovations lexicales, dont on verra qu'elles sont relativement nombreuses de manière hapaxique, diffusent parfois au sein de la communauté linguistique, et il faut se faire une idée précise de la nature des communautés linguistiques pour expliquer les chemins menant parfois à l'adoption de certaines innovations.

Avec cette conception initiale, nous allons pouvoir préciser dans le chapitre suivant les notions d'unités lexicales et d'innovations lexicales.

Enfin, nous n'avons pas, dans cette étude, suffisamment étudié les discours, ou plus exactement les situations de communication, qui sont le lieu d'émergence et de diffusion des innovations. Il s'agira d'une piste primordiale dans un travail futur car il est probable qu'une caractérisation adéquate de ces situations de communication, des moins normées aux plus normées (les genres de texte littéraires mais également toutes les situations de communication de la vie sociale) nous donnent encore d'autres lumières sur les lieux d'émergence privilégiés des innovations, sur le rôle de telle ou telle situation dans la diffusion, etc.

Chapitre 2

Unités linguistiques et innovations lexicales

Sommaire

Pour approcher l'innovation lexicale, nous avons caractérisé le changement linguistique dans l'économie générale des langues. Nous avons repris à notre compte une vision de la langue comme *energéia*, énergie, qui considère les systèmes linguistiques comme des systèmes dynamiques, qui redonne au changement linguistique toute sa place. Cette approche nous a fait reconsidérer les articulations langue-discours et synchronie-diachronie : ce n'est que par les discours que les langues existent, et ce sont les discours qui préservent les langues et les font évoluer. Les langues sont des produits continument évolutifs. Les discours sont des événements et c'est la mémorisation de ces événements qui permet de construire au fur et à mesure le trésor commun qu'est la langue. Contrairement à ce que pensait Saussure, la langue est en elle-même dynamique car elle est le résultat de l'action de mémorisation des nouveautés qui surviennent continument par notre activité discursive. Cette approche nous a amené à considérer que l'étude linguistique devait s'appuyer sur une approche pluridisciplinaire : une première approche centrée sur le matériau linguistique lui-même (la linguistique interne), mais également une approche

qui tient compte des communautés d'individus qui permettent aux langues d'exister (la sociolinguistique) et également une approche qui tient compte de l'appareil cognitif humain, dont la faculté linguistique n'est qu'un composant. Le chapitre 3 détaillera plus avant les apports d'une approche sociolinguistique notamment pour réviser la notion de langue.

A partir de ce modèle général, il nous faut maintenant détailler la notion d'unité lexicale et ses caractéristiques, puisque les innovations lexicales sont d'abord des unités lexicales et héritent de leurs propriétés. Ensuite, nous devrons détailler la notion d'innovation lexicale, en insistant sur ses propriétés distinctives. Une conclusion fera la synthèse de ce parcours.

2.1 Notion d'unité lexicale

Si nous concevons les langues comme une structure, il est probable que cette structure s'organise autour d'unités linguistiques. Quelles sont-elles ? Dans la tradition structuraliste, l'unité linguistique essentielle est le *signe linguistique*, l'association arbitraire d'une image acoustique (le signifiant) à un concept (le signifié). Saussure ne détaille aucunement d'autres unités linguistiques, et reprend - sans s'y attarder- une conception classique distinguant dans le système linguistique les unités lexicales (le lexique) et les règles de leur combinaison (la syntaxe). Revenons brièvement sur ces éléments afin de caractériser plus précisèment les unités lexicales.

2.1.1 Mot, phrase et énoncé

Le langage commun utilise en français contemporain trois termes principaux pour évoquer les unités linguistiques : *mot, phrase, texte*. La plupart des grammaires modernes ont ajouté des notions qui rendent compte de la dimension discursive des langues. Par exemple, (Riegel *et al.*, 1994), s'appuyant sur le schéma de communication de Jakobson (Jakobson, 1963, p.213), isole tout d'abord l'énoncé, conçu comme :

> « une forme linguistique signifiante dont l'interprétation requiert une double aptitude. L'allocutaire doit, bien sûr, connaître le sens codé des formes linguistiques simples et complexes (mots, groupes de mots, phrases et types de phrases). Mais il lui faut aussi procéder à des calculs (ou inférences) à partir de la signification proprement linguistique de l'énoncé et des connaissances qu'il estimera pertinentes pour aboutir à une interprétation plausible de cet énoncé dans la situation où il lui a été adressé. »((Riegel *et al.*, 2018, p.3)

En nous concentrant sur les unités intraphrastiques, en deçà de l'énoncé, on trouve donc deux unités : les mots et les phrases, avec une unité intermédiaire, les groupes de mots. Ces unités sont définies plus loin dans le manuel :

> « Une phrase est d'abord une séquence de mots que tout sujet parlant non seulement est capable de produire et d'interpréter, mais dont il sent aussi intuitivement l'unité et les limites. »(ibid, p.103)

Après cette définition intuitive, (Riegel *et al.*, 1994) commencent par réfuter les trois critères généralement avancés pour définir cette unité : les critères graphique et phonétique (« une phrase est délimitée par deux pauses importantes et caractérisée par une intonation qui varie avec le type de phrase. »(ibid, p.103)) et le critère sémantique (la « complétude sémantique »). Ils proposent de s'appuyer sur le critère syntaxique : une phrase est un « assemblage de mots grammatical »(ibid, p.105) et :

> « (...) la phrase constitue l'unité supérieure, à la fois complète et autonome, susceptible d'être décrite au moyen d'un ensemble de règles morpho-syntaxiques. Elle est formée de constituants (elle est construite) sans être elle-même un constituant (elle n'entre pas dans une construction syntaxique d'ordre supérieur et n'a donc pas de fonction grammaticale au sens ordinaire du terme). Cette double propriété fait de la phrase le cadre à l'intérieur duquel se déploient et se décrivent le réseau de relations (**les fonctions grammaticales**) et les classes d'unités simples (les **parties du discours**) et complexes (les **groupes de mots**) qui constituent l'architecture syntaxique des énoncés. »(ibid, p.104-105)

Ils reprennent également à leur compte la définition de Benveniste, qui insiste sur le double rôle de la phrase, comme une prédication, au sens logique, c'est-à-dire affirmation de quelque chose sur quelque chose, et comme acte communicatif, c'est-à-dire énoncé :

> « La phrase est l'unité de discours. Nous en trouvons confirmation dans les modalités dont la phrase est susceptible : on reconnaît partout qu'il y a des propositions assertives, des propositions interrogatives, des propositions impératives, distinguées par des traits spécifiques de syntaxe et de grammaire, tout en reposant identiquement sur la prédication »(Benveniste, 1974, p.130)

C'est en définissant la notion de phrase canonique (ou phrase de base), c'est-à-dire une phrase assertive, simple et neutre, qu'il est possible d'isoler la composante prédicative, et la composante communicative, par exemple : prédication : *le chat est sur le toit*, acte communicatif : *Pierre pense que le chat est sur le toit, Le chat est sur le toit!, etc.*.

La phrase est conçue comme une suite hiérarchique de constituants, et une analyse en constituants immédiats permet de passer des mots aux phrases par le biais de l'unité syntagmatique (les groupes de mots ou syntagmes).

La définition de la phrase contient plusieurs éléments pour définir les mots (et les groupes de mots) : d'une part, les mots ont une nature (une catégorie grammaticale) et d'autre part, ils ont une fonction, définie comme « le rôle que cet élément joue dans la structure d'ensemble de la phrase où il est employé. »(ibid, p.106) les catégories grammaticales (ou parties du discours) sont donc considérées comme une unité de niveau supérieur au mot, puisqu'il s'agit d'une classe de mots avec, pour le français, traditionnellement, neuf catégories (nom, article, adjectif, pronom, verbe, adverbe, préposition, conjonction, interjection).

L'identification de ces catégories est possible par un faisceau de critères :

— d'un point de vue morphologique, les verbes, les adjectifs, les noms et les pronoms portent des flexions spécifiques qui permettent de les distinguer entre elles, et de

les distinguer des autres catégories qui ne portent pas d'information morphologique ; mais les autres catégories ne peuvent pas être distinguées par ce critère ;

— d'un point de vue syntaxique, on peut définir des classes distributionnelles sur la base des comportements syntaxiques de chacune des catégories, par exemple le fait qu'un nom pourra être sujet ou objet d'un verbe, ou qu'un article sera dépendant d'un nom ; mais là encore, il est difficile de déterminer les comportements discriminants des différentes catégories, même s'il est possible de déterminer des comportements typiques. L'impossibilité qu'a démontré la grammaire générative à décrire ses règles de fonctionnement syntaxique montre bien qu'il n'existe pas de fonctionnement exhaustivement identificatoire pour chacune des catégories, chaque élément lexical appartenant à une catégorie ayant un fonctionnement spécifique ;

— d'un point de vue sémantique, les noms désignent *plutôt*des êtres vivants et des choses, les adjectifs des propriétés et des manières d'être, les verbes des actions ou des états, etc. Mais les noms peuvent également désigner des actions ou des événements (*course, élection, état*, ce critère n'est donc pas non plus suffisant.

Si nous remontons au fondateur de l'approche distributionnelle, et à la notion de classe distributionnelle, Harris avait conscience du caractère continu des catégories de discours. Sa réponse est de ne décrire que les cas « centraux », en laissant de côté les cas « limites » :

> *In a situation of this sort, there is something else that can be done. One can look for the constraints on combinations, meaning what it is that precludes certain combinations from occurring, what prevents the randomness of combination of words. If it turns out that we can do this, that we can find constraints that are describable precisely, the fuzziness of grammar is located in a particular section of this, because it is not in the definition of the constraints, it's in the domains of the constraints, so that this makes it possible to have at least part of the structure of language can be completely precise.*(Harris, 1988)

Si nous essayons de définir les mots eux-mêmes, en partant de la caractérisation générale de Saussure d'une association arbitraire signifiant-signifié, nous rencontrerons de même de grandes difficultés à déterminer les conditions nécessaires et suffisantes à leur identification. (Saussure, 1916, p.154) reconnaissait lui-même le paradoxe du mot : « ... le mot, malgré la difficulté qu'on a à le définir, est une unité qui s'impose à l'esprit, quelque chose de central dans le mécanisme de la langue. »(Riegel *et al.*, 1994, p.531) le définissent, en première approximation, comme une « unité préconstruite, ou précodée »que la langue fournit pour construire des énoncés, en citant à la fois des mots simples et complexes (*chemin de fer, pis-aller*).

Cette unité lexicale préconstruite, en dehors de sa forme, porte une information d'ordre grammatical, puisque les mots sont tous nécessairement une partie du discours, et une information d'ordre dénominatif, puisqu'ils dénotent un type de référent spécifique. Cette notion de préconstruction, qui se traduit notamment par l'insécabilité, permet de les distinguer des groupes de mots (exemple : *petite table ronde*, qui peut être décomposé en unités plus petites).

Cependant, les expressions semi-figées (*roman à l'eau de rose, prendre la mouche, etc.*) sont-elles alors à considérer comme des syntagmes ou comme des mots, puisqu'elles sont décomposables ? À l'inverse, les "mots" *du, des, au, aux* sont constitués de deux mots (de + le/les et à + le/les) et ne rentrent pas non plus dans les définitions proposées. Le critère de séparation graphique, souvent utilisé également, ne tient pas pour les expressions totalement figées (*pomme de terre, effet de serre, prendre note, au fur et à mesure*), ni pour les unités discontinues (*ne...pas, auxiliaire + V-participe passé : a ... chant-é*).

Néanmoins, c'est sur ces bases que la linguistique a forgé un terme technique pour désigner les mots : on parlera de lexie [1], ou de lexème. Les traditions linguistiques ont également convergé sur des couches de description linguistique : phonétique et phonologie, morphologie, syntaxe, sémantique (lexicale). C'est sur ces bases que sont construits la très grande majorité des manuels de langue, et sur ces bases que les dictionnaires décrivent les lexies.

2.1.2 De la lexie au morphème lexical et grammatical

Il reste un autre problème à résoudre, qui concerne la non insécabilité, qui ne tient pas non plus pour les lexies apparemment canoniques, puisque, en français, les noms, les adjectifs et les verbes se présentent sous différentes formes de par les flexions (*parler, parlerai, parlerions*) et peuvent donc être décomposés en deux formants.

Pour traiter ce problème, on doit à Bloomfield la distinction entre morphème libre (ou autonome) et morphème lié (ou non-autonome) :

> « 9. Def. A minimum form is a morpheme ; its meaning a sememe.
>
> Thus a morpheme is a recurrent (meaningful) form which cannot in turn be analyzed into smaller recurrent (meaningful) forms. Hence any unanalyzable word or formative is a morpheme.
>
> 10. Def. A form which may be an utterance is free. A form which is not free is bound.
>
> Thus, *book, the man* are free forms ; *-ing* (as in *writing*), *-er* (as in *writer*) are bound forms »(Bloomfield, 1926, p.155)

Ces deux notions permettent de rendre compte d'une différence entre des éléments porteurs d'une valeur sémantique référentielle, les lexies proprement dites (ou morphèmes lexicaux libres), qui sont autonomes, et les morphèmes liés, qui ne peuvent s'employer indépendamment et se rattachent aux morphèmes libres.

On distingue également parmi ces derniers éléments les flexions (ou morphèmes grammaticaux ou affixes flexionnels), porteuses d'une information sémantique générale liée à une partie du discours spécifique (mode, temps, aspect, personne, nombre, genre, voix pour les verbes par exemple), et les affixes (ou morphèmes lexicaux liés), qui sont éga-

1. Dans la suite, nous utiliserons le terme de lexie pour désigner les unités dotées d'une forme identifiable, d'un sens spécifique et d'une instruction syntaxique (partie du discours, fonctionnement combinatoire prototypique, mode de conceptualisation).

lement porteurs d'une valeur sémantique générale, mais ne sont pas nécessairement liés à une partie du discours (*anti-, non-, -iste, etc.*.

Cependant, là encore, les frontières sont floues : par exemple, un certain nombre d'affixes sont à la fois liés et libres, selon les emplois (*un anti, anti-social, un ex, ex-capitaliste*. De même, si l'on considère que la liste des affixes est fermée, comment traiter des formants comme *e-, cyber-* qui sont à l'évidence liés mais devraient être considérés comme des morphèmes libres ? De même, que faire des formations du type *dramatico-comique*, s'agit-il d'un morphème libre, de deux morphèmes libres ou d'un morphème lié et d'un morphème libre ?

Comme on peut le constater, quel que soit le critère utilisé, il paraît impossible de déterminer des conditions nécessaire et suffisantes pour caractériser de manière univoque les lexies (ou morphèmes libres) et les morphèmes liés (affixes et flexions). Dans le même temps, nous avons vu qu'entre les lexies et les groupes de lexies (les syntagmes) se placent toute une série de groupes de mots plus ou moins figés dont le statut est équivoque. Cela infirme-t-il les définitions proposées ? Il nous semble au contraire que cet état de fait signale une caractéristique essentielle des langues, ici appliquée aux unités linguistiques, à savoir qu'il existe un continuum entre les types d'unités linguistiques, et que les catégories forgées par la science linguistique permettent de caractériser des prototypes autour desquels tout un continuum de cas se rencontre. Certains critères ou certaines catégories ne sont peut-être pas adéquates, mais il faudrait aller plus loin dans l'étude pour s'en assurer (voir notamment le chapitre sur les néologismes formels, dans lequel nous reviendrons sur les notions d'affixes et de flexions). L'important ici concerne le caractère graduel entre les catégories d'unités linguistiques intra-propositionnelles, dont nous retiendrons les pôles : morphème lié (flexions et affixes), morphème libre (lexies proprement dite). Mais le continuum qui existe entre ces catégories nous amènent à une caractérisation plus générique, proposée par les grammaires de constructions, et qui nous paraît rendre compte plus adéquatement de la réalité linguistique.

2.1.3 De la lexie aux constructions

Les grammaires de construction (GC), initiées par (Fillmore *et al.*, 1988; Goldberg, 1995; Croft, 2001) (voir (Goldberg, 2013; Croft, 2007) pour une vision synoptique des différentes approches des grammaires de construction) sont liées d'une part au courant de la linguistique cognitive, qui remet la cognition au centre de l'analyse linguistique, et également - pour la plupart des courants - au paradigme de la linguistique basée sur l'usage. Selon (Goldberg, 2013), les GC partagent quatre hypothèses théoriques fortes et une cinquième généralement partagée:

— L'unité élémentaire de la grammaire est la construction, qui est une association conventionnelle forme-sens ; cette définition étend la notion classique de signe linguistique à des unités plus larges que le mot traditionnel mais également au-delà des locutions proprement dites, en correspondance avec la seconde hypothèse ; [2]

2. « Constructions are defined to be conventional, learned form-function pairings at varying levels of complexity and abstraction (...). This definition is meant to highlight the commonality between words

— La représentation sémantique est associée directement aux formes des constructions, sans aucune dérivation par un niveau intermédiaire syntaxique ; cette hypothèse forte, qui est également une réaction aux grammaires génératives, signifie que les langues sont uniquement composées de constructions, au moins jusqu'au niveau propositionnel ;
— Comme d'autres systèmes cognitifs, les langues constituent un réseau, composé de nœuds (les constructions) et de liens entre ces nœuds, de type hiérarchique (catégorisation) et associatif, permettant ainsi d'organiser les constructions selon leur plus ou moins grande abstraction et selon leurs affinités ;
— les langues sont soumises à des variations qui peuvent être décrites de différentes manières, notamment par des processus cognitifs et sociaux ;
— La structure des langues est construite par l'usage qui en est fait par les locuteurs. L'un des principaux facteurs d'identification des constructions conventionnalisées - adoptées par la communauté linguistique - est l'entrenchment (implantation cognitive) ;

Les deux premiers éléments sont les plus originaux. Le second est lié au rejet de la distinction classique - et formalisée à l'extrême par les grammaires génératives - entre un composant lexical et un composant grammatical, ce dernier dominant l'ensemble de l'architecture des phrases, le lexique étant simplement une liste d'items qui occupent des places grammaticales. Ce rejet est fondé sur la constatation de très nombreux cas limitrophes, et notamment l'étude de cas de constructions lexico-syntaxiques dont les schémas syntaxiques sont contraints au niveau des arguments lexicaux (par exemple le verbe *arriver* dans le sens 'parvenir dans un lieu' peut être décrit par la structure syntaxique $GNarriverPREPGN$, mais il faut aussi décrire les contraintes qui pèsent sur les différentes places argumentales et sur la préposition. De même, pour décrire la locution *prendre appui sur*, il faut décrire les possibilités d'insertion adverbiale et les contraintes sémantiques sur le complément.). Sur la base de ces cas particuliers, qui ne peuvent être résolus que par une analyse *conjointement* lexicale et grammaticale, la notion de construction a été généralisée et le niveau syntaxique génératiniste (règles de dérivation) rejeté. Cela aboutit à une conception des langues dont on pourrait décrire les unités dans un 'constructicon'. Les signes linguistiques, qu'ils appellent donc *constructions*, comprennent toute unité graphique ou sonore liée à un sens, du morphème au schéma syntaxique. Goldberg donne plusieurs exemples de constructions : morphèmes (*re-, -isme, -ait*), lexies (*Iran, another, banana*) expressions figées (*give the Devil his due, going great guns*) et semi-figées (*Jog <someone's> memory*), phraséologismes (*The Xer the Yer: the more you think about it, the less you understand*), constructions grammaticales (*Subj V Obj1 Obj2 : he gave her a fish taco, he baked her a muffin*). On assiste donc à une extension maximale, mais le critère décisif est qu'il s'agit, quel que soit le niveau, d'une paire forme-sens bien identifiable et spécifique. La différence réside dans la forme, qui peut être non variable ou « substantielle »(*substantial*), correspondant à la lexie, ou bien variable ou mieux « procédurale », c'est-à-dire - en termes harrissiens - intégrant une classe distributionnelle sur l'un ou plus de ses composants. Mais nous restons dans

and larger phrasal units. »(Goldberg, 2013, p. 1)

le cadre d'une parie forme-sens.

Certains tenants des GC n'emploieront pas le terme de construction de manière aussi radicale que Goldberg. Par exemple, (Croft, 2007, p.463) parlera de deux principes essentiels : « (a) a pairing of complex structure and meaning, (b) association of these pairings in a network. ». (Fillmore *et al.*, 1988) parle encore de structure argumentale syntaxique. De même, dans les analyses, il n'est pas rare de trouver une décomposition des informations selon la triparition traditionnelle (phonologie, syntaxe, sémantique). L'important ressortit à la disparition d'un niveau syntaxique indépendant, d'une part, et à l'extension formelle des unités linguistiques intrapropositionnelles au-delà de la lexie, jusqu'à couvrir l'ensemble des constructions syntaxiques conventionnalisées.

Ces extensions expliquent les dimensions permettant de distinguer les types de constructions (voir figure 2.1).

Size	Atomic	Complex	Intermediate
	red, -s	*pull strings, on top of*	*bonfire*
Specificity	Substantive	Schematic	Intermediate
	dropout, -dom	N, SAI	V-*ment*
Concept	Contentful	Procedural	Intermediate
	red, N	*-s*, SAI	*way*-construction

FIGURE 2.1 – Dimensions des constructions (Traugott et Trousdale, 2013, p.13)

La première dimension est liée à la forme des constructions, qui peut aller d'une taille minimale de type morphémique (flexions, affixes, lexies simples) à des entités complexes, composées de plusieurs unités morphémiques simples (locutions figées et semi-figées), tout en couvrant les cas intermédiaires où l'on peut partiellement décomposer l'unité en morphèmes (*bonjour, biocarburant, homme-sandwich, etc.*).

La seconde dimension concerne l'axe spécifique-schématique, et est liée au sémantisme de la construction, couvrant les unités ayant un sens spécifique (les lexies proprement dites), jusqu'aux catégories du discours et aux constructions syntaxiques (N V N pour la transitivité), en couvrant là encore les cas intermédiaires où se mêlent une composante spécifique et une composante schématique (par exemple la construction adverbiale en V-ment, ou encore toutes les affixations (PREF-X et X-SUFF) et flexions (généralement X-FLEX).

La troisième dimension, enfin, rend compte de l'instruction sémantique liée à l'utilisation en discours (ou plus simplement l'instruction syntaxique) qui permet de tracer un axe entre les unités à sens référentiel (les noms et dans une moindre mesure les adjectifs) et les unités à valeur grammaticale (les autres parties du discours, mais aussi les morphèmes liés, qui comportent également une instruction syntaxique en opérant sur des lexies).

Comme on le constate, l'un des intérêts de cette conception est lié à son pouvoir descriptif, car toutes les paires formes-sens sont couvertes, et à sa capacité à rendre compte de pôles de propriétés typiques sur un continuum.

Trois propriétés permettent ensuite de comprendre le réseau que constituent les constructions : la schématicité, la productivité et la compositionnalité.

La **schématicité - spécificité** permet de lier les constructions par des relations hiérarchiques : plus une construction est schématique, plus elle a de chances de comprendre des sous-constructions plus spécifiques. Cela vaut à la fois pour les constructions lexicales et pour les constructions procédurales. Par exemple, la construction 'objet manufacturé' est plus schématique que la construction 'fauteuil' ou 'piano'. De même, les constructions $N_{Sujet}V$ (verbe intransitif) et $N_{Sujet}VN_{Objet}$ (verbe transitif) sont moins schématiques que la construction V (verbe). Cette schématicité est l'un des principes d'organisation du réseau des constructions, les propriétés d'une construction plus schématique étant héritées par défaut par les constructions moins schématiques, qui en sont des instantiations. Les constructions les plus schématiques servent de modèles, par analogie, pour la création d'autres constructions. Mais il est évidemment possible que des sous-constructions rompent avec le schéma (et ses contraintes) sur lequel elles sont construites, ce qui oblige à une ré-analyse sur la base des propriétés de la sous-construction, et potentiellement une innovation. Dans ce cadre, on appellera *construct*, une instance discursive d'une construction, c'est-à-dire son emploi dans un contexte discursif spécifique, qui entraîne une charge d'information supplémentaire liée au contexte et à la situation d'énonciation et peut, en cas de répétition, aboutir à une nouvelle construction.

La **productivité** : cet aspect concerne l'ensemble des constructions, dès lors qu'elles ont un degré de procéduralité, même minimal. Dès lors, elles ont une plus ou moins grande productivité [3]. Par exemple, la flexion *-er*, pour former des verbes en français est très productive, au contraire de la flexion en *-ir*. Mais cela concerne toutes sortes de formations, par exemple, les suffixoïdes *-gate, -bashing* sont également productifs.

Enfin, la **compositionnalité** concerne la propriété des constructions de pouvoir ou non avoir un sens transparent. Pour les constructions simples, il ne peut pas y avoir de compositionnalité puisqu'elles sont construites sur des morphèmes indécomposables. Mais il y a tout un espace de l'opacité à la transparence, puisque pour certaines lexies simples, il subsiste une possibilité de décomposition (*bonjour, promener*, et il existe différents degrés de figement dans les constructions complexes.

La présentation des constructions que nous venons de faire établit une continuité entre les morphèmes liés et les constructions syntaxiques. Pour autant, il ne s'agit pas de nier l'existence d'une couche intermédiaire, *syntaxique* : comme un barycentre, la lexie simple est porteuse à la fois d'un sens spécifique et d'une instruction syntaxique, par la partie du discours à laquelle est appartient obligatoirement. On peut alors postuler que les parties du discours sont juste des "abstractions nommées" de classes distributionnelles, qui transmettent trois types d'instructions :

- une **instruction morphologique** : pour trois parties du discours (nom, adjectif et verbe), des règles de flexion sont spécifiquement ajoutées, dénotant des catégories sémantiques générales ;
- une **instruction syntaxique**, indiquant des règles génériques de la combinatoire de la lexie avec les autres lexies/parties du discours ;

3. Sur cette notion, nous renvoyons au chapitre sur les néologismes formels

— une **instruction sémantique**, sous la forme de catégories conceptuelles génériques (nom = objet ou entité considéré comme une unité conceptuelle délimitée, adjectif : propriété ou combinaison de propriétés applicables à des objets ou entités, verbe : état, relation, action ou événement impliquant des objets ou entités et/ou des propriétés). L'exemple de *critic-* est à cet égard illustratif : dans *une critique*, le sens est considérée comme une entité délimitée ('un type de discours particulier'), dans *une approche critique*, on a bien à faire à une propriété complexe, et dans *critiquer son voisin* à l'action de parole correspondante ('un type d'action spécifique'). On pourrait considérer, d'ailleurs que l'instruction syntaxique est dérivée de l'instruction sémantique : il s'agirait de la linéarisation de la conceptualisation sémantique des trois catégories principales de concepts.

D'un point de vue terminologique, nous ne retiendrons pas le terme de construction, trop marqué, à l'instar de Goldberg, pour désigner toutes les paires forme-sens. Nous adopterons une vision plus traditionnelle : morphèmes liés (flexion et affixe), morphèmes libres (lexie proprement dite, ou lexie simple), composés et constructions (dès le moment où des lexies simples sont combinées en obéissant aux règles génériques de la combinatoire des lexies).

Les passages entre les "pôles" ainsi définis sont nombreux : nous verrons spécifiquement ce qu'il en est entre flexion et lexie dans le chapitre 5. Entre les lexies et les constructions syntaxiques, il existe un réservoir lexical de constructions figées et semi-figées qui sont dans la zone limitrophe et montrent le continuum qui existe entre les deux pôles lexies et constructions [4]. Remarquons également qu'il existe, au-delà de la prédication-proposition simple où nous fixons la limite des "constructions", d'autres types de constructions généralement appelées pragmatèmes ou routines discursives, qui montrent bien que les constructions, comme paires forme-sens ayant à la fois des contraintes syntaxiques et sémantiques, occupent tout le spectre de l'expression discursive (voir par exemple (Blanco, 2015)).

Enfin, une autre hypothèse centrale des GC concerne la variabilité et la dynamique des langues : les constructions sont des unités conventionnalisées à un moment donné du temps et pour une communauté linguistique donnée. Et, par les discours, et le phénomène d'entrenchment, elles peuvent évoluer. Ce sera le sujet de la prochaine section que d'essayer de définir les innovations lexicales.

Mais nous ferons tout d'abord un rapide résumé des propriétés des morphèmes liés, des lexies et des constructions.

2.1.4 Propriétés des morphèmes liés, des lexies et des constructions

La tradition linguistique, si elle ne parvient pas à établir de critères nécessaires et suffisants pour distinguer les types d'unités linguistiques intra-propositionnelles, a

4. Voir synthèse, chapitre 1 pour un développement sur les unités polylexicales. Pour une description fine, on consultera (Mel'čuk, 2011).

néanmoins identifié des unités prototypiques sur lesquelles nous pouvons nous appuyer. Il s'agit, de l'unité formellement la plus petite à l'unité la plus grande : des morphèmes liés (flexions et affixes), des morphèmes libres (lexies proprement dites), des composés et des constructions. Nous délimitons les constructions aux formations générant une proposition, au sens logique de prédicat (verbe et ses dépendances obligatoires). Hors de ce champ, nous passons à la proposition énoncée (phase ou énoncé chez Riegel). Nous n'aborderons pas ici cette notion de proposition énoncée, dont l'analyse a été brillamment faite par (Adam, 1990; Adam, 2005).

2.1.4.1 Description des lexies

Pour décrire les lexies, nous nous appuierons sur les propriétés traditionnellement explicitées : formelles (phonologique et graphique/orthographique), morphosyntaxiques et sémantiques. Nous détaillons rapidement ces propriétés:

Propriétés formelles - phonologiques: Chaque lexie a des particularités phonologiques qui peuvent être transcrites en Alphabet Phonétique International. Nous aurons parfois recours, dans la description des innovations lexicales, même si cela n'est pas au cœur de cette étude, ni suffisamment développé, étant donné que ce travail ne comporte pas de corpus oral, à une transcription phonétique, notamment lorsque des modifications phonologiques se produisent par la création lexicale.

Propriétés formelles - (ortho)graphiques : En complément des spécificités phonologiques, les lexies ont une signature graphique-orthographique dont nous rendrons compte systématiquement, d'autant que nos corpus sont écrits. Notamment, nous indiquerons quand cela est nécessaire, les variantes orthographiques rencontrées.

Propriétés mophosyntaxiques - partie du discours : Concernant les propriétés morphosyntaxiques, l'information principale est l'assignation d'une partie de discours : le terme est pratique (et largement ancré dans la tradition linguistique) pour dénommer les *classes distributionnelles* ayant des propriétés communes de trois points de vue : du point de vue des flexions possibles, puisque les flexions sont liées à une partie du discours spécifique (pour ce qui concerne les noms, adjectifs et verbes) ; du point de vue du comportement syntaxique (combinatoire) de la lexie dans les propositions : à chaque partie du discours est attaché un ou des fonctionnement(s) combinatoire(s) prototypique(s), par exemple pour les noms d'être généralement précédés d'un déterminant, de pouvoir être modifié, d'être la tête des syntagmes nominaux, de pouvoir avoir de ce fait une fonction sujet, objet direct, objet indirect etc. vis-à-vis de la tête verbale ; enfin, du point de vue *sémantique*, imposant une conceptualisation spécifique (par exemple pour les noms de représenter un objet ou une entité de manière circonscrite, pour les adjectifs de représenter le sens spécifique comme une qualité, pour les verbes de représenter le sens spécifique comme une relation ou un processus [5]). Nous ne rentrerons pas ici dans le débat de savoir si c'est l'instruction syntaxique qui détermine la valeur sémantique (ou conceptuelle) ou l'inverse, nous souhaitons seulement décrire ces deux propriétés.

5. Nous renvoyons au travail initial de (Langacker, 1987) sur la sémantique des classes de mots. on pourra aussi consulter (Croft, 2001) qui étudie l'universalité des classes de mots et en déduit des valeurs conceptuelles

Il faut ajouter que, en dehors des lexies proprement dites (les morphèmes libres), d'autres unités intra-propositionnelles ont été introduites : en deçà de le lexie simple, les morphèmes liés, avec deux catégories principales: les morphèmes liés lexicaux (affixes) et les morphèmes liés grammaticaux (flexion). au-delà de la lexie simple, les composés et les amalgames (limités aux juxtapositions graphiquement marquées entre deux ou plus lexies simples : *homme-loup, attrape-feuille, aouthlétisme, optipessimisme*), les constructions lexicales figées ou semi-figées (appelées également unités polylexicales *perdre la face, effet de serre, etc.*), les constructions lexico-syntaxiques qui obéissent à la fois à une combinatoire syntaxique spécifique mais dont les places sont contraintes sémantiquement (*arriver*, dans le sens 'atteindre un lieu', pourra ainsi être décrit par *SN (agent) arriver PREP (locatif) SN (locatif)*), et enfin les constructions syntaxiques (c'est-à-dire sans contraintes paradigmatiques et dont on peut se demander si elles sont effectivement réalisées, en dehors de quelques mots grammaticaux (par exemple les déterminants définis et indéfinis, qui s'appliquent à tous les noms, ou la construction des formes composées et surcomposées des verbes mais *être* et *avoir* se partagent le champ des verbes au participe passé). Pour ces différentes *constructions*, elles emportent également une partie du discours, mais deux cas se présentent : soit la construction constitue dans sa totalité une partie du discours (ce qui est le cas de toutes les constructions en-deçà des unités polylexicales), soit elle constitue une construction au sens plus classique du terme, dont l'un des éléments est la tête et les autres éléments des dépendances ou des arguments (*perdre la face, jouer gros, mine déconfite, etc.*). Dans ce dernier cas, c'est la tête qui donne la partie du discours au tout.

Propriétés mophosyntaxiques - flexions : La propriété flexionnelle concerne les trois parties du discours lexicales, les lexies correspondantes intégrant sous forme suffixale des informations sémantiques génériques liées à chaque partie du discours : nombre pour les noms (le genre étant intégré et non variable sauf homonymie), genre et nombre pour les adjectifs, personne, nombre, genre, temps et mode pour les verbes. La caractérisation des possibilités flexionnelles pour les unités lexicales est importante, notamment quand des blocages se présentent. De même pour les innovations lexicales.

Propriétés morphosyntaxiques - combinatoire : Comme indiqué plus haut, la partie du discours emporte une caractérisation des possibilités combinatoires *génériques* de la lexie. En suivant les approches cognitivistes et constructionnelles, il est raisonnable de penser que chaque lexie répond certes au comportement syntaxique de la classe distributionnelle générale (nommée par la partie du discours) mais ce comportement est contraint sémantiquement (au niveau des arguments ou des modifieurs possibles dans un sens donné prototypique et conventionnalisé). Théoriquement, tout nom a la possibilité d'être modifié par tout adjectif, préférentiellement à droite. Mais le nom, dans un sens donné, est plus précisément décrit par un paradigme d'adjectifs prototypiques : on ne ressent aucune difficulté à adjoindre à *voiture* des propriétés prototypiques de cet objet : *voiture rouge, voiture puissante, voiture électrique, etc.* ; si je dis *voiture abstraite, politique, amphibie, ironique etc.*, je déroge aux propriétés typiques et attendues du concept 'voiture', et je suis obligé de ré-analyser le concept initial pour en faire une nouvelle interprétation. Nous nous dirigeons alors vers l'innovation lexicale. La vision générativiste

de parties du discours dont on pourrait décrire le fonctionnement générique n'est donc pas conforme à la réalité linguistique. Les parties du discours sont plutôt des abstractions prototypiques de propriétés discriminantes, morphologiques (flexions), combinatoires et sémantiques (nom = plutôt objet ou entité, adjectif = plutôt propriété, attribut, verbe = plutôt état, événement, action) qui sont *indéniables* (puisque chaque lexie doit pouvoir s'insérer dans ce système réglé), mais les lexies ont généralement à la fois des emplois limitrophes (par exemple les adjectifs qui peuvent tout à fait avoir une fonction nominale : *un bateau bleu* > *le bleu*) et des restrictions dans les emplois des règles génériques s'appliquant aux partie du discours. Il convient donc de spécifier le plus précisèment possible à la fois les combinatoires possibles pour une lexie, et surtout les contraintes sémantiques pouvant exister. La néologie sémantique se produit très souvent par une extension des contraintes sémantiques par rapport à un sens existant.

Propriétés sémantiques : Du point de vue de la description sémantique, nous opterons pour une vision prototypique, encyclopédique et holistique (multimodale) du sens, en accord avec la plupart des conceptions cognitivistes. Cet aspect mériterait un développement spécifique. Pour l'heure, nous nous cantonnerons à donner une définition générique et prototypique de la lexie ou de l'innovation lexicale, en nous appuyant sur la distinction classique dénotation (genre prochain et différences spécifiques) et connotation. Une information de domaine peut également permettre de distinguer des sens.

2.1.4.2 Éléments de formalisation

Nous aboutissons, avec ces propriétés, à la représentation schématique suivante (par exemple pour *maison*) (encart (2.1)) :

$$\left[\begin{array}{l}\text{Forme : } /API/, maison \\ \text{Morphosyntaxe :} \\ \quad\text{partie du discours : NOM (commun, féminin)} \\ \quad\text{flexion : singulier - Ø, pluriel - maison-s} \\ \quad\text{combinatoire : règles génériques du Nom + restrictions} \\ \text{Sémantique : 'lieu d'habitation de l'homme'}\end{array}\right] \tag{2.1}$$

Si nous passons maintenant aux morphèmes liés, la description sera évidemment différente, puisque ils fonctionnent à la fois comme des lexies, puisqu'ils portent une information sémantique, mais ne correspondent pas à une partie du discours, puisqu'ils ne sont pas libres. Ils portent une *règle*, étant donné qu'ils ne s'emploient qu'en combinaison avec des morphèmes libres. La règle qu'ils contiennent porte à la fois sur la morphosyntaxe du mot construit d'arrivée, et sur sa sémantique (voir (Booij, 2005a)).

Nous pouvons ainsi décrire les morphèmes liés comme indiqué dans les encarts (2.2),(2.3) et (2.4) (le formalisme est inspiré de Booij).

morphème lié : flexion

$$[X_{cat,sem_1}] + [Y_{,sem_2}] \Longrightarrow [XY_{cat,sem_1+sem_2}]$$

$$[arriv(er)_{V,\text{parvenir dans un lieu}}] + [-ons_{-,\text{présent, 1, pl}}]$$
$$\Longrightarrow [arrivons_{V,\text{parvenir dans un lieu, présent, 1, pl}}] \quad (2.2)$$

morphème lié : préfixe

$$[X_{-,sem_1}] + [Y_{cat,sem_2}] \Longrightarrow [X(-)Y_{cat,sem_1 \text{ REL } sem_2}]$$

$$[anti-_{-,\text{'contre, opposé à'}}] + [bruit_{N,\text{'son imprévu, sans harmonie'}}]$$
$$\Longrightarrow [anti-bruit_{(\text{Adj ou N}),\text{'contre les sons imprévus et/ou sans harmonie'}}] \quad (2.3)$$

morphème lié : suffixe

$$[X_{cat_1,sem_1}] + [Y_{\rightarrow cat_2,sem_2}] \Longrightarrow [XY_{cat_2,\text{'}sem_1 \text{ REL } sem_2\text{'}}]$$

$$[absurd_{Adj,\text{ 'contraire au sens commun'}}] + [\text{-ité}_{\rightarrow N(fem),\text{'état, chose'}}]$$
$$\Longrightarrow [\text{absurdité}_{N(fem),\text{'chose qui est contraire au sens commun'}}] \quad (2.4)$$

Dans ce formalisme, nous retrouvons les trois propriétés formelles, morphosyntaxiques et sémantiques. Par exemple, dans l'encart (2.4) :

- $[X_{cat_1,sem_1}]$ représente la forme lexicale (X) et ses propriétés morphosyntaxique (cat_1) et sémantique (sem_1),
- $[Y_{\rightarrow cat_2,sem_2}]$ représente la forme du suffixe (Y) et la règle (ou instruction) morphosyntaxique ($\rightarrow cat_2$) qui signifie que le préfixe n'a pas de partie du discours, mais qu'il génère une partie du discours, et la valeur sémantique (sem_2),
- $[XY_{cat_2,\text{'}sem_1 \text{ REL } sem_2\text{'}}]$ signifie que la lexie (morphème libre) générée est une concaténation des formes X et Y, que la propriété sémantique est héritée du suffixe, et que la propriété sémantique est une le résultat de la mise en relation REL de sem_1 et de sem_2,
- sur l'exemple de *absurdité*, la règle portée par le suffixe *-ité* est l'instruction morphosyntaxique de générer un Nom féminin ($N(fem)$) et la règle sémantique consiste à appliquer la notion générale 'état, chose' à la valeur sémantique de la lexie, ce qui aboutit au sens 'chose qui est contraire au sens commun'.

Pour les composés, on peut de même synthétiser leurs fonctionnement (encart (2.5)) :

composition simple de type V-N

$$[X_{V,sem_1}] + [Y_{N,sem_2}] \Longrightarrow [X(-)Y_{N(masc),\text{'Objet/Entité qui effectue } sem_1 \text{ appliqué à } sem_2\text{'}}]$$

$$\begin{aligned}&[porte_{V,\text{ 'tenir, soutenir'}}] + [avion_{N,\text{'appareil qui vole'}}]\\ &\Longrightarrow [porte-avion_{N(masc),\text{'objet qui soutient des appareils qui volent'}}]\\ &\Longrightarrow [porte-avion_{N(masc),\text{'navire qui transporte des avions etc.'}}]\end{aligned} \tag{2.5}$$

Nous avons ajouté dans ces formules une ligne supplémentaire, car la règle générale ne permet pas de rendre compte du sens attesté indiqué en dernière ligne. Il s'agit d'un cas de réanalyse fréquent. Ces formalismes seront détaillés dans le chapitre 6, quand nous étudierons plus en détail les dérivés et les composés. Nous essaierons également de préciser les règles de fonctionnement.

S'agissant maintenant des constructions, nous en avons explicité quatre types principaux : les constructions figées (c'est-à-dire celles qui sont totalement contraintes du point de vue de leurs constituants et de leur ordre), les constructions semi-figées (sans doute les plus fréquentes et qui, sur l'un ou plusieurs des aspects descriptifs admettent une variation), les constructions lexico-syntaxiques (admettant un paradigme pour un ou plusieurs des composants, plus la possibilité de transformations), les constructions syntaxiques (schémas dont au moins l'un des composants correspond à l'ensemble d'une des parties du discours). Nous ne les détaillerons pas ici, et renvoyons à l'étude préliminaire faite dans la synthèse (chapitre 1, section 2) ainsi qu'au chapitre 5 du présent document.

2.2 Notion d'innovation lexicale

Dans ce modèle général, nous définirons l'innovation lexicale, en première approximation, de la façon suivante : « une paire forme-sens qui dévie de l'usage mémorisé pour la communauté linguistique considérée. » Décomposons chacun des éléments de cette définition :

2.2.1 Paire forme-sens vs néologie formelle / néologie sémantique

Nous avons défini la notion de paire forme-sens dans la section précédente. Il est temps maintenant de revenir sur la distinction classiquement faite entre néologie formelle et néologie sémantique. En effet, cette opposition est méthodologiquement pratique mais théoriquement non valide : en effet, lorsqu'une innovation comporte la création d'une forme nouvelle (*email, courriel, macroniste, bureau-frigo*), il se produit également et nécessairement un sens nouveau (même si dans certains cas le concept nouveau peut entrer en concurrence avec un concept existant, par exemple *email* versus *courriel* – mais il subsistera même dans ce cas des différences même ténues au moins dans la perspective). De l'autre côté, dans le cas de la néologie sémantique et donc lorsqu'une ou plusieurs formes lexicales existantes sont utilisées, en se plaçant du point de vue

d'une conception du sens comme usage, il y aura également création à la fois d'une forme nouvelle entendue dans le sens de *construction nouvelle* et d'un sens nouveau : l'innovation *un tsunami d'applaudissements*, apparue dans les années 1970, est bien un nouvel emploi de la lexie *tsunami*. Au final, il s'agit d'une nouvelle forme (le sens nouveau n'est pas présent dans *tsunami* mais dans la construction *tsunami de N* (dénotant une collection de N surgissant brusquement et violemment, ces deux derniers traits provenant du sens initial), cette construction devenant, par extension du paradigme initial (limité à des noms comptables ou qui le deviennent par réanalyse dans la construction : *?un tsunami de bonté*), un déterminant complexe. Il y a certes une différence de conception de la forme entre la néologie formelle (forme au sens morphologique, en ajoutant les dérivés, les composés et les emprunts lexicaux) et la néologie sémantique (forme au sens d'une expression polylexicale plus ou moins figée, comme plus haut, mais aussi au sens plus large de constructions lexico-syntaxiques, dans le cas des emplois métaphoriques et métonymiques : *Apple se trouve au premier feu à droite, Apple décide de lancer son nouveau smartphone*. Ici, ce qui distingue les deux mentions d'Apple du sens initial 'société' ressortit à la combinatoire prototypique à laquelle les deux emplois ici proposés dérogent.

2.2.2 Déviation par rapport à une norme linguistique

Le second élément définitoire concerne la déviation introduite par l'innovation. Cette notion est bien évidemment ce qui permet de distinguer une innovation lexicale d'une unité lexicale déjà mémorisée. Cette déviation peut être reformulée en terme de variation : une innovation lexicale est en premier lieu une variation par rapport à l'usage en vigueur. Dans une perspective sociolinguistique et variationnelle, on ne peut par ailleurs pas se satisfaire de la notion de langue unique. Nous verrons dans le chapitre 2 que cette déviation pourra être définie par rapport à une langue fonctionnelle et la norme qu'elle définit. Notons dès à présent que le « sentiment néologique », variable d'un individu à l'autre, dépend justement de la ou des variétés que chaque individu a mémorisées (voir (Gardin *et al.*, 1974), (Sablayrolles, 2002)). L

2.2.3 Périmètre de l'innovation lexicale

Un autre point définitoire concerne le périmètre de l'innovation lexicale, au regard de son émergence et de sa diffusion. Certains linguistes, anglo-saxons notamment, limitent les innovations lexicales aux lexies en cours de diffusion, mais non encore entrées dans la norme. Cette définition permettrait de les distinguer des *nonce-words*, ou créations forgées pour une occasion particulière (on rencontre également les notions d'occasionnalisme et de protologisme). Dans notre conception, l'innovation lexicale débute dès sa première apparition ou émergence, l'histoire de sa diffusion (ou non) faisant partie du périmètre de son étude, la lexicalisation ou entrée dans l'usage étant la seule limite à l'innovation lexicale. L'opposition entre innovation lexicale (non entrée dans l'usage) et unité lexicale (entrée dans l'usage) doit encore être complétée en distinguant innovation lexicale et particularisme, lorsque la lexie n'est entrée que dans l'usage d'une variété (d'un

point de vue diastratique et/ou diatopique), innovation lexicale et idiotisme, lorsque la lexie ne concerne qu'un individu, (Gérard 2010), innovation lexicale et terme, lorsque la lexie ne concerne qu'une communauté professionnelle de pratiques.

2.2.4 Typologie des procédés néologiques

Une fois fixé le périmètre des innovations lexicales, il faut, pour les décrire, expliciter les différents mécanismes disponibles pour créer de nouvelles lexies et de nouvelles constructions. En effet, la création lexicale n'est pas un phénomène purement hasardeux et sans règle. Au contraire, on peut tout à fait isoler des mécanismes internes à chaque langue - autre preuve de son dynamisme intrinsèque - mis à disposition des locuteurs pour créer de nouvelles paires forme-sens. Il existe également un autre phénomène lié au contact des langues qui permet l'import dans une langue des mots d'une autre langue. Nous détaillerons ces mécanismes dans le chapitre 6, en nous appuyant sur la typologie proposée par (Sablayrolles et Pruvost, 2016; Sablayrolles, 2017).

2.2.5 Cycle de vie des lexies

Enfin, on ne peut parler des innovations lexicales sans évoquer leur cycle de vie : nous avons vu que nous considérons une innovation lexicale dès son émergence initiale, mais là ne s'arrête pas la vie des innovations, en tout cas la vie de certaines, l'histoire des mots est, après tout, comme le disait Alain Rey, l'histoire de la néologie, puisque tous les mots ont bien été créés à un moment, et ceux qui nous restent ont été mémorisés puis transmis. Il faut donc essayer de comprendre comment, à partir d'une création initiale, une diffusion peut s'en suivre puis une éventuelle adoption par la communauté linguistique. On peut prendre trois points de vue sur l'évolution des lexies : linguistique, sociolinguistique et psycholinguistique.

2.2.5.1 Approches linguistiques

Sans vraiment modéliser les phases du cycle de vie des néologismes, les linguistes reconnaissent généralement l'existence de trois phases saillantes : l'émergence, la diffusion puis la lexicalisation. Mais ils s'intéresseront principalement aux contextes d'émergence des néologismes, et aux critères permettant d'établir la lexicalisation, sans vraiment d'ailleurs y parvenir. Voir cependant (Sablayrolles, 2000a). Dans le domaine anglo-saxon, tout un pan de recherche a été consacré à l'institutionnalisation (adoption d'une unité par une communauté) et à la lexicalisation (le fait pour une création lexicale de devenir une lexie simple, avec perte de son caractère compositionnel), liant à la fois les approches linguistiques et sociolinguistiques. Par exemple (Bauer, 1983, 42-61), (Bauer, 2001, 33–47), (Hohenhaus, 2005) et (Lipka *et al.*, 2004).

2.2.5.2 Approches sociolinguistiques

La sociolinguistique s'est beaucoup intéressée aux modalités de diffusion des innovations lexicales. (Meillet, 1904), un des précurseurs de la discipline, expose un modèle

de la diffusion des changements linguistiques qui se trouve déjà implicitement dans son introduction : « ce qui le montre [que la langue est extérieure, d'une certaine façon, aux individus qui la portent], c'est qu'il ne dépend d'aucun d'entre eux de la changer et que toute déviation individuelle de l'usage provoque une réaction ; cette réaction n'a le plus souvent d'autre sanction que le ridicule auquel elle expose l'homme qui ne parle pas comme tout le monde ». L'émergence est individuelle, et la diffusion est le (long) chemin du passage de la nouveauté d'un groupe de locuteurs à un autre, avec les résistances des différentes communautés dont sont constituées les sociétés. Jusqu'à l'adoption éventuelle par l'ensemble de la communauté, la diffusion est marquée par deux processus : une diffusion facilitée au sein d'un groupe social pour autant qu'un membre influent adopte l'innovation, et une résistance "naturelle" pour la diffusion hors du groupe, qui peut être levée lorsqu'un membre appartenant aux deux groupes devient influent dans le second.

On retrouvera un modèle similaire et plus détaillé avec (Rogers, 2010). Le sociologue détaille de manière générique le modèle général d'une innovation réussie, c'est-à-dire adoptée par la communauté dans son entier (voir figure 2.2) : Dans ce schéma, la courbe

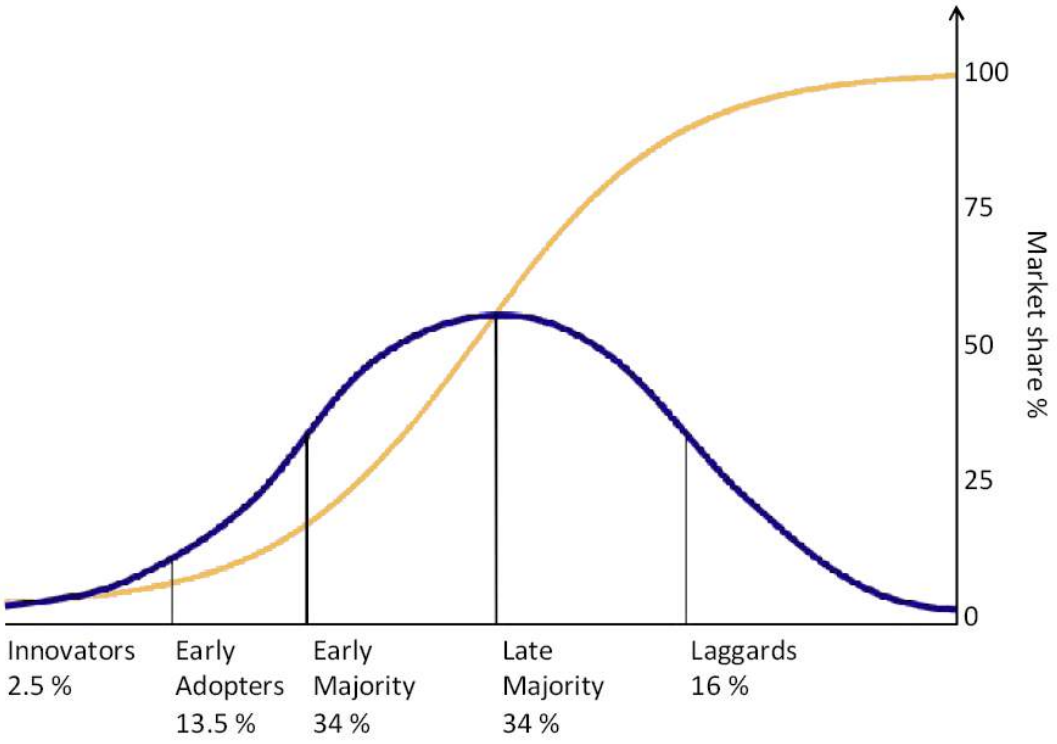


FIGURE 2.2 – Modèle de l'innovation réussie, d'après (Rogers, 2010)

en cloche représente la distribution des membres de la société vis-à-vis de l'innovation, divisés en cinq catégories : les innovateurs (2,5%), les adopteurs précoces (13,5%), la majorité précoce (34%), la majorité tardive (34%) et les retardataires (16%)). La courbe jaune indique les phases temporelles correspondant à l'adoption cumulative par chacun

des groupes, jusqu'à l'obtention de 100% des parts de marché (adoption définitive de l'innovation). Sur le principe, on retrouve là les distinctions de Meillet, avec un affinement des rôles des individus et des groupes, notamment la catégorie des adopteurs précoces qui sont a priori favorables aux innovations, et la division entre une majorité précoce et une majorité tardive, qui constitue en cumul plus de 70% de la population. La courbe d'adoption peut être modélisée par la courbe sigmoïde, ou courbe en S [6] Pour Rogers, le temps de l'adoption, pour chaque individu, comprend plusieurs phases temporelles:

- **une phase d'information**, durant laquelle l'individu est exposé à la nouveauté, et où sa réaction à l'innovation dépend essentiellement de ses caractéristiques vis-à-vis de l'innovation en général ;
- **une phase de persuasion**, durant laquelle l'individu commence à se positionner vis-à-vis de l'innovation, en fonction des caractéristiques de l'innovation dont il a connaissance (avantage relatif en terme économique et social, compatibilité avec les valeurs de son groupe d'appartenance, complexité de l'innovation, possibilité de tester l'innovation, visibilité de l'adoption vis-à-vis des autres) ;
- **une phase de décision**, durant laquelle soit il adopte l'innovation et peut expérimenter par lui-même les avantages et inconvénients de cette innovation, soit il la refuse, ce qui laisse ouvert son adoption future ;
- **une phase de confirmation**, durant laquelle il va renforcer son choix d'adoption ou de refus par des arguments complémentaires.

Comme on le voit, la décision d'adoption est essentiellement basée sur l'information et sur l'expérimentation propre de l'innovation. Par ailleurs, Rogers complète le processus en précisant qu'une adoption n'est jamais définitive, puisque l'utilisateur peut éventuellement revenir sur sa décision s'il constate des désagréments ultérieurement à sa décision initiale. La décision est par ailleurs basée sur l'utilité pratique mais aussi sociale.

Ce modèle théorique sert de base à beaucoup de stratégies marketing dans le domaine des technologies numériques. Il a également été mis à l'épreuve des données linguistiques, dans de nombreuses études mais de manière extrêmement simplifiée, puisqu'elles évaluent seulement la pertinence de la courbe sigmoïde (voir (Blythe et Croft, 2012) pour une revue et (Nevalainen, 2015) pour l'une des dernières études).

D'autres travaux, principalement issus des travaux fondateurs de (Labov, 1966; Labov, 1994; Labov, 2001), par le biais d'enquêtes, tentent de combiner l'approche purement quantitative avec des informations sur le rôle des différents individus (sont-ils des leaders d'opinion - des influenceurs- ou des suiveurs précoces ou tardifs ?) et la structure de leur réseau social (quelle place ont-ils dans leur réseau social, quels liens ont-ils avec d'autres réseaux ? etc.). L'étude conjointe de la structure des réseaux sociaux et des cheminements des innovations dans ces réseaux sera modélisé par (Milroy et Milroy, 1985). Nous reviendrons sur ces modèles lors de l'analyse des évolutions sémantiques.

6.
$$\theta(x) = \frac{1}{1 + e^{-\theta^T x}} \tag{2.6}$$

où θ est une fonction quelconque (ici la diffusion).

2.2.5.3 Approches cognitives et socio-cognitives

(Schmid, 2008; Schmid, 2017) présente une modélisation qui se base sur la notion d'implantation cognitive (*entrenchment*) (voir chapitre 1, section 1.4.2.2) pour établir des phases de la vie des néologismes. Il considère que trois perspectives doivent être prises : une perspective linguistique, une perspective cognitive et une perspective socio-pragmatique. Reprenant une conception classique, il considère que le cycle de vie des néologismes connait trois phases saillantes : l'émergence, la diffusion et l'adoption. mais pour chacune des perspectives, l'analyse est distincte : du point de vue linguistique, l'émergence est la première apparition ; durant l'éventuelle diffusion, la lexie se stabilise , au niveau de l'ensemble de ses propriétés ; l'adoption correspond à l'entrée dans l'usage de l'ensemble de la communauté ; du point de vue cognitif, l'implantation cognitive permet de passer du pseudo-concept au concept proprement dit ; du point de vue socio-pragmatique, il s'agit de prendre en compte la diffusion de l'innovation au sein de la communauté. On pourra aussi consulter (Winter-Froemel, 2013).

2.2.5.4 Conclusion

Les différents travaux présentés, peu ou prou, aboutissent à définir trois phases dans les changements lexicaux : l'émergence, la diffusion et la lexicalisation. Nous pouvons compléter cette vision, en tenant compte des phases de baisse d'usage - jusqu'à la disparition des mots et/ou des sens- en proposant six phases :

1. **Émergence** : le moment d'apparition d'une forme lexicale nouvelle, ou d'un usage nouveau ; cette apparition est située, c'est-à-dire qu'elle doit être liée à un énonciateur et à une situation de communication spécifiques ; dans la pratique, il est généralement complexe d'identifier le moment d'émergence, parce qu'il est probable que la majorité des innovations émergent d'abord à l'oral ou en tout cas dans des situations de proximité communicative (Koch et Oesterreicher, 1985). De ce fait, plutôt que d'identifier le moment d'émergence, il sera plus raisonnable de parler de période d'émergence, lorsque la lexie nouvelle apparaît dans un environnement discursif particulier ; cependant, *stricto sensu*, nous sommes déjà dans la phase suivante, d'autant plus si nous sommes dans un cadre énonciatif non oral, et ayant un auditoire étendu (par exemple l'audience d'un journal) ; une autre caractéristique de la période d'émergence est l'emploi de la lexie en mention, avec une glose : le terme nouveau étant ressenti comme nouveau, on éprouve le besoin de le gloser et de le marquer spécifiquement ;
2. **Diffusion** : la phase suivante est la diffusion, qui permet à l'innovation de sortir de son environnement discursif initial ; cette phase sera identifiable par des changements dans les situations de communication, un emploi hors du domaine d'origine, et par contrecoup, une fréquence accrue et plus dispersée ;
3. **Adoption** : l'adoption est le moment ou la période où l'on peut dire que la lexie n'est plus du tout ressentie comme nouvelle, par l'ensemble des membres de la communauté linguistique ; cette situation n'est en réalité que rarement rencontrée,

étant donné qu'il n'est pas raisonnable de penser qu'il existe une langue unique ou normée identique dans chacun des membres de la communauté ; les lexicologues s'appuient sur l'entrée dans les dictionnaires pour identifier l'adoption ; on pourrait ajouter d'autres signes de cette adoption : absence totale de marquage métalinguistique, stabilité linguistique de la paire forme-sens (d'un point de vue orthographique, morphosyntaxique : partie du discours, combinatoire et sémantique) ; stabilité de la fréquence (mais les événements extérieurs peuvent perturber cette stabilité) ;

4. **Préservation** : la phase de préservation est le moment durant laquelle la lexie adoptée se maintient sans évolution notable, ni au niveau de sa forme, ni au niveau de sa combinatoire, ni au niveau de son sens, ni au niveau de ses domaines d'application ; une modification de l'une de ces caractéristiques renvoie la lexie vers l'émergence d'un nouvel emploi, ou bien vers la phase de dégénérescence ;
5. **Dégénérescence** : cette phase est identifiable par une baisse de fréquence par rapport à la fréquence constatée lors de la phase de préservation ; il n'y a pas de création de sens nouveau, mais soit une perte de référence qui explique la baisse d'emploi, soit l'émergence d'un concurrent qui tend à remplacer la lexie en question ;
6. **Disparition** : il s'agit de la constatation de l'absence complète d'emploi.

Il faut noter que la vie des mots n'est pas limitée à un seul cycle des phases précédentes. nombre de mots disparaissent - ou semblent disparaître - puis reparaissent. Un exemple frappant est le verbe *pimper*, qui était présent en ancien français, seul *pimpant* ayant perduré en français moderne, jusqu'à ce que le verbe reparaisse par l'entremise de l'anglais (où le verbe *pimper* était passé par emprunt lors de la guerre de cent ans). Enfin, la triple perspective proposée par (Schmid, 2008) nous semble la plus adéquate pour rendre compte des différentes modalités de ce cycle de vie.

2.3 Conclusion et perspectives

Nous avons étudié dans ce chapitre les notions d'unité lexicale et d'innovation lexicale. La conception dynamique des langues s'accompagne d'un principe de continuité entre les unités linguistiques, qui ne peuvent être conçues que comme des prototypes que la tradition a diversement nommé. En partant d'une conception traditionnelle des grammaires du français, distinguant mot, phrase et texte, nous avons tenté de spécifier les caractéristiques spécifiques des unités lexicales, tout d'abord en traçant sa frontière supérieure, la proposition énoncée. Nous inspirant de la conception de (Adam, 1990; Adam, 2005), nous avons indiqué que cette unité combine une prédication, un point de vue énonciatif et une inscription contextuelle et co-textuelle. En deçà de la proposition énoncée, nous inspirant de la notion de *construction*, nous avons montré que l'unité lexicale, ou paire forme-sens, s'étend des morphèmes liés (flexion et affixe) aux morphèmes libres (les lexies au sens traditionnel) et au-delà aux unités polylexicales plus ou moins figées et aux constructions lexico-syntaxiques et syntaxiques. Les formes en sont donc

diverses, mais elles partagent toutes cette propriété d'associer une forme à une représentation mentale déterminée qui n'est pas de l'ordre de la proposition énoncée. Nous avons également établi quelques unes des propriétés prototypiques de ces différentes unités, qui ressortissent à trois dimensions : la forme elle-même, les propriétés morphosyntaxiques (partie du discours, formes flexionnelles éventuelles et règles de combinatoire) et la représentation mentale. Nous avons émis l'idée que les parties du discours ne sont que des abstractions de fonctionnements combinatoires prototypiques des lexies qui permettent leur inscription dans les prédications. Les langues isolantes disposent de lexies pures dont la combinatoire est matérialisée par des formes spécifiques isolées et l'ordre des mots permet les relations de dépendances entre les lexies, tandis que les langues agglutinantes agglomèrent ces informations sur les lexies proprement dites. Les langues casuelles et les langues flexionnelles proposent une situation mixte, notamment pour les dernières en agglomérant des propriétés sémantiques spécifiques et générales pour trois parties du discours (nom, adjectif et verbe). Il n'en reste pas moins que les unités lexicales doivent être décrites en distinguant les trois dimensions formelles, combinatoire (à défaut d'une meilleure dimension, car il s'agit là également d'une information sémantique) et sémantique ou représentationnelle. Nous avons ensuite décrit les particularités spécifiques des flexions, des affixes et des lexies proprement dites.

Après cette modélisation générale, nous avons établi les propriétés essentielles des innovations lexicales : il s'agit d'unités lexicales qui divergent par rapport à l'usage d'une communauté linguistique donnée, à la fois au niveau de la forme (qu'il s'agisse de la forme au sens classique, orthographique et morphologique, auquel cas nous avons des innovations formelles ; ou qu'il s'agisse d'une forme au sens d'une combinatoire inhabituelle, auquel cas nous avons avons une innovation sémantique, même si le terme ne rend pas justice aux modifications formelles) et au niveau du sens, puisqu'une nouvelle conception mentale apparaît. Nous avons également indiqué que nous considérons l'innovation lexicale dès le moment de son émergence et quel que soit son sort futur dans la langue, en limitant la dénomination au moment où l'unité lexicale est adoptée par l'ensemble d'une communauté linguistique et où donc la divergence n'est plus ressentie. Nous avons indiqué l'importance d'effectuer une typologie des procédés d'innovation lexicale, typologie qui sera présentée plus exhaustivement dans le chapitre 5. Enfin, nous avons présenté les principales phases du cycle de vie des innovations, l'émergence, la diffusion et l'adoption, en nous appuyant sur les points de vue lexicologique, sociolinguistique et psychologique.

Chapitre 3

Langue et société : langue, variations, variétés

Sommaire

La langue est un *fait social*, disait Saussure. Effectivement, les langues sont manipulées par des groupes d'individus, et leur existence dépend d'une communauté pour les préserver et les faire évoluer selon les besoins pratiques de la communication. Les flux d'informations linguistiques qui se produisent dans une communauté donnée sont à la source de la préservation de la langue et à la source de ses évolutions, mais il existe des variations dans les usages au niveau de l'individu comme des communautés qui rendent inadéquate la notion de langue unique.

Ce chapitre révisera donc la notion de langue en abordant les notions de **variation et de variété de langue** : à l'évidence, il existe dans toute langue vivante des variations, à

tous les niveaux : phonologique, morphologique, syntaxique, lexical et même au-delà avec les routines discursives. Quand cet ensemble de variations est suffisamment conséquent et lié à une sous-communauté linguistique déterminée, on peut alors parler de variété de langue. Plusieurs variétés, plusieurs normes coexistent au sein d'une même communauté linguistique, et chacun d'entre nous maîtrise plusieurs variétés de langues. Mais comment caractériser et identifier ces variétés, quels sont les rapports qu'elles entretiennent, et comment articuler cette notion de variété avec celle de langue unique, puisque nous avons malgré tout le sentiment de tous parler la même langue ?

On peut considérer une innovation lexicale comme étant d'abord une variation par rapport à une norme établie, et mieux comprendre les phénomènes de variations et de variété permettra par exemple de faire le départ entre ce qui relève du particularisme et de l'innovation : *amender* dans le sens de 'mettre une amende' est d'usage courant en Afrique francophone, mais constitue un particularisme de cette variété du français et non une innovation lexicale ; de même pour *char* et une multitude d'autres lexies en français du Canada. De plus, nombre d'innovations lexicales sont introduites par des groupes sociaux spécifiques, y restent parfois cantonnées, et parfois diffusent dans l'ensemble de la communauté par des chemins et des mécanismes qu'il convient de décrire. c'est par exemple le cas de nombre d'anglicismes actuels, qui émergent dans la presse féminine ou dans la presse informatique avant d'être adoptés par tous les corps sociaux. Comprendre qui sont les innovateurs, qui sont les diffuseurs et comment les membres de la communauté linguistique adoptent ou non des innovations est un pan de l'analyse du phénomène qui mérite donc notre attention, et pour cela une meilleure compréhension des phénomènes sociolinguistiques de variations, de variétés et de l'organisation générale de ces variétés est capitale.

Le propos sera organisé en quatre sections principales :

— nous présenterons tout d'abord les notions de variation et de variété, en nous appuyant principalement sur (Weinreich *et al.*, 1968) et surtout les travaux de Coséru (Coseriu, 1952; Coseriu, 1958; Coseriu, 1981), qui a posé les jalons pour l'analyse de ces phénomènes.

— puis nous aborderons une première manière de caractériser variations et variétés, par l'explicitation de dimensions de la variation : on doit à (Flydal, 1951), repris et complété par Coseriu dans de nombreux écrits (Coseriu, 1952; Coseriu, 1958; Coseriu, 1981; Coseriu, 1998), d'avoir explicité trois dimensions pour caractériser les variations : les dimensions diatopiques, diastratiques et diaphasiques ;

— nous évoquerons ensuite une autre perspective sur les variations, qui se veut complémentaire de la précédente, proposée par (Koch et Oesterreicher, 1985; Koch et Oesterreicher, 2001) ;

— enfin, nous aborderons une autre manière d'aborder la variation qui prend appui sur l'étude des flux de communication au sein du réseau social : ce sont en effet les flux de communication au sein et entre les groupes sociaux et les variétés qu'ils représentent qui permettent d'expliquer comment, à partir d'une innovation lexicale apparue ici ou là, elle se diffuse (éventuellement) au travers des différents groupes pour finalement être (éventuellement) adoptée par toute la communauté. La mo-

délisation de ces flux et des rôles des individus au sein de ces flux (innovateurs, diffuseurs, etc.) est donc essentielle à la bonne compréhension des changements notamment lexicaux.

3.1 Langue, variations, variétés

Le constat de l'existence de variétés de langue n'est pas une découverte récente, mais pendant très longtemps, la pensée dominante assignait une valeur "linguistique" aux seules langues véhiculaires (le latin, le grec, puis l'hébreu et l'arabe), puis, après leur mise en place, aux langues dites nationales, formées de façon autoritaire à partir de la Renaissance en Europe. Dès lors, les parlers locaux, appelés dialectes ou patois, ont généralement été dévalorisés. Un premier sursaut se produit dans la dernière période de la linguistique historique, avec la naissance de la dialectologie européenne (fin XIXème) : il s'agissait de caractériser les différents parlers d'une zone géographique déterminée [1]. Il faudra cependant attendre les travaux de (Labov, 1966; Weinreich *et al.*, 1968; Coseriu, 1952) pour voir une première élaboration du concept de variation et la naissance de l'école dite variationniste qui deviendra rapidement la sociolinguistique.

L'hypothèse fondamentale consiste à renverser la conception structuraliste homogène et unique de la langue. Au contraire, (Weinreich *et al.*, 1968) concevront les langues comme des *systèmes dynamiques* caractérisés par une « hétérogénéité ordonnée »(*orderly heterogeneity*) : il peut se présenter - et il se présente de façon continue - des divergences d'usage au sein de la communauté linguistique, et l'état d'hétérogénéité qui en découle se résout par l'abandon d'une des variantes au profit de l'autre, ou d'un équilibre des emplois (par exemple une répartition des contextes d'utilisation). (Coseriu, 1980, p.5) affirme également que : « Le locuteur (...) se trouve confronté, dans son expérience réelle, à l'état d'une langue historique, dont la synchronie est différenciée des points de vue diatopique, diastratique et diaphasique. Tout locuteur, s'il ne connaît pas la langue historique dans son ensemble, connaît, au moins jusqu'à un certain degré, plus d'un dialecte et plus d'un niveau de langue ; et tout locuteur maîtrise plusieurs styles de langue. » Il existe donc non pas une langue unique (le français, l'italien, l'espagnol, etc.) mais une série de variétés linguistiques qui coexistent et s'interpénètrent.

Ce principe de variabilité intrinsèque des langues est facile à constater : d'abord, chacun d'entre nous, même dans un environnement monolingue, est compétent dans plus d'un code linguistique : nous n'utilisons pas le même vocabulaire ni les mêmes formulations selon que nous sommes dans une situation intime, familiale, amicale ou professionnelle. À l'écrit, nous n'employons ni le même vocabulaire ni la même syntaxe s'il s'agit de rédiger une lettre pour l'administration, un blog ou un travail académique. Les différences parfois considérables entre l'oral et l'écrit sont une autre preuve de l'existence de différentes variétés d'une même langue. Ensuite, il existe à l'évidence des cercles

1. La première étude dialectologique est à attribuer à Johann Andreas Schmeller, avec ses *Dialects of Bavaria*, qui incluait un atlas linguistique. D'autres travaux suivront, aboutissant à des descriptions exhaustives des parlers locaux en Angleterre (English Dialect Dictionary, 1905, Josph Wright), en Allemagne (Deutsche SprachAtlas, 1926) en France (Atlas linguistique de la France, 1911, Gilléron).

sociaux disposant d'une variété de langue spécifique : la langue des jeunes, la langue des bobos ou des hipsters, la langue des geeks, la langue des bouchers, etc.

L'approche variationniste va mettre en évidence la variation et le changement linguistique en introduisant la notion de *variable linguistique*, lorsque « deux formes différentes permettent de dire "la même chose", c'est-à-dire lorsque deux signifiants ont le même signifié et que les différences qu'ils entretiennent ont une fonction autre, stylistique ou sociale. »(Calvet, 1998: p.76) Ces variables linguistiques touchent toutes les couches de la langue : phonologique (le r roulé ou non en français), syntaxique (*aller au coiffeur, aller chez le coiffeur*), lexicale (*bagnole, caisse, voiture, chiotte, etc.*), phraséologique/pragmatique (formule de politesse en fin de courrier : *je vous prie d'agréer, veuillez recevoir,* , *bonjour, salut, hello, etc.*), etc. La variation et le changement linguistique sont deux facettes du même phénomène : en synchronie, il y a des variations, qui sont ensuite éventuellement résolues en diachronie par l'adoption d'une des variantes, ou une redistribution du champ sémantique. Deux méthodes de mise en évidence de ces changements linguistiques sont disponibles : la première, qui sera introduite par les variationnistes, travaille sur le "temps apparent", c'est-à-dire sur les usages, à un moment donné du temps, de groupes de populations distincts (par âge, par classe sociale, etc.) ; la seconde, dite méthode par le "temps réel" consiste à étudier des corpus d'usages entre deux périodes temporelles, et a été introduite par la linguistique historique.

Pour caractériser les variations et les variétés, deux voies ont été explorées : la première a consisté à établir une série de dimensions puis, pour chaque dimension, à mettre en place une méthodologie pour établir les propriétés permettant de caractériser les communautés linguistiques concernées et les traits linguistiques pertinents. Il s'agit alors, principalement par sondage auprès des locuteurs ou d'intermédiaires culturels [2], de s'informer sur leurs caractéristiques socio-économico-professionnelles et leurs usages linguistiques, puis d'agréger les données recueillies pour identifier les communautés linguistiques semblables et dissemblables, et éventuellement établir les particularités qui se propagent d'un groupe à l'autre. C'est la voie suivie dès le départ par la dialectologie, puis par la sociolinguistique labovienne. La seconde voie, bien plus récente - et peut-être issue pour partie des apories rencontrées par la première méthode - consiste à considérer les communications linguistiques en elles-mêmes, et à étudier les messages qui circulent, les réseaux qui se constituent et se défont, les acteurs et leurs rôles, pour obtenir une vision des variations qui émergent et se diffusent et les réseaux concernés. Nous présenterons dans cette section la première approche. La seconde fera l'objet d'une section distincte.

3.2 Dimensions de la variation

Dans la première méthode, la caractérisation des variétés de langue s'appuie d'abord sur une typologie des dimensions à prendre en compte. On trouve une première formu-

2. (Gadet, 2007) désigne par là les prêtres locaux qui ont compilé les données permettant à l'Abbé Grégoire en 1792 de se faire une idée des compétences linguistiques de la population française, et plus tard les maîtres d'école joueront ce rôle.

lation de ces dimensions chez Coseriu[3], qui les empruntent à (Flydal, 1951) avec les notions de diatopie (liée à la proximité géographique, aboutissant à des dialectes) et de diastratie (liée à l'appartenance à un ou à des groupes sociaux spécifiques, la variation aboutissant à des sociolectes), auxquelles il ajoute la diaphasie (liée à l'individu dans différentes situations de communication, et qui aboutissent aux *styles*, ou registres de langue).

Coseriu détaille ces trois dimensions :

> 3.1.1. [...] dans chaque langue historique on constate normalement trois grands types de différenciation interne : a) différences dans l'espace géographique ou *différences diatopiques* ; b) différences entre les diverses couches socioculturelles de la la communauté linguistique ou *différences diastratiques* et c) différences entre les types de modalité expressive correspondant aux circonstances constantes de la parole (locuteur, destinataire, situation ou occasion de la parole et "chose" dont on parle) ou *différences diaphatiques*. (Coseriu, 1998, p.28)

Il explicite ensuite les dimensions correspondantes en synchronie :

> 3.1.2. À ces trois types de différences correspondent dans le sens inverse (donc, dans le sens de la convergence et de l'homogénéité des traditions langagières) trois types de systèmes unitaires (ou, au moins, plus ou moins unitaires) d'isoglosses, à savoir : les unités *syntopiques*, que l'on peut continuer à appeler *dialectes*, étant donné qu'elles sont effectivement un type particulier de "dialectes" ; les unités *synstratiques ou niveaux de langues* (par exemple : "langage cultivé", "langage moyen", "langage populaire", etc.) ; et les unités *symphatiques* ou *styles de langue* (par exemple : "langage familier", langage formel", etc.). Au domaine des styles de langue appartiennent aussi les langages des "groupes" qu'on peut distinguer à l'intérieur du même niveau socioculturel (ou indépendamment des niveaux) : d'une part, les "langages" des grands groupes "biologiques" ("langage des hommes", "langage des femmes", très différents dans certaines communautés) et des générations ("langage des adultes", "langage des enfants", etc.), d'autre part, les "langages" des groupes sociaux ou professionnels. Les types très généraux de styles apparentés correspondant à des aspects généraux de la vie et de la culture et à des types similaires de circonstances (par exemple : "langue parlée", "langue écrite", "langue littéraire") peuvent être appelés *registres idiomatiques*. (Coseriu, 1998, p.28-29)

Les trois dimensions permettent donc de définir des *unités de langue*, marquées par une position spécifique dans l'une des dimensions, mais cela n'implique nullement la même unité pour les deux autres dimensions : une variété diatopique peut se scinder en

3. Il est difficile, étant l'immense bibliographie de Coseriu, et son caractère partiellement parcellaire, de retrouver les premières mentions de ces notions et/ou les écrits les plus significatifs à leur propos. Citons, en français et en espagnol (Coseriu, 1981; Coseriu, 1958; Coseriu, 1982; Coseriu, 1980; Coseriu, 1998). Nous renvoyons au site des *Archives Coseriu* à Tübingen pour une bibliographie complète : `http://www.coseriu.de/`

différentes unités aux niveaux diastratiques et diaphasiques, ce qui est le cas de la plupart des langues nationales. Lorsque une unité se créé sur les trois dimensions, on parlera alors de variété de langue ou encore langue fonctionnelle. Ces langues fonctionnelles sont des réalisations concrètes de ce qu'il nomme des langues historiques : « Une langue historique est une dénomination pratique pour parler d'une langue d'un point de vue abstrait. mais elle se décompose en langues fonctionnelles, correspondant à des valeurs unitaires au niveau des dimensions de la variation [...] une langue fonctionnelle est un système autosuffisant minimal à l'intérieur d'une langue historique « (Coseriu, 1998, p.29). La notion de langue fonctionnelle (qu'il appelle également isoglosse) est alors une unité pouvant être caractérisée selon les trois dimensions de la variation et différentes langues fonctionnelles peuvent être regroupées au sein de la notion de langue historique, correspondant notamment pour les langues de l'Europe occidentale aux langues dites nationales.

Coseriu détaille ensuite l'ontogenèse de ces variétés et de la langue historique : une langue historique constitue d'abord un des dialectes. Par exemple, le français comme langue historique était primitivement un des dialectes parlés, celui de la cour. Cette variété est ensuite devenue une langue standard (ou commune). Elle peut ensuite se re-décomposer en dialectes. Par exemple, pour le français, se sont constitués des régions de la francophonie. Coseriu appelle dialectes primaires, les dialectes sans forme commune, dialectes secondaires les dialectes ayant une forme commune et dialectes tertiaires les dialectes issus d'une langue standard.

La formation des langues historiques par les dimensions de la variation est "orientée" ou "ordonnée" : la dimension principale est la diatopie (le dialecte), puis vient la diastratie (le niveau), puis la diaphasie (le style) : les langues historiques se forment d'abord sur un territoire donné, puis forment une norme diastratique standard, la diaphasie restant généralement en situation de variation. Les langues fonctionnelles permettent ensuite de rendre compte d'unités aux niveaux diastratiques et diaphasiques. Il est également possible d'avoir des langues historiques qui deviennent des niveaux diastratiques, par exemple la langue française en Angleterre dans l'aristocratie aux XIIIème et XIVème siécles, ou le latin pour le clergé durant tout le moyen-âge.

Les dialectes forment des systèmes complets (ils sont unifiés dans tous les aspects de la langue), tandis que les niveaux et les styles sont en général des systèmes incomplets, car ils prennent pour substrat une langue fonctionnelle au niveau diatopique et ne se distinguent que par quelques traits spécifiques.

Essayons maintenant de préciser ce qu'il faut entendre par diatopie, diastratie et diaphasie.

3.2.1 Dimension diatopique

La dimension diatopique, aboutissant à la création de topolectes (ou dialectes, au sens strict du terme, et c'est le terme qu'emploiera Coseriu de manière préférentielle), a fait l'objet des premiers travaux linguistiques, avec la dialectologie : il s'agissait, par sondages auprès de populations géographiquement circonscrites de déterminer des variétés ou isoglosses, de langues (appelées patois ou dialectes). Les dialectologues matérialisent

les caractéristiques linguistiques communes par des courbes d'isoglosses marquant territorialement les différentes variantes. Ces isoglosses correspondent très souvent à des frontières naturelles (montagne, cours d'eau, etc.) et/ou humaines (formation d'un diocèse, d'une municipalité). Très souvent également, les tracés isoglossiques explicitent des frontières floues.

On pourrait penser que cette dimension, au moins dans les sociétés urbanisées, est la moins pertinente pour caractériser des variétés : d'une part, la mobilité humaine s'est considérablement développée avec l'exode rural massif et global dès l'entre deux guerres et l'avènement des moyens de transport (voiture, train, avion), aboutissant à des sociétés essentiellement urbaines et donc cosmopolites ; les moyens de communication actuels rendent également l'isolement géographique beaucoup moins prégnant : télévision, radio, puis communications numériques amoindrissent considérablement l'impact de la géographie sur la naissance des variétés linguistiques. Cependant, en dépit de ces modifications techno-socio-culturelles, il n'en reste pas moins que le critère de proximité géographique joue toujours un rôle de premier plan dans la construction de variétés de langue : nous passons certes de plus en plus de temps devant nos ordinateurs, mais nous passons tout de même encore la très grande majorité de notre temps dans un lieu géographique relativement restreint : il existe bien des variantes du français parlé en régions, à l'évidence, comme il existe des variétés du français parlé dans les différentes régions du monde où cette langue est implantée. Les études de (Labov *et al.*, 2008) pour construire l'*Atlas of North American English* (ANAE) montrent bien également que la dimension géographique reste prédominante dans l'existence des variétés. Les travaux de (Gadet, 2007) sur les variations en français également. Dans beaucoup de régions du monde, enfin, la dimension géographique reste prédominante dans la création et la préservation de variétés. On pourra consulter (Britain, 2010) pour une description détaillée des méthodes et acquis des travaux de sociolinguistique liés à la diatopie. On consultera (Reutner, 2017) pour la situation du français à cet égard.

3.2.2 Dimension diastratique

La différenciation géographique est le premier déterminant de la création de variétés. La dimension diastratique permet ensuite de déterminer des sociolectes, c'est-à-dire des variétés corrélées à des sous-groupes *sociologiquement marqués* de la communauté linguistique, en relation étroite avec l'organisation interne des sociétés. Mais comment déterminer les caractéristiques de ces sous-groupes ? S'agit-il de caractéristiques biologiques (âge, sexe, race etc.) ethnico-culturelles (origine, appartenance à une communauté partageant des pratiques communes), économiques (revenu, propriété foncière, etc.) ?

Pour cette présentation, nous nous sommes appuyés notamment sur les analyses proposés dans (Ritzer, 2004), (Garcia *et al.*, 2017) et (Chambers et Schilling-Estes, 2013). L'analyse de (Eckert, 2012), notamment les deux premières vagues des travaux en sociolinguistiques (les études macro-structurelles essentiellement basées sur la notion de classe sociale, puis des études locales essentiellement basées sur l'ethnicité et les communautés de pratiques) correspondent aux deux pans que nous allons évoquer.

Coseriu parle de la diastratie comme marquant un niveau de langue (voir citation

ci-dessus), ce qui s'explique par la prégnance de la notion de classe sociale à l'époque : ce mode de découpage de la société en classes a été introduite par Karl Marx - qui la basait exclusivement sur la propriété foncière et financière - a été par la suite affinée par Max Weber (1864–1920) qui a élaboré une caractérisation multidimensionnelle (mais essentiellement socio-économique) combinant la propriété et d'autres facteurs (pouvoir, prestige et nombre d'interactions sociales). Dans les années 30, en sociologie, Lloyd Warner a élaboré des grilles de paramètres au travers d'une étude minutieuse des strates de la société américaine, en prenant la ville de Yankee comme échantillon représentatif (1930-1934), qui a aboutit à l'*Index of Status Characteristics* (ISC) (Warner, 1960). Les grilles ont été raffinées, jusqu'aux travaux du National Opinion Research Council (NORC) aboutissant à définir près de 500 activités socio-professionnelles. Ces échelles socioéconomiques ont été utilisées pour déterminer la variation et les changements linguistiques dans les premiers travaux de la sociolinguistique : (Labov, 1966), pour étudier les variations phonologiques dans le East Side à New York, a placé chacun des individus sur une échelle comprenant dix "classes sociales", déterminées par trois facteurs principaux : le nombre d'années d'études, l'activité professionnelle du chef de famille et le revenu du ménage.

D'autres études suivront (aux États-Unis et en Angleterre : Wolfram 1969, Trudgill 1974, Macaulay 1977) à Panama (Cedergren 1973) en Iran (Modaressi 1978). utilisant les mêmes facteurs (en suivant les classements)proposés par le NORC). Dans ces études, l'objectif est de corréler une variété linguistique (*vernacular*) à une classe sociale déterminée socio-économiquement, les classes les plus basses étant généralement corrélées à des variétés stigmatisées, les classes les plus hautes à des variétés standard "prestigieuses". Un second enseignement de ces études concerne les acteurs du changement et sa directionnalité : dans les classes laborieuses et moyennes, certains membres sont en phase d'ascension sociale : ils font toujours partie de leur communauté d'origine, mais ont également accès aux communautés socio-économiquement plus avancées, et leurs interactions avec ces derniers groupes en font des vecteurs de diffusion d'innovations d'un groupe à l'autre. Les classes moyennes, et dans ce groupe les individus en phase d'ascension sociale, seraient les principaux vecteurs de la diffusion de variations ou d'innovations, même si la vision binaire des variétés de Labov (vernaculaire versus langue standard) implique un rejet des variétés marquées socialement. Les changements sont directionnels, et reposent sur l'opposition vernaculaire (stigmatisé) / langue standard (valorisé) : les acteurs du changement, dans leur entreprise d'ascension sociale, vont adopter les usages de la langue valorisée.

En partant d'un état initial des variations au sein des communautés, deux types de changements linguistiques ont été identifiés : le changement « par dessus »(*change from above*), qui se produit par le contact avec d'autres communautés linguistiques, audelà de la conscience sociale de la classe, et le changement « par en-dessous »(*change from below*), qui se produit au sein même de la communauté linguistique. Deux facteurs principaux influent sur le premier changement, comme signalé plus haut : le prestige ou la stigmatisation attaché à certaines formes linguistiques, qui entraînent un rejet ou une adoption. Pour ce qui concerne le second type, “there is no important distinction between stigmatized and prestige forms: the speech form assumed by each group may

be taken as an unconscious mark of self-identification" (Labov, 1966, p.331). Il s'agit alors d'affirmer l'identité de son groupe, en réintroduisant des spécificités linguistique éventuellement oubliées (voir le cas de l'île de Martha's Vineyard, où les pêcheurs ont réintroduit la centralisation des deux diphtongues /ay/ et /aw/).

En dehors des caractéristiques socio-économiques, ces études ont également étudiées les variations et le changement en liaison avec les caractéristiques biologiques (sexe et âge). (Wolfram, 1969 ; Trugdill, 1974 ; Macaulay, 1977) ont ainsi montré, dans des populations américaines et anglaises, que les femmes de basse couche sociale ont une tendance plus marquée à utiliser la langue standard que les hommes, ce qui dénoterait la tendance plus grande des femmes à s'élever socialement et donc à se conformer aux pressions sociales des couches plus élevées.

Les premières études sont généralement des macro-études, cherchant à identifier différentes variétés dans un lieu géographique cosmopolite - urbain -, plus favorable à la coexistence de différentes variétés, de par la forte immigration et l'urbanisation intense. Elles utilisent quasiment exclusivement des paramètres biologiques et socio-économiques.

Par la suite, des études sur des groupes beaucoup plus restreints seront menées), pour montrer la dynamique variationnelle locale. Les premières études dans ce sens sont dues à (Milroy et Milroy, 1978) qui a étudié les variations phonologiques dans des réseaux sociaux à Belfast. L'approche essaie de montrer, contrairement aux approches précédentes, que l'utilisation des vernaculaires n'est pas juste le reflet passif, inconscient de l'appartenance à une classe sociale, mais au contraire l'affirmation consciente d'une identité revendiquée. Elle montre ainsi que les individus disposant d'un réseau social dense et diversifié ont une tendance marquée à revendiquer et utiliser les traits phonologiques identifiant leur classe sociale d'origine, et donc à les diffuser. D'autres études iront dans le même sens (par exemple : (Edwards, 1992; Knack, 1991; Rickford, 1986)) montrant que si les vernaculaires sont stigmatisés au niveau global, ils sont valorisés au niveau local. Ces études aboutiront au concept de communautés de pratiques (Lave *et al.*, 1991; Wenger, 1998). Une communauté de pratiques est un groupe social défini par un engagement mutuel de ses participants, un ou des objectifs communs et un répertoire linguistique commun. Par exemple, un groupe de jeunes ou des collègues de travail ayant régulièrement des activités communes, les membres d'un club sportif sont autant de communautés de pratiques. Ces communautés sont généralement locales, mais plusieurs études ont montré qu'elles peuvent être locales mais reproduites en d'autres lieux dans des conditions sociologiques similaires (exemple des groupes de jeunes ayant des conditions de vie similaires dans les banlieues par exemple (Irvine et Gal, 2009)) ou même être totalement virtuelles et n'avoir aucune délimitation géographique (exemples des adeptes de certains jeux électroniques, voire les réseaux sociaux numériques). Au sein de ces groupes, il peut y avoir des leaders, mais le caractère non contraignant fait que les membres peuvent y être en position périphérique. Chaque membre fait souvent partie de plusieurs communautés, ces communautés elles-mêmes ayant une existence éventuellement fugace. (Lave *et al.*, 1991, p.52-53) indique clairement que cette approche « suggests a very explicit focus on the person, but as person-in-the-world, as member of a sociocultural community ».

Toutes les études présentées pour étudier la dimension diastratique cherchent à établir des corrélations entre des traits linguistiques et l'appartenance à un groupe social, et à étudier les mécanismes du changement linguistique. Que ce soit au niveau de macro-caractéristiques structurants les sociétés humaines, ou de micro-caractéristiques agissant au niveau plus local. ces études ont mis au jour un certain nombre de paramètres pour caractériser des communautés sociolinguistiques : paramètres biologiques (âge, sexe, etc.), ethnico-culturels (race, appartenance ethnique marquée par des pratiques spécifiques) ou socio-économiques (revenu et type d'activité professionnelle) (Mallinson, 2007). À un niveau structurel, on cherche à identifier des classes socio-économico-culturelles, à un niveau local des communautés de pratiques, avec tout un continuum entre ces deux pôles.

Cependant cette approche présente plusieurs difficultés :

— tout d'abord, d'un point de vue méthodologique, il paraît difficile de stabiliser les caractéristiques des individus et des groupes à prendre en compte : ces paramètres dépendent-ils essentiellement des conditions locales d'organisation des sociétés ? Ces caractéristiques ne sont-elles pas variables également dans le temps, puisque les sociétés évoluent, et les spécificités linguistiques suivent (ou mènent ?) ce mouvement ? Quoi qu'il en soit, aujourd'hui, les études menées sont principalement effectuées par questionnaires et sondages auprès de communautés restreintes et même si les résultats semblent convaincants sur les phénomènes linguistiques étudiés, on doit se demander si et comment on pourrait généraliser ces approches, notamment dans une perspective computationnelle ;
— Ensuite, la méthode définit, de par les questions qui sont posées aux sujets, *a priori* les caractéristiques sociales qui seront mises en corrélation avec les phénomènes linguistiques, et cette méthode hypothético-déductive biaise sans doute la réalité des corrélations entre langue et société ;
— De plus, du point de vue des paramètres de classification des groupes, articulés autour de la notion de classe sociale, ils ont des sous-bassements idéologiques clairs : les premiers travaux pré-définissaient ces classes selon des critères adaptés aux pays économiquement développés marqués par les idéologies marxiste puis capitaliste. De même, la vision de la structure familiale, dans laquelle le chef de famille était généralement considéré comme le pourvoyeur de revenus (alors que les études ont montré que, dans nombre de cas, le salaire des femmes est supérieur, pour un niveau d'études équivalent) est sans doute adapté à l'étude des sociétés occidentales de l'après-guerre, mais est sans doute moins pertinente dans les sociétés contemporaines et dans d'autres sociétés ;
— Enfin, de manière plus globale, il s'agit chaque fois d'une caractérisation parcellaire des individus, alors qu'à l'évidence, chaque individu n'est pas identifiable seulement par son appartenance à une classe sociale, mais également par ses interactions individuelles avec d'autres classes sociales. Cela a amené plusieurs autres modèles, dits relationnels, tenant compte non pas de la position statique de chacun déterminée par des facteurs socio-économiques et biologiques, mais par les relations entretenues vis-à-vis de sa propre classe et des classes dans l'envi-

ronnement (voir entrée classe sociale dans (Ritzer, 2004) pour une typologie des approches et (Mallinson, 2007) pour une analyse complémentaire).

On remarquera finalement que, historiquement, on est passé d'études macro-structurelles à des études locales, aboutissant à la notion de communauté de pratiques, qui permet d'identifier des groupes instables et restreints, et rapproche la caractérisation des variations et des changements du facteur *individu*. on peut se demander si l'évolution de ces approches n'est pas liée à l'évolution de la société elle-même, de moins en moins définie sociolinguistiquement - en tout cas dans les sociétés occidentales, par la géographie et par les classes sociales, mais par des communautés de pratiques diverses et non pérennes. Le même mouvement s'accompagnerait d'un effacement progressif des variétés marquées de langue, par leur uniformisation dans une langue standard, des variétés moins clairement identifiables (et clairement incomplètes, dans les termes de Coseriu) subsistant par les communautés de pratiques.

Avec les études locales, on voit également apparaître une approche différente des variations et des variétés : dans ces dernières, il s'agit moins de construire des modèles de classes sociales, mais à considérer le réseau social comme une structure communicationnelle, et d'étudier les flux d'information entre les membres, dont la densité permet d'identifier des groupes. Dans ce cas, on peut déterminer les rôles divers que peuvent y jouer les individus et les groupes sociaux qu'ils peuvent représenter, selon la structure de leur réseau social. Nous y reviendrons dans une section spécifique.

3.2.3 Dimension diaphasique

Cette dimension correspond, selon les termes de Coseriu, à un « registre de langue ». 'A notre connaissance, Coseriu n'a pas détaillé la diaphasie. Il utilise régulièrement la notion de registre, notion couramment utilisée, notamment dans les grammaires, avec différents pôles orientés : registres vulgaire, argotique, populaire, relâché, familier, courant, soutenu, formel etc. Nous passons ici à une dimension essentiellement liée au locuteur, tandis que les deux autres portent sur la localisation géographique et la seconde sur des regroupements de locuteurs. Certains chercheurs, étant donné le lien de cette dimension avec les situations de discours (un même individu, selon la situation linguistique, pouvant user de tel ou tel registre), préfèrent parler de dimension *diasituationnelle* (Halliday, 1978). Nous verrons en effet qu'il existe un flou conceptuel sur l'objet de la diaphasie. La conceptualisation théorique de la notion de registre (ou style, dans la sphère anglo-saxonne) est assez pauvre en linguistique théorique, mais a été abordé dans le cadre de l'analyse du discours ordinaire (Moirand, 2007, Charaudeau 1997, Boutet 1995) et dans les travaux sur les genres (Bronckart 1996, Adam 1999). La sociolinguistique a de son côté élaboré différents modèles pour rendre compte de cette dimension.

La théorie initiale dite ***Attention to Speech***, est dûe à (Labov, 1972) qui considère le registre (ou style) comme l'expression d'une plus ou moins forte formalité du discours, essentiellement conditionnée par une adaptation à l'auditoire. Il s'agit d'une dimension secondaire par rapport aux deux premières dimensions, qui est notamment abordée méthodologiquement pour la mise en œuvre des questionnaires : « To obtain the data most important for linguistic theory, we have to observe how people speak when

they are not being observed. »(Labov, 1972, p.113). En effet, le principe de l'adaptation à l'auditoire s'applique dès l'enquête : les sondés ne vont répondre avec leur langage habituel (le vernaculaire) que s'ils n'ont pas conscience d'être interrogés [4]. L'objectif de Labov est en effet d'atteindre la langue la plus usuelle des locuteurs, la moins contrôlée, ce qu'il appelle la langue vernaculaire, qui serait la plus représentative de chacun des locuteurs (et de leur groupe social). Les autres styles seraient donc de simples contraintes de formulation liées aux situations de communication, elles-mêmes liées aux relations socio-économiques sous-jacentes entre classes, et le style est exclusivement défini sur l'axe formalité-non-formalité. Ce modèle a été critiqué sur plusieurs points, notamment son unidimensionnalité - voire sa binarité - et sur la primauté qu'il donne au langage non-contrôlé (ou inconscient), alors qu'à l'évidence, un locuteur maîtrise plus d'un style, et ce répertoire de styles fait partie de ses caractéristiques.

Pour prendre en compte d'autres paramètres définissant le style, et rendre compte du caractère volontaire et actif des compétences stylistiques, des modèles d'adaptation au discours (*Speech Accomodation Model*) ont été proposées (Giles et Powesland, 1975), dans lesquels on considère le style comme le résultat d'une adaptation du discours en fonction des interlocuteurs et des objectifs interactionnels. Cette approche sera généralisée par (Bell, 1984) puis affinée dans (Bell, 2001) dans la théorie dite ***Style as Audience Design***, dans laquelle l'audience des discours est le facteur déterminant des styles. Cette théorie détaille trois types d'interlocuteurs, ayant chacun une influence sur le style du locuteur : les auditeurs proprement dits (*auditors*), à savoir les interlocuteurs présents durant le discours et à qui s'adresse directement le discours, les auditeurs lointains (*overhearers*, c'est-à-dire les personnes présentes et qui peuvent entendre, mais ne font pas partie des personnes à qui s'adresse directement le discours, et les *eavesdroppers*, c'est-à-dire les personnes non présentes mais qui pourraient l'être. Ces distinctions permettent par exemple de modéliser une situation de communication entre deux interlocuteurs dans un restaurant. Ce modèle général a notamment été appliqué pour rendre compte des variations de style de présentateurs radio, selon qu'ils s'adressaient à un public national ou local. En dehors de la modélisation de l'audience, le sujet, les circonstances du discours et le média utilisé sont également pris en compte, mais en tant que facteurs secondaires. Ce modèle laisse ouvertes plusieurs questions : dans quelle mesure un locuteur s'adapte-t-il à son auditoire et en considérant quelles caractéristiques de celui-ci (ou de ceux-ci) ? On peut se poser plus globalement la question de la part respective du style comme réaction à la structure de l'auditoire et comme initiative : on peut en effet concevoir que dans nombres de cas, le locuteur va chercher à avoir une influence sur son auditoire, en effectuant non pas une adaptation pure et simple, mais une adaptation liée à ses objectifs de communication.

Pour rendre compte des objectifs du locuteur lui-même dans les interactions, un troisième modèle, dit du Locuteur (***Speaker Design***) a été proposé. déjà, pour rendre compte de certaines situations, (Bell, 2001) avait ajouté un nouveau paramètre, l'initia-

4. D'où la fameuse question de mise en danger pour faire oublier la situation de sondage : "Have you ever been in a situation where you were in serious danger of being killed, where you thought to yourself, This is it ?"

tive, mais qui restait secondaire. Dans le nouveau modèle, Il s'agit de détailler les motivations du locuteur dans les interactions verbales, et les conséquences que cela entraîne sur le ou les registres qu'il peut utiliser. Dans ce cadre, le style devient un paramètre central dans l'étude de la variation, puisqu'il devient le moyen de communiquer une identité revendiquée de manière consciente, plutôt que la simple expression d'une identité et d'une appartenance communautaire inconsciente. Dans le second modèle, en effet, le focus est essentiellement sur l'auditoire et sur ce qu'il représente, et le style une *adaptation* à ses caractéristiques socio-économico-culturelles. Dans ce troisième modèle, l'individu, par son style, ne rend pas compte des caractéristiques socio-économico-culturelles de sa classe sociale, mais au contraire construit sa propre identité et éventuellement des communautés de pratiques ((Wenger, 1998)). Cette dernière approche renverse donc complètement l'ordre entre les dimensions : le diatopique, le diastratique et le diasituationnel sont des dimensions qui doivent être combinées, mais dont l'importance peut varier d'une époque à l'autre, et selon les situations linguistiques.(Gadet, 1998, p.62) soutient, par exemple, que la dimension stylistique et les communautés de pratiques sont beaucoup plus importantes pour l'identification de variétés dans le domaine français à l'ère contemporaine. La dimension diatopique, jusqu'à l'ère pré-industrielle, était la dimension dominante pour caractériser les variétés, puis, avec l'urbanisation et l'industrialisation, et l'importance des relations sociales (et des conflits sociaux), le diastratique. Avec l'avènement d'une classe moyenne plus large, et la moindre compartimentation des classes, la dimension diaphasique a pris le pas, notamment depuis les années 60, en tout cas dans les sociétés occidentales. Pour plus de détail sur cette dernière approche, on consultera (Eckert, 2012) et (Chambers et Schilling-Estes, 2013).

Ces différents modèles aboutissent à une conception du style qui se débarrasse de l'axe formel-non formel, comme de l'axe traditionnel soutenu-relâché (populaire). En combinant le second et le troisième modèle, on caractérise le style par l'ensemble des paramètres des situations de discours : l'émetteur (sa volonté, son savoir linguistique et encyclopédique), le locuteur et l'auditoire, et certains paramètres de la situation de communication. Cependant, cette situation de communication n'est pas modélisée en tant que telle et dans son fonctionnement interne. Il faudra se tourner vers l'analyse du discours et les travaux en linguistique textuelle pour voir émerger des modèles permettant de définir les différentes situations de communication. c'est dans ce cadre que nous abordons une approche présentée par ses auteurs comme s'intégrant au modèle cosérien, mais qui nous semble devoir entrer dans une dimension plus intimement liée aux situations de communication dans leur structure interne.

3.3 Proximité et distance : le modèle de Koch et Oesterreicher

3.3.1 Présentation du modèle

En 1985, paraît (Koch et Oesterreicher, 1985) qui sera ensuite traduit en français (Koch et Oesterreicher, 2001) [5]. Les auteurs se placent résolument dans la lignée des travaux de Coseriu, et proposent une nouvelle dimension pour caractériser les langues fonctionnelles, la dimension de la proximité-distance communicative [6].

Les auteurs partent de la distinction entre les codes écrit et parlé, qui est généralement associée à des variations marquées dans les langues, le code parlé étant considéré comme plus relâché et le code écrit, plus formel (voir par exemple la dimension diamésique proposée par (Gadet et Guérin, 2008)). Ils considèrent que cette assimilation n'est pas valide, notamment parce qu'il existe des styles écrits relâchés voire vulgaires (par exemple les romans de Céline, San Antonio, etc.) et que certaines pratiques orales sont formelles (une soutenance de thèse, une plaidoirie d'avocat, etc.). Pour résoudre cette aporie, il faut distinguer deux niveaux d'analyse, l'un lié au médium utilisé ou aspect médial (code parlé et code écrit) et un aspect conceptionnel (ou « allure linguistique »du texte) qui définit une distance communicative. Dans l'analyse médiale, il y a une opposition discrète entre l'écrit et le parlé, tandis que du point de vue de la distance communicative, il y a un continuum entre l'immédiat communicatif et la distance communicative. Certes, le parlé est par nature plus enclin à la proximité communicative, et l'écrit plus enclin par nature, à la distance communicative. Mais dans moultes situations parlées, il y a distance communicative (communication avec un supérieur hiérarchique, dans une situation normée, etc.) et, de même, dans moultes situations écrites, nous sommes dans la proximité communicative (lettre intime, roman autobiographique etc.).

Pour rendre compte de cette conceptualisation, les auteurs proposent un schéma récapitulatif (figure 3.1) :

Dans ce schéma, les codes sont bien distincts (partie haute et partie basse du schéma séparées par la ligne horizontale). Par contre la dimension conceptionnelle forme un continuum dont les extrêmes sont l'immédiat communicatif et la distance communicative. En prenant cet axe comme référence, on peut placer l'ensemble des situations de communications, en croisant la distance et le médium. Les lettres minuscules indiquent certaines situations, par exemple : (a) conversation spontanée entre amis, (c) lettre personnelle entre amis, (e) interview de presse, (g) conférence scientifique, (a') transcription d'une conversation spontanée entre amis, (i') lecture à haute voix d'un texte de loi. Les lettres majuscules dessinent des secteurs (A : code oral + proximité, B: code oral + distance, C: code écrit + proximité, C : code écrit + distance). Évidemment, les langues seulement orales ne sont pas concernées par la partie haute du schéma. Il faut dès à

5. Le sujet sera également abordé dans (Koch et Oesterreicher, 1990) et le texte initial traduit en espagnol et en roumain.

6. En allemand, les deux notions sont *NäheSprache/DistanzSprache* mais la traduction française du premier terme est *immédiat (communicatif)*. Nous renvoyons à (Krefeld, 2015) pour une discussion sur ce point, nous emploierons pour notre part indifféremment proximité ou immédiat communicatif.

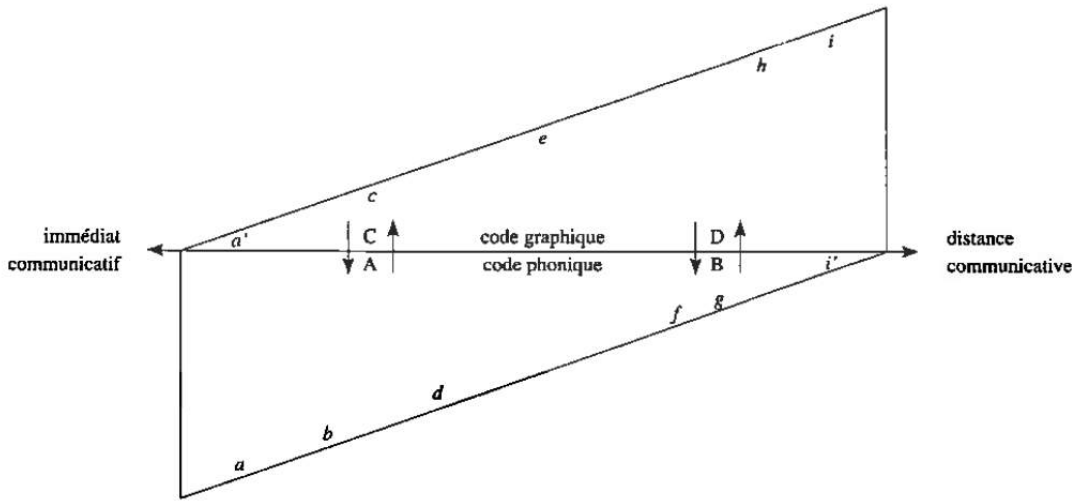


FIGURE 3.1 – Immédiat communicatif/distance communicative et code phonique/graphique (Koch et Oesterreicher, 2001, p.586)

présent noter que ce schéma reconnaît au code oral de permettre une proximité 'absolue" (dans la conversation entre amis ou intime par exemple), le code écrit ne pouvant se situer que dans le continuum de la proximité (par la transcription ou l'imitation de la conversation orale, mais qui perd alors son caractère d'immédiateté). C'est pour cela que Koch et Oesterreicher conservent dans leur analyse de la proximité les deux niveaux, celui du médium et celui de l'attitude conceptionnelle.

Pour caractériser chaque situation de communication, les deux auteurs proposent un certain nombre de paramètres, présentés également comme des axes continus (voir figure 3.2, nous reprenons la légende de la figure originale) :

① communication privée	communication publique ❶
② interlocuteur intime	interlocuteur inconnu ❷
③ émotionnalité forte	émotionnalité faible ❸
④ ancrage actionnel et situationnel	détachement actionnel et situationnel ❹
⑤ ancrage référentiel dans la situation	détachement référentiel de la situation ❺
⑥ coprésence spatio-temporelle	séparation spatio-temporelle ❻
⑦ coopération communicative intense	coopération communicative minime ❼
⑧ dialogue	monologue ❽
⑨ communication spontanée	communication préparée ❾
⑩ liberté thématique	fixation thématique ❿
etc.	etc.

FIGURE 3.2 – Paramètres pour caractériser le comportement communicatif des interlocuteurs par rapport aux déterminants situationnels et contextuels (Koch et Oesterreicher, 2001, p.586)

En utilisant ces différents paramètres, on peut ainsi caractériser le profil "distanciel' de n'importe quelle situation (voir l'exemple de l'entretien professionnel (ibid, p.

587). En reprenant à leur compte la tripartition cosérienne (Coseriu, 1958, p.25-28) entre langue universelle (la faculté humaine), langue historique (les langues qui se sont formées et transmises dans le temps) et langue individuelle (les énoncés produits par un individu dans une langue historique fonctionnelle), les auteurs proposent une catégorisation complémentaire des différentes situations de communication :

- au niveau universel, il existe une différence marquée entre le médium oral et le médium écrit, car les conditions de l'acte de parole ne sont pas identiques ;
- au niveau des langues historiques, il faut distinguer deux aspects : d'une part, l'existence, dans toute langue historique de *traditions discursives* qui imposent évidemment leurs règles propres mais qui ont également des profils distanciels spécifiques (lettre commerciale, oraison funèbre, mode d'emploi, sonnet, comédie, etc.), d'autre part, il faut distinguer les variétés écrites et orales des langues historiques, puisqu'il s'agit toujours de deux variétés distinctes dans les langues historiques ;
- au niveau individuel, il s'agit de l'acte de parole actuel, qui se place dans le cadre des traditions discursives plus ou moins formées et les soumets à variation individuelle.

La notion de tradition discursive est particulièrement importante, car elle permet d'identifier, au niveau des situations de communications, des pôles qui doivent être décrits de manière plus complète que par la seule dimension de la proximité. Koch et Oesterreicher voient dans ces traditions discursives, notamment celles liées à la distance, l'un des signes de formation d'une langue historique. De même, ils voient bien la variabilité interne de la plupart des traditions discursives, en continuum avec les situations de communication moins normées. (ibid, section 3, p. 601-604). Nous reviendrons sur ces éléments dans le chapitre suivant sur les situations de communication.

Les auteurs analysent ensuite les relations entre écrit/oral, immédiat/distance et innovation : l'innovation survient plus facilement en situation d'immédiateté, puisque l'innovation est favorisée par l'expressivité individuelle, ce qui explique la plus grande propension de l'oral à la créativité. Mais le modèle proposé permet également d'expliquer l'émergence d'innovations au sein de l'écrit, qui là encore interviendra plutôt dans des situations où la distance communicative est faible.

Enfin, Koch et Oesterreicher tentent de placer l'axe proximité-distance dans le modèle cosérien des dimensions de la variation. Ils proposent alors d'ajouter aux trois dimensions diatopique, diastratique et diaphasique, la nouvelle dimension.

Ils considèrent que la dimension immédiat-distance détermine l'espace variationnel tout entier, quelle que soit la dimension considérée : par exemple, la distinction *automobile / bagnole / chiotte* est généralement associée à la diaphasie, mais si l'on ajoute les pôles oral/écrit et immédiat/distance, on aboutit à une analyse différentes : dans le code écrit, *automobile* sera considéré comme "soutenu" *bagnole* "populaire" et *chiotte* "vulgaire" ; dans le code oral, par contre, *automobile* sera considéré comme "recherché" *bagnole* "familier" et *chiotte* "populaire". Les axes oral/écrit et proximité/distance sont donc universels. Contrairement aux autres dimensions, ils sont toujours présents, quelle que soit la situation de communication, puisqu'ils ressortissent aux conditions univer-

selles d'utilisation de la langue dans les paroles.

3.3.2 Discussion : des nouveaux médias

Dans le schéma proposé par (Koch et Oesterreicher, 2001), il n'existe que deux médias, le code phonique et le code graphique. Mais on peut imaginer placer les médiums apparus dans l'ère moderne : téléphone, radio, télévision puis média numériques. Selon Koch et Oesterreicher, ces différents média peuvent tous être rattachés à l'un des codes, écrit ou oral.

On trouve dans (Krefeld, 2015) une critique de ce positionnement. Il insiste d'abord sur le primat de la phonie : « il est impossible, d'un point de vue anthropologique, de mettre en doute le primat de la phonie : la communication linguistique élémentaire et naturelle est produite par un système articulatoire et perçue par les modalités sensorielles, l'audition avant tout, mais avec un support de la vision non négligeable »(ibid, p.265). De ce point de vue, la communication orale se déroule sans support technique extérieur, ce qui la distingue essentiellement des autres modalités, elle est immédiate, et donc non médiatisée. Elle se distingue par trois critères : son ancrage actionnel et situationel, son ancrage référentiel dans la situation, la coprésence spatio-temporelle.

Cet immédiat communicatif, non médiatisé, s'oppose aux deux grands types de médiatisation développés dans l'histoire, les médias graphique et tactile (code braille). La communication orale est intrinsèquement le mode de la proximité, tandis que les communications médiatisées sont toujours dans une situation, même minimale, de distance. À l'appui de son argumentation, Krefeld insiste également sur les transformations de la langue impliquées par la médiatisation de la communication : développement d'une langue plus élaborée au niveau du vocabulaire et de la syntaxe (par exemple les subordonnées complexes sont propres à l'écrit), et émergence d'une variété standard à tous les points de vue : phonologiques, morphologiques, syntaxiques et lexicaux. Avec l'émergence du téléphone, on peut considérer qu'il s'agit d'une communication immédiate, mais sans la coprésence (qui enlève la référence situationnelle partagée). Radio et télévision, par contre, restent apparentés à l'écrit, puisqu'il s'agit toujours d'une communication monologique, comme l'écrit traditionnel, mais se rapprochent de la communication verbale, par l'entremise de l'immédiateté (au moins ressentie) de la voix (radio) ou de l'audiovisuel (télévision). Avec les médias numériques, nous nous rapprochons encore plus de l'immédiat communicatif, qu'il s'agit des courriels, du chat ou plus encore de la visioconférence. Les différents média sont des inventions humaines modifiant les possibilités de communication. En faisant de la science-fiction, on pourrait imaginer un nouveau médium construit à partir de la visioconférence qui ajouterait la possibilité d'une coprésence spatiale complète (et donc permettant des interactions tactiles et le partage de l'espace référentiel commun. Mieux encore, l'ubiquité permettrait à la fois de reproduire à l'identique les conditions de la communication orale, tout en ayant les caractéristiques essentielles de la communication écrite, la distance communicative et la reproductibilité (sauf pour ce qui concerne l'inscription spatio-temporelle de l'acte de communication, qui est propre à l'oral). Nous donnons dans le tableau 3.1 quelques pistes d'analyse en explicitant pour chaque type de communication quelques caractéristiques. Celles-ci

montrent bien que d'une part, la communication orale est bien distincte des communications médiatisées, et d'autre part, que les inventions de communication tendent à se rapprocher de la communication orale.

Type	Canaux sensoriels	Co-présence temporelle des interlocuteurs	Co-présence spatiale des interlocuteurs	Contexte référentiel commun	Implications
Oral	auditif, visuel, autres sens	Oui	Oui	Oui	interactivité immédiate possible de la communication, non reproductibilité
Ecrit	visuel	Non	Non	Non	reproductibilité du message, impossibilité de l'interactivité immédiate
Tactile	tactile	Non	Non	Non	reproductibilité du message, impossibilité de l'interactivité immédiate
Radio	auditif	Oui	Non	Non	reproductibilité du message, impossibilité de l'interactivité immédiate
Télévision	auditif, visuel	Oui pour les émissions en direct, non pour les émissions en différé	Partiellement pour les émissions en direct	Partiellement pour les émissions en direct	reproductibilité du message, impossibilité de l'interactivité immédiate
Chat	visuel	Oui	Non	Non	reproductibilité du message, impossibilité de l'interactivité immédiate
Courriel	visuel	Non	Non	Non	reproductibilité du message, interactivité décalée
Téléphone et audioconférence	auditif	Oui	Non	Non	reproductibilité du message, interactivité immédiate possible
Visioconférence	auditif, visuel	Oui	Partiellement	Partiellement	reproductibilité du message,interactivité immédiate possible

TABLE 3.1 – Caractéristiques des types de communication

3.3.3 Discussion : des paramètres caractérisant la distance communicative

Pour caractériser la distance communicative Koch et Oesterreicher établissent une liste de 10 axes qui permettraient de dresser le profil distanciel de toute situation de communication. Ils n'affirment pas que cette liste soit close. Mais arrêtons-nous plutôt aux critères qu'ils proposent. En effet, il est sans doute possible de rationaliser cette liste, en les ramenant à quelques conditions nécessaires de la communication.

On peut d'ors et déjà reprendre l'argumentaire de (Krefeld, 2015), qui définit la communication verbale par trois des paramètres :(4) ancrage actionnel et situationnel, (5) ancrage référentiel dans la situation, (6) coprésence spatio-temporelle. On peut aller plus loin et replacer les paramètres dans un schéma de communication. Au lieu de prendre pour modèle le schéma proposé par Jakobson, qui ne tient pas compte des paramètres situationnels de la communication, nous proposons d'utiliser celui, plus complet, de Hymes (Hymes, 1982; Hymes, 1984).

Le modèle SPEAKING se compose de huit paramètres (chacun correspondant à une lettre de son libellé) :

— la situation (*Setting*), qui englobe à la fois le « cadre » (le moment et le lieu d'un échange) et la « scène » (sa définition culturelle : « une scène de séduction », « un repas d'affaires »...) ;
— les participants (*Participants*), qui comportent, outre le destinataire et le destinateur, tous ceux qui assistent à la rencontre et qui, par leur présence, influent sur son déroulement ;
— les finalités (*Ends*), qui désignent à la fois les «objectifs-intentions » (l'effet que l'on vise par la communication) et les «objectifs-résultats » (ce qui a effectivement lieu) ; il conviendrait sans doute de rapprocher ces objectifs de la bipartition valeur illocutoire et effet perlocutoire (Austin, 1970) ou encore des études sur l'intentionnalité dans les actes de langage ;
— les actes (*Acts sequences*), qui comprennent à la fois le contenu du message et sa forme ;
— le ton (*Keys*), qui rend compte « de l'accent, de la manière ou de l'esprit dans lequel l'acte est accompli » ; élément important dans la mesure où des actes identiques, dans un même cadre, peuvent différer par le ton et, donc, avoir un effet divergent (c'est le cas du ton ironique qui transforme une insulte en plaisanterie) ;
— les instruments (*Instrumentalities*), qui regroupent à la fois les « canaux » et les « formes » de la parole (un canal linguistique peut ainsi être utilisé pour parler, chanter, psalmodier, mais aussi pour se servir d'un code compris par toutes les personnes présentes - la langue du pays - ou bien d'un dialecte connu d'un seul, ou encore d'une expression n'ayant de sens que pour un intime, etc.) ;
— les normes (*Norms*), qui comprennent à la fois les normes d'interaction (voir les maximes conversationnelles et la théorie de la pertinence) et les normes d'interprétation qui font référence aux habitudes culturelles (« Comment allez-vous ? » n'est pas une incitation à parler de sa santé, mais une phrase rituelle d'ouverture de la communication qui ne requiert que la réponse rituelle complémentaire : « Très bien, merci ») ;
— le genre (*Gender*), qui s'applique à la catégorie formelle dans laquelle s'inscrit un message (poème, conférence, lettre commerciale...).

Ce modèle plus complexe mais tenant compte notamment du cadre de la situation, des objectifs communicatifs et des moyens linguistiques pour les réaliser, d'une triparition des moyens, et des normes, nous paraît l'une des pistes de travail pour compléter l'analyse proposée par Koch et Oesterreicher, qui déterminent un axe majeur d'une quatrième dimension de la variation linguistique, liée aux situations de communications elles-mêmes.

3.4 Flux de communication au sein du réseau social

(Milroy, 1980; Milroy et Milroy, 1985; Milroy et Llamas, 2013; Milroy et Milroy, 1992)

L'approche de la variation par les flux d'informations linguistiques entre les individus et des réseaux sociaux qu'ils construisent s'est développée à partir d'une étude (Milroy et Milroy, 1978) sur les flux de communications entre locuteurs à Belfast, et le

champ sera théoriquement défini dans (Milroy, 1980; Milroy et Milroy, 1985; Milroy et Milroy, 1992). L'étude des réseaux sociaux est une manière d'approcher la dynamique des interactions sociolinguistiques entre les locuteurs qui complète les études cherchant à identifier des groupes sociaux par leurs caractéristiques. Cette approche est peut-être plus adaptée pour suivre le cycle de vie des innovations lexicales, puisqu'elle permet théoriquement de tracer les situations de communications qu'elles traversent et de débusquer les innovateurs, les diffuseurs et les adopteurs en analysant les réseaux construits par les interactions linguistiques entre individus. Elle rend sans doute également mieux compte de la variabilité et de la dynamicité des groupes de locuteurs que les regroupements a priori d'individus dans des classes sociales. Cette approche est également en relation avec l'approche des communautés de pratique, l'identification des réseaux denses de communication étant évidemment un signe de leur existence. Historiquement, le champ s'est développé en réaction aux macro-études cherchant à corréler les variations et les changements linguistiques à des "classes sociales". Dans la suite de cette présentation, nous nous nous appuierons principalement sur les travaux fondateurs du champ (Milroy, 1980; Milroy et Milroy, 1985; Milroy et Llamas, 2013; Milroy et Milroy, 1992).

3.4.1 Réseaux et flux d'informations

L'hypothèse principale de cette approche est que les individus créent un réseau personnel qui leur fournit un cadre pour résoudre l'ensemble des problèmes de la vie quotidienne [7].

Ce réseau est constitué de liens de différentes natures : on distingue les liens forts (parents et amis) - déterminés par une grande fréquence d'interactions et par l'implication de tous les domaines de la vie des individus -, et les liens faibles (autres connaissances). (Milardo, 1988, p.26-36) propose une autre bipartition entre liens d'échanges (généralement parents et amis, avec des interactions fréquentes et un support moral, affectif et matériel) et liens interactifs (qui n'engagent pas de support moral, affectif et matériel, par exemple une relation commerçant-client). (Li, 1994) identifie également un lien "passif" qui permet de rendre compte de relations plus intermittentes mais impliquant un investissement personnel (les parents et amis éloignés, pour les migrants et les individus mobiles).

On distingue dans le réseau les relations de premier ordre, et les relations d'ordre supérieur, qui permettent de constituer le réseau complet des réseaux entre tous les individus. Notons que la théorie des petits mondes estime à un ordre 6 la mise en réseau de tous les individus. Le réseau d'un individu peut être dense ou moins dense, selon sa structure : un réseau dense est défini par le nombre de connexions d'un individu et le nombre de connexions entre les individus avec lesquels il est connecté. On considère généralement qu'un réseau dense constitué de liens forts a la capacité de supporter ses membres aux niveaux pratique et symbolique : il sera par conséquence plus apte à maintenir l'existence de particularités linguistiques et à résister aux pressions d'autres réseaux.

7. « a fundamental postulate of network analysis is that individuals create personal communities that provide them with a meaningful framework for solving the problems of their day-to-day existence »(Milroy et Milroy, 1992, p.2)

La situation est évidemment généralement plus complexe, puisque chaque membre peut également avoir des relations hors du réseau dense. Lorsqu'à l'inverse un individu a un réseau constitué majoritairement de liens faibles, il sera un pont entre les réseaux denses (Granovetter, 1973).

3.4.2 Modèles de réseaux sociaux pour le changement linguistique

L'application des réseaux sociaux a été très tôt appliquée en sociolinguistique, puisque (Gauchat, 1905) en rendait déjà compte dans son étude des variations à Charney (Suisse) et Durkheim évoquait également l'importance du réseau social des individus dans la construction des groupes sociaux. Mais on associe généralement la naissance du champ aux travaux de (Granovetter, 1973) qui seront ensuite appliqués en sociolinguistique dans (Milroy et Milroy, 1978) : cette étude a montré le rôle des réseaux sociaux denses à liens forts dans le maintien de six variables phonologiques spécifiques à la communauté urbaine de Belfast. Elle sera suivie de très nombreuses autres, focalisant sur des réseaux denses ou moins denses (voir (Milroy et Llamas, 2013)).

Le modèle dit du lien faible (Granovetter, 1973), a également été appliqué dans ces études (Milroy et Milroy, 1985). Ce modèle prédit que les innovateurs-diffuseurs sont généralement des individus ayant des liens faibles, parce qu'il servent de pont entre des réseaux plus denses. En effet, la diffusion peut difficilement se produire dans un réseau dense à liens forts, car la communauté ainsi construite a une tendance naturelle à préserver ses habitudes, d'autant plus qu'elle dispose de peu de liens faibles lui donnant accès aux autres réseaux et aux idées et spécificités qui peuvent y circuler. Au contraire, les individus sans réseau dense ou en marge des réseaux denses sont plus enclins à l'innovation, d'abord parce qu'ils ont accès à plus d'informations nouvelles en provenance des différents réseaux, et ils sont eux-mêmes les plus à même de diffuser les innovations. Plusieurs études (Milroy et Milroy, 1978; Lippi-Green, 1989; Trudgill *et al.*, 2000) semblent montrer le rôle des individus à liens faibles dans la diffusion des nouveautés.

(Labov, 2001, p.363-365), dans son étude sur les changements linguistiques à Philadelphie, par contre, considère que les diffuseurs ne sont pas tant les individus à liens faibles que les individus ayant à la fois un réseau social dense à liens forts et un réseau dense à liens faibles, et ils jouent le double rôle d'innovateurs et de diffuseurs : en effet, dans cette position, ils sont capables d'accéder aux nouveautés en provenance d'autres réseaux, de diffuser ces innovations dans leur propre communauté, et de diffuser les innovations provenant de leur communauté (ou d'eux-mêmes) à d'autres réseaux denses (dans l'étude de Labov, ce rôle était tenu par des femmes appartenant aux classes supérieures).

On obtient en simulation informatique des résultats contradictoires sur le rôle de la structure des réseaux sociaux sur les changements linguistiques : (Fagyal *et al.*, 2010), reprenant les travaux des Milroy, considèrent que deux rôles sont essentiels pour la diffusion des innovations : les "loners", n'ayant que peu de liens, généralement faibles, et les "hubs", qui ont des réseaux denses à liens forts. les deux rôles seraient essentiels, à la fois pour diffuser d'un réseau dense à l'autre, puis pour faire adopter les innovations au sein des réseaux denses. Ils montrent que cette situation est la plus favorable

pour la diffusion des innovations. Cependant, leur simulation n'est pas réaliste, puisqu'il suffit qu'un changement parvienne à un individu pour qu'il l'adopte. (Kirby et Sonderegger, 2013; Clem, 2016) ont pour leur part testé différentes configurations de réseaux, et démontrent que, quelle que soit la structure du réseau, le succès d'une innovation ne dépend que de la manière dont on conçoit l'adoption d'une innovation pour un individu donné [8], le réseau comprenant à la fois des "loners" et des "hubs" n'affectant alors que la rapidité du changement.

Comme on le voit, les travaux initiés par les Milroy sur les réseaux sociaux ne donnent pas encore de résultats véritablement convaincants. Les travaux initiaux ont focalisé sur des réseaux locaux, avec des méthodes de mesure approximatives [9]. Il est vrai qu'il est difficilement imaginable de pouvoir mesurer et modéliser le réseau social des individus de manière automatique et purement quantitative, car il n'est pas imaginable de "tracer" les individus dans l'ensemble de leurs interactions verbales et socio-culturelles. De même, les études plus massives qui ont été menées jusqu'à présent, principalement sur les réseaux sociaux numériques, ne sont au mieux que des approximations de ces réseaux eux-mêmes, plutôt qu'une approximation des réseaux sociaux réels. Il semble cependant se dégager deux situations idéales pour la diffusion des innovations : la première comprend un taux élevé d'individus à réseaux denses et liens forts, et un taux élevé d'individus à liens faibles ; la seconde un nombre conséquent d'individus ayant à la fois un réseau dense à liens forts, et un réseau dense à liens faibles. Il s'agit dans ce dernier cas de la définition exacte de l'influenceur idéal que débusquent aujourd'hui les sociétés commerciales dans le cadre de l'exploitation des réseaux sociaux numériques. Cependant, il ne semble pas que le réseau en lui-même soit le principal déterminant des changements linguistiques (ou des habitudes de consommation).

Il reste des pistes d'amélioration pour pleinement exploiter la notion de réseau social. Les études, jusqu'à présent, se sont focalisées sur les réseaux sociaux des individus entre eux, et ne tiennent pas compte d'autres informations en provenance des différents média. Pourtant, il est clair que les moyens de communication actuels (radio, télévision, internet sous toutes ses formes) sont également partie prenante des réseaux sociaux, en tout cas pour ce qui concerne une étude des innovations : il conviendrait d'ajouter ces éléments pour véritablement obtenir une vision des cheminements des innovations dans les réseaux sociaux.

On pourra consulter (Kee, 2017) pour une présentation générale des approches de la diffusion/adoption en Sciences Humaines et Sociales.

8. L'étude (Clem, 2016) teste plusieurs modèles : modèle de l'adaptation-convergence à l'interlocuteur (Giles et Powesland, 1975), modèle de la fréquence (Trudgill *et al.*, 2000), modèle de la sélection du plus simple (Andersen, 1988).

9. (Milroy et Milroy, 1978) ont développé une échelle de densité de réseau (*Network Steength Scale*) à partir de cinq indicateurs : appartenance à un réseau dense local (type club de sport ou de loisirs) ; appartenance à un réseau familial habitant au moins dans deux endroits différents ; travail avec au moins deux liens forts ; travail avec au moins deux liens forts du même sexe ; participation à une association locale.

3.5 Conclusion prospective

3.5.1 Résumé

L'étude qui vient d'être menée, au travers des conceptions de différents auteurs, permet de réviser la notion de langue. Contrairement à la vision unitaire proposée par le structuralisme du CLG, il faut reconnaître l'existence de variations et de variétés qui peuvent être modélisées selon trois dimensions :

— une dimension diatopique, qui est sans doute le fondement de la création des langues, sur des bases géographiques ;

— une dimension diastratique, qui permet d'isoler des variations voire des variétés en liaison avec des sous-groupes de la communauté linguistique : de ce point de vue, la situation se révèle complexe car, historiquement, en tout cas dans les sociétés occidentales, on peut identifier d'abord des variétés conditionnées par la stratification par classes sociales, puis, à l'époque moderne et contemporaine des variétés conditionnées par des communautés de pratiques sans doute plus temporaires ; à ces deux situations correspondent deux méthodologies : la première essaie de corréler les variétés à une macro-structure sociale, en se basant sur des propriétés biologiques, ethnico-culturelles et économiques ; la seconde s'intéresse aux réseaux sociaux locaux, en complétant la première approche d'une étude des réseaux de communication entre les individus, aboutissant à identifier des réseaux denses et moins denses, et identifier des rôles spécifiques permettant de décrire l'émergence et la diffusion des innovations linguistiques ;

— une dimension diaphasique, qui permet de rendre compte de "styles", ou variations individuelles, liées aux interactions des individus dans des situations de communication diverses. Là encore, la situation est complexe, et différentes approches ont été proposées : une approche pour laquelle le style est déterminé par une adaptation à la classe sociale de l'interlocuteur, qui peut être appliqué dans le cadre de relations sociales conflictuelles où certaines variétés sont stigmatisées et d'autres valorisées, et où chacun des individus est le représentant d'une classe sociale et d'un vernaculaire spécifique ; une second approche considère que le style est le résultat d'une adaptation complexe à l'auditoire - cet auditoire étant multiple (les auditeurs, l'auditoire secondaire et l'auditoire imaginé) - et est dépendant des objectifs de communication ; enfin, une dernière approche considère le style comme une expression volontaire déterminée par des objectifs communicatifs.

Dans l'approche sociolinguistique, l'innovation linguistique se définit par l'émergence d'une variante linguistique. Le processus d'émergence aboutit en synchronie à une variation, qui se résoudra en diachronie par l'adoption d'une des deux variantes ou la recomposition du champ sémantique. Les sociolinguistiques ont proposé différents modèles, permettant de définir le mécanisme de sélection et d'adoption des variantes. Labov distingue le changement par en-dessous et le changement par au-dessus : celui par en-dessus concerne l'adoption d'une innovation d'un groupe social considéré comme plus prestigieux et/ou du rejet d'innovations provenant d'un groupe stigmatisé ; il est conscient

et conditionné par une volonté consciente d'adopter les pratiques des groupes sociaux les plus valorisés ; le changement par en-dessous est inconscient et consiste à introduire des marqueurs linguistiques propres à un groupes social par les membres eux-mêmes : il s'agit d'un processus inconscient qui est conditionné par l'identification au groupe social. (Giles et Powesland, 1975) a introduit la notion d'adaptation-convergence à l'interlocuteur/auditoire (textitCommunication Adaptation Theory) (Giles et Powesland, 1975) (et (Gallois et Giles, 2015) pour une revue des aménagements effectués depuis lors) pour expliquer l'adoption de certaines innovations, tandis que (Trudgill *et al.*, 2000) a insisté sur l'importance de la fréquence d'exposition (qui sera plus approfondie par les psycholinguistes). Enfin, (Andersen, 1988) a indiqué qu'aux facteurs précédents pouvait s'ajouter un facteur de "simplification" pour expliquer le choix d'une des variantes. Ces modèles s'intéressent essentiellement aux innovations aboutissant à une variation : du coup, elles ne tiennent pas compte des cas d'innovations linguistiques notamment lexicales liées à des innovations technologiques qui doivent être nommées, et pour lesquelles ne se pose pas le problème de la concurrence. De même, la dimension ludique de l'innovation linguistique n'est pas considérée. Il est vrai qu'elle n'est généralement pas suivie de diffusion.

Au niveau des acteurs de l'innovation, la sociolinguistique a identifié deux structures de réseaux sociaux favorisant l'adoption des innovations : Labov considère que la situation la plus favorable est liée à l'existence d'individus ayant un réseau dense à liens forts et un réseau dense à liens faibles : dans cette configuration, ils sont les diffuseurs par excellence, de par leur prestige au sein de leur propre communauté, et leurs liens nombreux avec d'autres communautés. n autre modèle dit du lien faible, considère que ce sont les individus à liens faibles qui sont les véritables diffuseurs des innovations, car ils sont des ponts entre les communautés, et sont également des innovateurs, de par leur ouverture à des communautés disparates et à leurs idées. Mais ce modèle peut être complété par l'ajout des "hubs", ces individus ayant un réseau dens à liens forts, qui permet la diffusion au sein de chaque communauté. ces différentes modèles ont été testés dans différentes configurations sociales, et, comme on peut s'en douter, les situations sont chaque fois différentes. Il a été démontré que ces configurations idéales ne jouaient pas véritablement un rôle dans l'adoption des innovations, elles permettent seulement d'en accélérer le rythme.

Nous avons également présenté l'approche de Koch et Oesterreicher qui explicitent une quatrième dimension, universelle et transversale aux trois autres, celle de la proximité-distance communicative, qui permet de caractériser toutes les situations de communication. Construite à partir de la distinction des codes écrit et oral, qui lui sert de fondement, elle permet d'associer à chaque situation de communication une valeur de proximité. Nous avons indiqué deux pistes pour préciser ce modèle : tout d'abord, de rétablir le caractère absolument immédiat de l'oral, alors que l'écrit se place dans le continuum proximité-distance ; ensuite, une proposition de rationalisation des critères permettant d'identifier la proximité des situations de communication. Il nous semble à cet égard que le schéma de communication de (Hymes, 1982) est le plus complet pour établir une liste complète de paramètres.

3.5.2 Perspectives

Nous avons également évoqué, au cours de l'exposé, différentes pistes de travail à explorer, notamment l'intérêt d'une étude des flux de communications entre individus à un niveau plus global que les études menées jusqu'à présent, en ajoutant les organes de diffusion d'information qui jouent, dans la période contemporaine, un rôle de diffuseurs non négligeables et font à l'évidence partie des situations de communication auxquelles nous sommes exposés. Cette approche inductive des variations nous semble plus objective que les études sociologiques par l'établissement a priori de caractéristiques individuelles pour déterminer les corrélations variations/changements linguistiques et les groupes d'humains.

Il subsiste d'autres points encore à éclaircir et qui méritent une exploration plus approfondie. Nous revenons ici sur le principal d'entre eux. Les études sur la dimension diaphasique (le *style*) en sociolinguistique aboutissent à l'associer à l'*individu en situation de communication*, et la dimension de la distance communicative est définie par les auteurs eux-mêmes comme « le comportement communicatif des interlocuteurs par rapport aux déterminants situationnels et contextuels »(voir figure 3.2). Il nous semble que pour rationaliser les études en la matière, deux voies sont possibles : soit nous conservons trois dimensions de la variation : la première est liée à l'environnement spatial (la diatopie), la seconde est liée aux groupes sociaux, l'individu pouvant être défini dans ce cadre (la diastratie), le troisième est lié aux situations de communications, plus ou moins normées, que nous proposons d'appeler, en empruntant le terme de Halliday, le diasituationnel ; soit nous conservons quatre dimensions : diatopique, diastratique (pour les variations liées aux groupes sociaux), diaphasique (pour les variations liées à l'individu) et diasituationnel. Mais dans ce cas, il faudrait préciser en quoi consiste la variation individuelle, qui ne semble être qu'un agrégat dans l'individu de variations diastratiques et diasituationnelles.

Coseriu décrit parfaitement comment probablement les langues se sont créées : une proximité géographique pour une communauté qui devient une communauté linguistique. Dès le début de la création d'une langue commune, il est probable que des variations diastratiques au sens large, se soient mises en place ; et plus la communauté s'agrandit, plus il est normal que des variations se développent, ces variations, qu'elles soient individuelles ou collectives, étant des expressions de variétés liées aux caractéristiques sociales de la communauté. Le style, dans ce cadre, n'est qu'une des composantes de la diastratie, une expression individuelle.

Le traitement qui a été fait de la diaphasie par la sociolinguistique nous semble clairement ressortir d'une analyse des situations de communications. De même, l'axe proximité-distance, comme le disent les auteurs eux-mêmes permet de « caractériser le comportement communicatif des interlocuteurs par rapport aux déterminants situationnels et contextuels ». Il s'agit donc bien de variations qui dépendent de la communication elle-même, et ils ont ainsi dans la dimension diasituationnelle, clairement identifié un axe universel. Mais qui nécessite d'aller encore plus loin et de tenter de décrire les situations de communication, qu'il s'agisse de traditions discursives ou de situations de communication moins normées. Ce sera l'objet d'un prochain travail.

Deuxième partie

Modèles opérationnels

Résumé

Dans cette partie nous abordons l'innovation lexicale du point de vue de l'automatisation des tâches de détection et de suivi du cycle de vie des lexies en corpus dynamique.
Dans le **chapitre 4**, nous présentons une architecture générale pour la détection, la description linguistique et le suivi de l'évolution des lexies sur corpus dynamique. Il s'agit de proposer un modèle général pour permettre l'automatisation des différentes tâches, tout en proposant une architecture combinant les processus automatiques et l'expertise linguistique. Après la présentation d'un programme de travail, nous présentons le travail effectué dans le cadre du projet de recherche *Néoveille*.
Dans le **chapitre 5**, nous focalisons sur la tâche de détection automatique des néologismes de forme, c'est-à-dire des néologismes qui se manifestent par l'émergence d'une nouvelle forme lexicale. Nous présentons une modélisation du phénomène, puis les différents algorithmes pour effectuer une détection automatique, enfin le travail effectué dans le cadre du projet *Néoveille*.

Chapitre 4

Plateforme pour l'étude des néologismes en corpus

Sommaire

Le modèle présenté dans les chapitres précédents doit être mis à l'épreuve de la réalité linguistique. Pour ce faire, les outils de traitement automatique des langues (TAL) peuvent être mis à profit pour automatiser certains processus impliqués dans le changement lexical. Dans ce chapitre, nous présentons un modèle global, puis les algorithmes et les réalisations développées dans le projet Néoveille pour détecter, décrire linguistiquement puis suivre le cycle de vie des lexies en corpus dynamique.

En accord avec la modélisation des langues présentée précédemment, 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 discours d'un point de vue socio-pragmatique ;
— 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 fréquentielle, linguistiques et/ou socio-pragmatiques des formes lexicales existantes dans les discours.

Enfin, étant donné qu'il s'agit pour nous, certes d'automatiser les différents 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 devra prévoir cette interaction entre les processus automatiques et les interventions humaines.

Nous détaillons ci-après, après quelques éléments méthodologiques et une présentation des travaux précédents sur le sujet, les choix que nous avons effectués dans le cadre de la plateforme Néoveille. Les modules spécifiques concernant la détection de la néologie formelle et de l'évolution des lexies seront présentés dans les deux chapitres suivants.

4.1 Éléments méthodologiques

Dans cette section, nous explicitons les hypothèses méthodologiques qui ont guidé les approches que nous avons développées en traitement automatique des langues.

4.1.1 Distributionnalisme et induction

Le modèle théorique que nous avons explicité dans la première partie fait du discours le centre des préoccupations linguistiques. Ce modèle considère que, psychologiquement, ce sont les informations glanées au cours de nos expériences antérieures, notamment l'ensemble des interactions discursives auxquelles nous avons été confrontées, qui nous permettent d'interpréter un nouveau message et spécifiquement d'y reconnaître une innovation lexicale. Cela trace pour nous le périmètre des travaux en TAL, qui doivent baser leurs algorithmes uniquement sur les discours qui apparaissent, en extraire des informations, les stocker puis les réutiliser pour traiter les discours nouveaux qui apparaissent.

Ce recentrage sur le discours trouve sa source dans les travaux du distributionnalisme harrissien, ensuite poursuivi par les travaux de (Firth, 1957), la linguistique de corpus et tous les travaux en TAL basés sur les statistiques et les probabilités, jusqu'aux derniers développements de l'apprentissage automatique et profond. Il n'est pas inutile de revenir à la méthode préconisée par Harris.

Le distributionnalisme harrissien, pour étudier les langues, établit une méthode ne prenant appui que sur les matérialisations écrites ou orales observables, et refuse de

recourir à des hypothèses sur la pensée sous-jacente ou sur la relation de référence qui lie le langage au monde. C'est le programme décrit dans l'un des articles fondateurs :

> Nous allons voir ici que, d'une part, on peut décrire toute langue par une structure distributionnelle, c'est-à-dire par l'occurrence des parties (et, en dernière analyse, des sons), relativement les unes aux autres et que, d'autre part, cette description n'exige pas qu'on fasse appel à d'autres caractéristiques, telles que l'histoire ou le sens. (Harris, 1954, trad. française, 1970, p.14)

Cette approche [1] se justifie d'une part par les difficultés ou même peut-être l'impossibilité d'accéder à la « pensée », d'une part, mais est aussi fondée sur le creuset d'informations contenues dans les réalisations matérielles du langage :

> « les parties d'une langue ne se rencontrent pas de façon arbitraire les unes par rapport aux autres ; chaque élément se rencontre dans certaines positions par rapport à certains autres éléments. »(Harris, 1954, trad. française, 1970, p.15)

Cette première constatation est fondamentale : la *distribution* des éléments du langage nous informe à la fois sur les unités linguistiques (les « classes distributionnelles ») et sur leurs caractéristiques :

> « Les contraintes concernant l'occurrence relative de chaque élément peuvent être décrites très simplement par un réseau d'interrelations, certaines de ces dernières étant formulées à partir des résultats de certaines autres, plutôt que par une simple évaluation de toutes les contraintes imposées à chaque élément considéré séparément. »(Harris, 1954, trad. française, 1970, p.16)

L'idée énoncée ici est celle de la possibilité d'une re-construction progressive des classes d'équivalence (ou classes de similarité), permettant d'identifier et de décrire les unités linguistiques à tous les niveaux. Ces principes fondent le développement des méthodes statistiques et probabilistes dans le champ de la linguistique, tout d'abord dans le giron de la linguistique de corpus, à partir des premiers travaux de Sinclair, puis tous les travaux en TAL aboutissant aux modèles de langues utilisés en apprentissage automatique et en apprentissage profond. Nous détaillerons certaines de ces applications dans le chapitre sur le suivi des évolutions lexicales, avec la représentation par *word embeddings* qui appliquent ce principe distributionnel pour accéder au sens des lexies.

Une seconde conséquence de cette approche concerne la méthode pour parvenir à identifier les classes d'équivalence, qui ne sont pas définies a priori, mais induites des discours : pour ce faire, une méthode basée sur le comptage des occurrences dans un corpus doit être adoptée, puis des algorithmes pour la découverte des classes d'équivalence, c'est-à-dire des unités qui se comportent de façon similaire. Cette approche semble tout à fait éloignée des préoccupations linguistiques puisque la découverte des unités et de leurs propriétés ne doit se baser sur aucune connaissance antérieure, et spécifiquement

1. Historiquement, cette approche se justifie aussi, d'une part, par les difficultés qu'ont rencontrés les linguistes américains pour décrire les langues amérindiennes dans le cadre de leur mission au sein du *Census Bureau*, et, d'autre part, par leur adoption de la théorie gestaltiste.

sur aucun a priori linguistique. Mais il s'agit cependant du principe de base de toutes les linguistiques basées sur l'usage. Plus, les travaux de la linguistique cognitive ont montré la centralité des *répétitions* dans les processus d'apprentissage et de stabilisation mémorielle (ou non) des paires formes-sens.

Nous reviendrons plus en détail sur les différentes méthodes de comptage disponibles et utilisables en TAL dans le chapitre 7.

Une troisième conséquence de cette approche est que la langue - ce que nous mémorisons - ne peut pas être formulée en termes discrets, mais en termes continus, puisque dans cette approche les phénomènes sont mémorisés au fur et à mesure de leur apparition, qui aboutissent donc à des représentations continues. Prenons un exemple : notre connaissance de la forme *maison* dépendra de toutes les expositions contenant ce mot auxquelles nous avons été confronté : les discours et les contextes de ces discours. De ces événements nous tirerons une représentation probable de l'association forme-sens, qui pourra évoluer dans le temps au fil des expositions. Nous verrons que la loi des grands nombres fait converger la représentation continue vers un ou des prototypes pour chacune des unités. Nous renvoyons à (Cartier, 2016a) pour une présentation plus détaillée des hypothèses et du modèle harrissien.

4.1.2 Théorie de l'information et entropie

L'approche distributionnelle, prise dans le mouvement structuraliste favorisant l'étude synchronique, ne verra pas l'intérêt d'utiliser l'historique des discours pour interpréter les discours nouveaux. On trouve encore aujourd'hui les traces de cet apriori puisque la très grande majorité des travaux en TAL cherchent exclusivement à établir des modèles de langue synchroniques. Pourtant, à l'évidence, l'interprétation d'un discours nouveau se base sur le traitement des discours antérieurs auxquels nous avons été confronté, et le jugement de nouveauté provient d'une comparaison des propriétés d'un discours nouveau avec les propriétés mémorisées de l'ensemble des discours antérieurs. On trouve l'application de ce principe dans la théorie mathématique de l'information proposée par (Shannon, 1948).

Pour Shannon, on peut partir de l'hypothèse que l'information est a priori aléatoire : si nous partons d'un vocabulaire composé de n symboles, et sans aucune autre information, toutes les combinaisons de ces symboles peuvent se produire de manière équiprobable. Chaque symbole i a une probabilité calculée comme sa fréquence constatée divisée par la fréquence de tous les symboles du corpus $p_i = 1/n$. Un message X composé de n symboles a une probabilité P qui est la somme des probabilités de chacun des symboles : $P(X) = \sum_{i=1}^{n} p_i$. Pour chaque symbole i, si nous pouvons calculer sa probabilité, on peut calculer la symétrique de cette probabilité, à savoir son incertitude (ou sa nouveauté) : $-log_2(p)$. Or, plus une information est incertaine, plus elle est intéressante et contient des informations nouvelles, et à l'inverse plus elle est certaine, et moins elle contient d'informations nouvelles. Shannon proposera d'appeler cette mesure de surprise ou de nouveauté l'*entropie* et en formulera une équation mathématique. Soit un message X composé de n symboles, un symbole i ayant une probabilité p_i d'apparaître, l'entropie H sera définie comme :

$$H(X) = -\sum_{i}^{n}(p_i)\log(p_i) \tag{4.1}$$

Cette mesure, qui sert de base dans nombre d'algorithmes d'apprentissage automatique, permet de calculer la surprise d'un événement nouveau sur la base de la probabilité calculée sur la base des événements passés. Elle est donc essentielle pour la détection de la nouveauté dans les messages linguistiques.

4.1.3 Automatisation et collaboration homme-machine

En dehors des éléments précédents, qui fondent une approche quantitative et probabiliste de la linguistique, nous sommes partis d'un autre postulat qui ressortit à l'actuelle incapacité des traitements automatiques à effectuer des traitements suffisamment efficaces pour être analysés tel quel. Il n'y a pas aujourd'hui de programmes de TAL qui effectuent un traitement avec une précision et un rappel de 100%. Peut-être ne sera-ce d'ailleurs jamais le cas, car les messages linguistiques contiennent peut-être intrinsèquement de l'ambiguïté et une entropie jamais nulle. Partant de ce principe, appuyé également sur les tâches que nous proposons d'effectuer, nous avons considéré que la plateforme à développer devait prévoir une collaboration homme-machine, permettant d'une part la correction humaine des traitements automatiques (notamment concernant la détection automatique des néologismes), et d'autre part l'ajout d'informations linguistiques (sur les lexies) et métalinguistiques (sur les corpus) non inférables des données textuelles traitées. Le processus de collaboration ne va pas que dans le sens machine-homme, mais également dans le sens homme-machine. Le système doit prévoir l'exploitation des données provenant de l'expertise manuelle dans les processus automatiques ultérieurs, créant ainsi une chaîne d'apprentissage actif (*Active Learning*[2]).

4.2 Outils disponibles pour l'étude des néologismes sur corpus

Comme indiqué dans l'introduction, il s'agissait dans le projet Néoveille de construire un système permettant de détecter (semi-) automatiquement, de décrire linguistiquement et de suivre le cycle de vie des lexies sur corpus dynamique contemporain.

Une étude préalable des systèmes existants nous a permis, en amont des développements, de préciser l'architecture générale de la plateforme visée et les différents modules nécessaires. Nous présentons tout d'abord les outils disponibles pour suivre les néologismes. À notre connaissance, il n'existe actuellement aucune plateforme permettant d'effectuer à la fois le repérage automatique (ou semi-automatique) des néologismes en corpus et de suivre l'évolution des néologismes en corpus diachronique. Il existe cependant des outils génériques permettant d'étudier les néologismes, ainsi que des systèmes

2. Pour plus d'informations sur ce domaine, nous renvoyons à (Settles, 2014; Olsson, 2009).

dédiés.

4.2.1 Outils génériques pour la recherche et le suivi des néologismes

Parmi les systèmes génériques, il faut citer trois types d'outils régulièrement utilisés par les linguistes pour effectuer une veille néologique (ou tout autre travail dit de linguistique de corpus) : les moteurs de recherche généraux ou spécialisés, les outils dédiés à la linguistique de corpus et au moins deux outils proposés par Google, *Google Trends* et *Google Ngrams*.

4.2.1.1 Moteurs de recherche génériques

Les moteurs de recherche généraux (type Google ou Bing) sont des outils pratiques pour effectuer des recherches d'attestations néologiques à partir du moment où l'expert a déjà repéré des candidats : ils sont simples d'utilisation, permettent d'accéder aux corpus numériques les plus conséquents, ces corpus sont dynamiques puisque toute nouvelle information publiée sur Internet apparait très rapidement dans les résultats, ils fournissent des contextes d'apparition avec une indication (approximative) du nombre d'occurrences. Des outils de tri permettent de classer les résultats (par date, par pertinence, par type de documents, etc.).

Ils présentent trois caractéristiques fondamentales pour la plateforme que nous visons :

— Ils sont gérés par une technologie éprouvée depuis près de 30 ans, permettant d'une part d'indexer les documents dans un format compact, et d'autre part d'interroger ces index de manière extrêmement rapide ;
— l'architecture de ces systèmes permet une accumulation dynamique des informations nouvelles publiées sur internet ;
— pour chacun des items d'informations ("les pages web"), des méta-informations sont stockés en plus du texte : date de publication, nom de domaine, voire type de document) ; même si les moteurs de recherche généralistes sont assez limités de ce point de vue, la technologie sous-jacente permet a priori de stocker un nombre illimité de méta-informations.

Ils présentent cependant plusieurs inconvénients :

— ils effectuent une recherche sur du corpus tout-venant, issu du web, sans qu'il soit possible de distinguer les types de documents accessibles : il peut s'agir de pages confidentielles truffées de fautes d'orthographe, de *mèmes*, ce qui introduit un bruit non négligeable empêchant toute conclusion objective sur le nombre d'occurrences et sur les contextes d'apparition ou sur le cycle de vie ; il est cependant possible de restreindre les recherches à un sous-corpus (par exemple, dans Google, de restreindre aux textes en langue française, aux sites appartenant à tel ou tel pays francophone, ou encore aux sites d'actualités ou aux blogs) : ces fonctionnalités additionnelles permettent de s'approcher des moteurs spécialisés, comme Europresse, qui proposent un moteur de recherche à partir de données plus contrôlées de la presse (voir section suivante) ;

— ils donnent des résultats sous forme d'extraits de texte, montrant généralement la seule première attestation de la lexie cherchée dans le texte, alors que le linguiste souhaiterait obtenir l'ensemble des occurrences ;
— les résultats, au-delà de la première ou de la deuxième page selon les cas, ne sont généralement pas fiables ;
— les possibilités d'interrogation des corpus sont limitées : dans Google, il n'existe pas, par exemple, de langage d'interrogation par expressions régulières, ou par chaîne de caractères tronquée (par exemple arriv.* pour retrouver tous les mots contenant la sous-chaîne *arriv*) ;
— les moteurs de recherche généralistes n'effectuent aucun traitement linguistique préalable des corpus, limitant par là les possibilités d'interrogation et la précision des recherches.

Au final, la limitation des moteurs de recherche généralistes doit également être rapportée au public cible, qui, essentiellement, cherche à accéder à des pages web répondant à une requête approximative. Il n'y a donc finalement de recherche linguistique que de manière secondaire. Ces outils - et principalement Google, puisque le nombre de pages indexées est sans commune mesure avec ce que peuvent proposer ses concurrents - permettent cependant d'obtenir une idée approximative du nombre d'occurrences d'une lexie, et, pour les lexies récemment apparues, d'obtenir généralement les contextes web pertinents. De plus, la technologie du moteur de recherche sous-jacent est bien évidemment l'état-de-l'art en la matière, pour stocker des masses considérables de données, ajouter des métainformations, et accéder de manière quasi-instantanée aux résultats de requêtes simples [3].

4.2.1.2 Moteurs de recherche spécialisés

Une partie des limitations présentées ci-dessus sont levées dans les moteurs de recherche spécialisés. Ceux-ci proposent, souvent de façon payante, un moteur de recherche permettant d'accéder à des corpus de texte beaucoup plus contrôlés. Citons *European Media Monitor, Europresse, Factiva* pour ce qui concerne la presse, les deux derniers étant parmi les nombreux aggrégateurs de contenu proposant de tels services.

Par exemple, *Europresse* permet de rechercher (avec des fonctionnalités généralement assez proches des moteurs généralistes) parmi l'ensemble de la presse française et anglaise depuis 1945. Des moteurs de recherche spécialisés existent dans nombre de domaines (médical, juridique, fiscal, brevets, etc.) de façon payante et plus rarement gratuitement (voir par exemple la base de données PubMed sur les publications dans le domaine médical). Citons également les moteurs de recherche proposés par les bibliothèques nationales, qui proposent à la fois un très bon contrôle des sources d'informations, ainsi que des fonctionnalités de recherche avancées.

3. Nous n'entrerons pas dans les détails techniques, mais les requêtes peuvent être beaucoup plus complexes. Pour un exemple de requêtes complexe,s nous renvoyons à la documentation technique du moteur de recherche Open Source *Apache Lucene* et sa surcouche *Apache Solr* : `http://lucene.apache.org/solr/`

Par exemple, le site internet de la Bibliothèque nationale de France, *Gallica* propose en libre accès, un moteur de recherche doté de nombreuses fonctionnalités :

— accès à des corpus de textes variés ;
— recherche tenant compte des types de documents et d'une série de méta-informations (auteur(s), date,mots-clés, domaine, etc.) ;
— recherche dotée de fonctionnalités avancées, généralement liées au moteur de recherche *Apache Lucene* ou ses surcouches *Apache Solr* ou *Elastic Search* ;
— possibilité de sauvegarde des résultats.

Par contre, ces moteurs de recherche sont encore limités, au moins sur trois points :

— ils ne permettent pas d'effectuer des recherches linguistiques (par exemple pour retrouver les adjectifs les plus fréquents associés à *voiture*) ;
— ils sont souvent basés sur une version automatiquement numérisée et océrisée des ouvrages papier, induisant un certain nombre d'erreurs dans les résultats ;
— l'accès au contenu des publications contemporaines est limité voire impossible, du fait des droits d'auteurs.

Les outils spécialisés, notamment les moteurs de bibliothèques publiques, sont donc parmi les outils au plus grand potentiel : d'une part, ils disposent des sources d'informations les plus importantes, au moins concernant les sources papier, et ils ont développé des technologies pour l'indexation des informations de plus en plus sophistiquées : réutilisation des travaux de classement des ouvrages, numérisation, océrisation, indexation dans les moteurs de recherche les plus récents. Reste cependant que ces moteurs de recherche restent généralistes et n'embarquent pas jusqu'à aujourd'hui d'outils spécifiques pour l'analyse linguistique.

Notons cependant des projets pilotes pour embarquer une analyse linguistique et permettre des interactions enrichies avec les résultats : projet Néonaute (Cartier *et al.*, 2018a; Cartier *et al.*, 2018b) (`http://tal.lipn.univ-paris13.fr/neonaute`), projet Européen NewsEye (`https://www.newseye.eu/`). On pourra également consulter les développements effectués dans le cadre de l'*International Internet Preservation Consortium* (IIPC) : `http://netpreserve.org`.

4.2.1.3 Outils pour la linguistique de corpus

Les outils dédiés à la linguistique de corpus ont été développé conjointement avec le mouvement de construction des corpus de référence, qui nécessitaient des outils conviviaux pour être exploités. Aujourd'hui, deux systèmes en Open Source sont disponibles et éprouvés : IMS Corpus Workbench (Evert et Hardie, 2011) et NoSketchEngine (Rychlỳ, 2007). Un autre système, SketchEngine, sur la base du précédent, a été développé dans un cadre commercial (Kilgarriff *et al.*, 2014).

Sans refaire l'historique des outils pour la linguistique de corpus (voir (Anthony, 2013)), il faut tout de même évoquer les principales spécificités technologiques de ces outils :

1. **Stockage de corpus de taille conséquente** : les trois outils précédents ont développé des modules de stockage de corpus étiquetés morphosyntaxiquement

voire syntaxiquement : sans véritablement permettre un ajout convivial de nouvelles données textuelles, ces outils permettent d'interroger, sous forme d'interface Web, les corpus et se rapprochent ainsi des possibilités des moteurs de recherche. Ils incluent également des méta-information sur les documents constituant les corpus ;

2. **Présentation des résultats sous forme de *Key Word In Context (KWIC)*** : les outils développés ont intégré une présentation spécifique des résultats de recherche sous forme d'alignement de la requête textuelle (mot ou locution) sous un format tabulaire permettant de visualiser les contextes de manière conviviale, ainsi que de les trier ;
3. **Statistiques lexicales** : les outils proposent également toute une série de statistiques lexicales pour obtenir le nombre d'occurrences de la séquence recherchée, pour obtenir les collocations du terme recherché (voir par exemple `http://collocations.de/AM/index.html`), ainsi que, pour SketchEngine, une première approximation de la notion de profil combinatoire des lexies (Kilgarriff *et al.*, 2010) ;
4. **Langage de requêtes évolué** : l'apport principal peut-être de ces développements consiste en la mise en oeuvre d'un language de requête permettant d'effectuer une recherche non pas par mots-clés, mais une requête structurée permettant d'obtenir par exemple les occurrences de la forme *marche*, en tant que verbe, suivi de deux mots quelconques au maximum, et d'un nom commun au singulier (voir (Hardie, 2012; Jakubicek *et al.*, 2010)).
5. **Interfaces web conviviale** : les trois systèmes précédents incluent une interface internet permettant une interaction avec les données qui se veut conviviale.

Citons également l'outil *WebCorp* (Renouf *et al.*, 2007) qui permet d'effectuer des recherches sur du corpus de presse dynamique, et propose des visualisations des évolutions d'usage dans le temps [4]. Les principales lacunes de ces outils ressortissent à la difficulté pour indexer les corpus, opération non triviale et la limitation des corpus en taille (mais (Evert et Hardie, 2015) annoncent des développements pour lever ces limitations).

4.2.1.4 Corpus numériques synchroniques et diachroniques

Les outils précédents sont utilisés dans de nombreux centres de recherche et des corpus ont été spécialement développés. Citons, dans le domaine français, le corpus FrWac : `https://corpora.dipintra.it/`, ainsi que des versions du Wikipedia annoté mis en oeuvre par Franck Sajous à l'ERSS-Toulouse (`http://redac.univ-tlse2.fr/corpus/wikipedia.html`). À notre connaissance, il n'existe par contre pas de corpus contemporain diachronique librement accessible [5]. Dans le domaine anglophone, bien plus avancé, les plus gros corpus sont disponibles via le site de Mark Davis (`https:`

4. `http://wse1.webcorp.org.uk/`
5. Il existe cependant un tel corpus récupéré par la BnF, dans le cadre de sa mission d'archivage du web.

//corpus.byu.edu/), avec notamment le corpus NOW, qui récupère depuis 2010 de manière dynamique les articles de presse de plus de 500 journaux anglophones.

4.2.1.5 Google Ngrams et Google Trends

Pour ce qui concerne le suivi des néologismes, Google Trends [6] est aujourd'hui un modèle pour étudier l'évolution fréquentielle des termes au cours du temps. Même si cette application ne donne des résultats que pour les mots issus des requêtes saisies dans Google, elle donne une idée de ce qu'il faudrait implémenter pour rendre compte des évolutions d'attestations dans un moteur de recherche spécialisé en veille néologique.

De même, l'application Google Ngrams [7] (Michel *et al.*, 2010; Aiden et Michel, 2014) propose une visualisation de l'évolution des emplois de lexies sur une période couvrant plusieurs siècles, selon les langues. Les corpus sont issus de la numérisation progressive des archives de la Bibliothèque du Congrès aux États-Unis, et l'interface permet également des requêtes semi-évoluées. Une application similaire a été développée récemment pour le finlandais (Birkenes *et al.*, 2015).

4.2.2 Outils spécifiques pour la recherche et le suivi des néologismes

Pour ce qui concerne les outils spécifiques, citons cinq applications de veille néologique [8] : Pompamo [9] (Ollinger et Valette, 2008), le Logoscope [10] (Gérard *et al.*, 2014), Obneo.Buscaneo [11] (Cabré *et al.*, 2003; Cabré *et al.*, 2004; Cabré et Nazar, 2011), NeoCrawler [12] (Kerremans *et al.*, 2012; Kerremans et Prokić, 2018) et le système (dorénavant SagotEtAl) proposé dans (Sagot *et al.*, 2013). Parmi ces outils, Pompaneo, le Logoscope et SagotEtAl fonctionnent exclusivement sur le français, NéoCrawler pour l'anglais, Obeno/Buscaneo étant principalement développé pour le catalan et l'espagnol, mais une extension aux autres langues romanes étant possible. Parmi les fonctionnalités proposées par ces outils, les plus avancés sont NéoCrawler et Obneo/Buscaneo, qui en plus d'effectuer une veille néologique sur du corpus internet en continu, proposent une interface Internet sécurisée de validation et de description des néologismes détectés. Dans le premier cas, le corpus provient de l'API de Google, dans le second de sites web de presse qui sont récupérés automatiquement à intervalle régulier. L'ensemble des outils utilisent des dictionnaires de référence pour la détection des néologismes formels : Pompamo par exemple, permet, à partir d'un texte étiqueté morphosyntaxiquement de

6. https://trends.google.fr/trends
7. http://books.google.com/ngrams
8. En réalité, de nombreux centres de recherche disposent d'outils de veille néologique, généralement confidentiels. La plupart fonctionnent comme des bases de données lexicales dans lesquelles les opérateurs humains saisissent les néologismes qu'ils détectent par veille manuelle. Par ailleurs, d'autres outils sont décrits dans diverses publications. Sans exhaustivité, citons (Breen, 2010; Costin-Gabriel et Rebedea, 2014; Janssen, 2012b; Lai et Ng, 2014; Halskov et Jarvad, 2010).
9. http://www.cnrtl.fr/outils/pompamo/
10. http://logoscope.unistra.fr/
11. http://obneo.iula.upf.edu/buscaneo/
12. http://www.neocrawler.anglistik.uni-muenchen.de/crawler/html/

repérer des candidats néologismes. Il utilise comme dictionnaire d'exclusion Morfalou [13] ainsi que des dictionnaires additionnels (noms propres et gentilés), éventuellement fournis par l'utilisateur. Cet outil est relativement simple, car il n'effectue aucune catégorisation des candidats néologismes, et ne permet de travailler que sur de petits fichiers à télécharger par Internet, ce qui en limite singulièrement l'intérêt. Le Logoscope, pour sa part, propose directement les candidats néologismes identifiés dans des corpus de presse, avec leur contexte d'apparition. Mais, là encore, l'application est très statique, et semble peu évolutive, l'utilisateur n'accédant qu'au résultat final qui comporte une grande part de travail manuel. Selon les auteurs, les résultats du système sont humainement triés avant présentation au public. Finalement, Obneo/Buscaneo semble l'outil le plus avancé puisque, en plus d'un détecteur automatique de néologismes formels, il propose toute une panoplie d'outils annexes permettant de stocker les néologismes validés dans une base lexicale comprenant des champs descriptifs additionnels, ainsi qu'un gestionnaire de dictionnaires de référence, permettant une adaptation aux corpus.

4.2.3 Plateforme de repérage et de suivi des néologismes : exigences d'un système idéal

Les fonctionnalités des outils précédents, combinés avec le modèle proposé dans les chapitres précédents, nous permettent d'expliciter trois caractéristiques essentielles d'une plateforme de détection et de suivi des évolutions lexicales, qui ressortissent tous d'un objectif de simulation de la compétence de compréhension des messages linguistiques par l'être humain :

- Le système doit articuler dynamiquement discours (ou parole ou corpus) et langue (ou ressources linguistiques mémorisées) dans un flux continu : comme nous l'avons explicité dans le chapitre 1, psychologiquement, nous mémorisons de façon continue le flux des événements linguistiques, et la plus ou moins grande exposition aux mêmes événements linguistiques ou à des événements similaires facilite la routinisation puis l'automatisation de l'accès aux formes-sens ; il est donc nécessaire de reproduire cette articulation entre les événements linguistiques (les discours constituant des corpus), le stockage de ces évènements et des ressources dérivées (dictionnaires de formants, de lexies et de patrons lexico-syntaxiques qui pourront servir de dictionnaires de référence, mais aussi, pour les formes non attestées de dictionnaires de néologismes) ;
- Le système d'analyse des corpus bruts doit pouvoir apprendre, de façon continue et itérative, à partir des seules ressources qui sont mises à sa disposition, à la fois les associations formes-sens, et leurs différentes propriétés linguistiques (combinatoire, distributions) et les informations socio-pragmatiques liées aux discours dans lesquels ils sont rencontrés ;
- Le système doit permettre une interaction entre les processus automatiques et l'expertise humaine : les résultats automatiques ne sont jamais fiables à 100%, et il est donc nécessaire d'inclure des possibilités de modification des résultats

13. `http://www.cnrtl.fr/lexiques/morphalou/`

automatiques par l'expert linguiste, qui pourra également ajouter de l'information linguistique aux informations extraites ; l'interaction doit également aller dans l'autre sens, les décisions humaines pouvant être réinjectées dans les processus automatiques qui, ainsi, pourront être améliorés de manière itérative.

Techniquement, l'articulation corpus - ressources linguistiques peut être mise en place en prévoyant un système récupérant de façon continue des corpus (par exemple sur le web) et stockant, d'une part, ces corpus et toutes les informations qui pourront en être dérivées (métainformations liées à la source d'information, analyse morphosyntaxique, syntaxique, sémantique des documents, etc.) dans un moteur de recherche, permettant des requêtes enrichies des corpus mémorisés ; stockant, d'autre part, dans des entrepôts spécifiques (bases de données), des informations dérivées de l'ensemble des corpus (par exemple, les formes qui se répètent, c'est-à-dire une ébauche de dictionnaire, ou encore des informations lexicales choisies dans les corpus par les linguistes, avec des informations additionnelles choisies par eux).

Concernant l'articulation processus automatique / expertise humaine, elle peut techniquement être mise en œuvre via un système client - serveur, les experts linguistes pouvant se connecter au serveur via une plateforme web puis interagir avec les données brutes ou analysées automatiquement, tandis que sur le serveur s'effectuent des traitements automatiques en continu, éventuellement guidés par les informations données par les linguistes.

Plus précisèment encore, la plateforme doit comporter deux ressources en interaction, gérées par l'expert linguiste :

— **Corpus** : il s'agit du matériau brut (les paroles) et la plateforme doit en permettre une gestion par les utilisateurs référencés, afin d'ajouter, de modifier, de supprimer les sources d'informations, et les décrire manuellement le plus précisèment possible ; les processus automatiques correspondent à la récupération continue de ces données, et à leur stockage dans un moteur de recherche ;
— **Ressources linguistiques** : ces données correspondent aux unités linguistiques mémorisées suite au traitement automatique et itératif des discours (corpus) (et éventuelle validation manuelle) : ces traitements peuvent être plus ou moins sophistiqués, mais visent à reproduire ce que l'esprit humain effectue dans les processus d'apprentissage, le traitement le plus basique consistant donc à détecter des répétitions de séquences, c'est-à-dire des unités linguistiques (lexicales, constructionnelles, phrastiques, etc.). Dans le cadre d'une plateforme de suivi des innovations lexicales, ces ressources stockeraient les innovations lexicales détectées et les informations glanées au fil des corpus dynamiques ; comme nous l'avons déjà indiqué, un tel système peut être étendu à toutes les unités lexicales, en considérant que la néologie n'est qu'une partie du changement lexical.

Sur ces principes, nous détaillons ci-après les différents modules développés dans la plateforme Néoveille.

4.3 Architecture générale de la plateforme Néoveille

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 accessible depuis 2016, 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'italien, l'espagnol, l'allemand et le néerlandais. Des modules sont régulièrement développés, la présentation ici faite détaillant les fonctionnalités au 30 juin 2018.

L'architecture de Néoveille est l'aboutissement actuel de plusieurs tentatives que j'ai menées pour réaliser une plateforme pour l'analyse linguistique. L'architecture générale de Néoveille est présentée sur la figure 4.1.

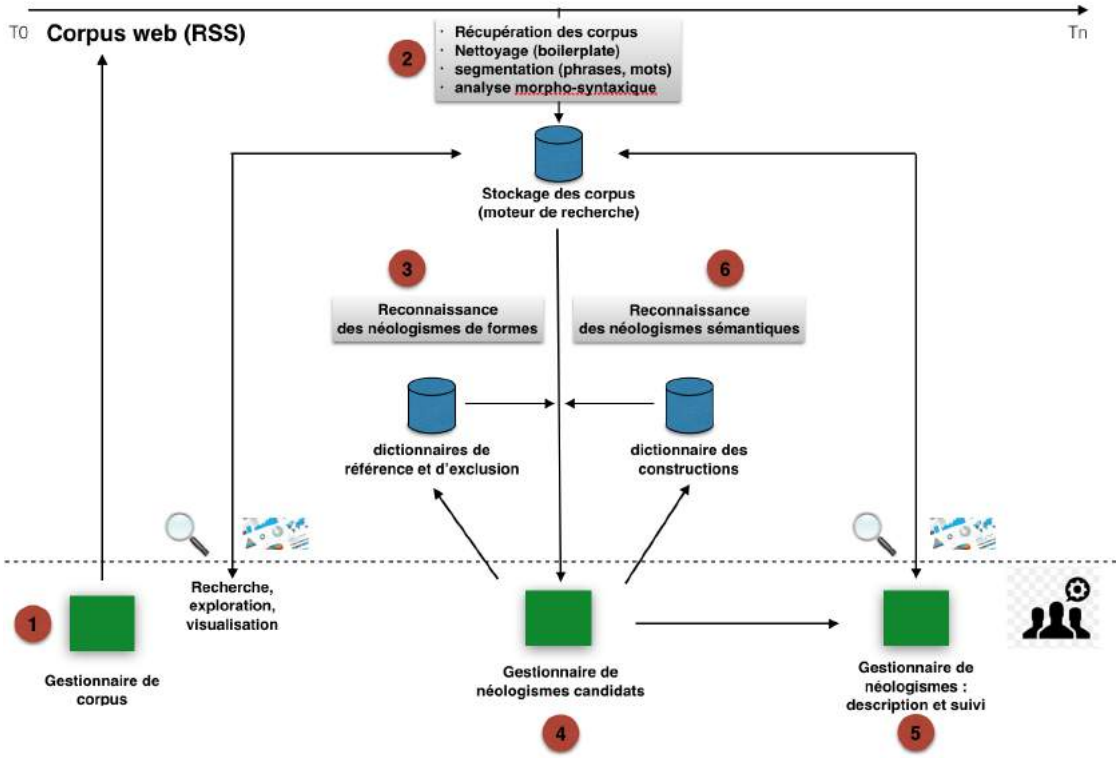


FIGURE 4.1 – Architecture générale de Néoveille

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 (processus automatiques).

On distingue six grands modules :

— **Gestionnaire de corpus** (pastille 1) : l'expert linguiste peut déterminer (ajouter, supprimer, modifier) les corpus qu'il souhaite faire analyser par le système, actuellement soit un fil RSS, soit un site web. Il peut expliciter par ailleurs un certain nombre de méta-informations : nom du journal, url d'entrée, catégorie des informations fournies (presse générale ou spécialisée à l'heure actuelle), domaine (informatique, santé, économie, mode, etc.), langue (parmi les sept langues du projet), pays du journal (cette information pourra servir ultérieurement à étudier des différences néologiques par pays pour une même langue), type de la ressource (site web ou fil RSS actuellement), fréquence de parution. Ces informations sont associées à chaque unité d'information (« article ») qui sera récupérée et pourront permettre de filtrer les résultats dans le moteur de recherche.

— **Récupération continue des fils RSS, des articles liés, analyse linguistique et stockage dans un moteur de recherche** (pastille 2) : ce module permet d'effectuer la récupération régulière des articles de presse explicités dans les fils RSS et les pages web et d'effectuer différents traitements linguistiques. Les

données sont enfin stockées dans un moteur de recherche, pouvant par la suite être directement interrogé ;

— **Détection automatique des néologismes de forme** (pastille 3) : ce module permet, de détecter dans les articles de presse stockés dans le moteur de recherche des candidats néologismes de forme par application de différents algorithmes ; ce module sera détaillé dans le chapitre 6 ;

— **Détection automatique des néologismes sémantiques** (pastille 6) : ce module permet, de détecter dans les articles de presse stockés dans le moteur de recherche des candidats néologismes sémantiques par application de différents algorithmes ; ce module sera détaillé dans le chapitre 7 ;

— **Gestionnaire des néologismes candidats** (pastille 4) : Les candidats néologismes (de forme ou sémantiques) sont détectés automatiquement, et 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 les systèmes automatiques ;

— **Le gestionnaire de néologismes** (pastille 5) est une base de données pour partie inspirée de projets développés en collaboration avec Jean-François Sablayrolles (Cartier, 2011a). Ce gestionnaire permet de décrire linguistiquement les néologismes validés, et d'obtenir, au fur et à mesure du temps, un suivi de leur cycle de vie ;

4.3.1 Gestionnaire de corpus

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 (voir figures 4.2 et 4.3). Actuellement, le système permet de récupérer des fils RSS [14], ainsi que des sites web. À chaque source d'information sont associées des méta-informations, permettant de les caractériser : nom du journal, url d'entrée, public visé (presse générale ou spécialisée 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.

14. Nous renvoyons par exemple à `http://www.rssboard.org/rss-specification` pour une présentation détaillée du format XML des flux RSS.

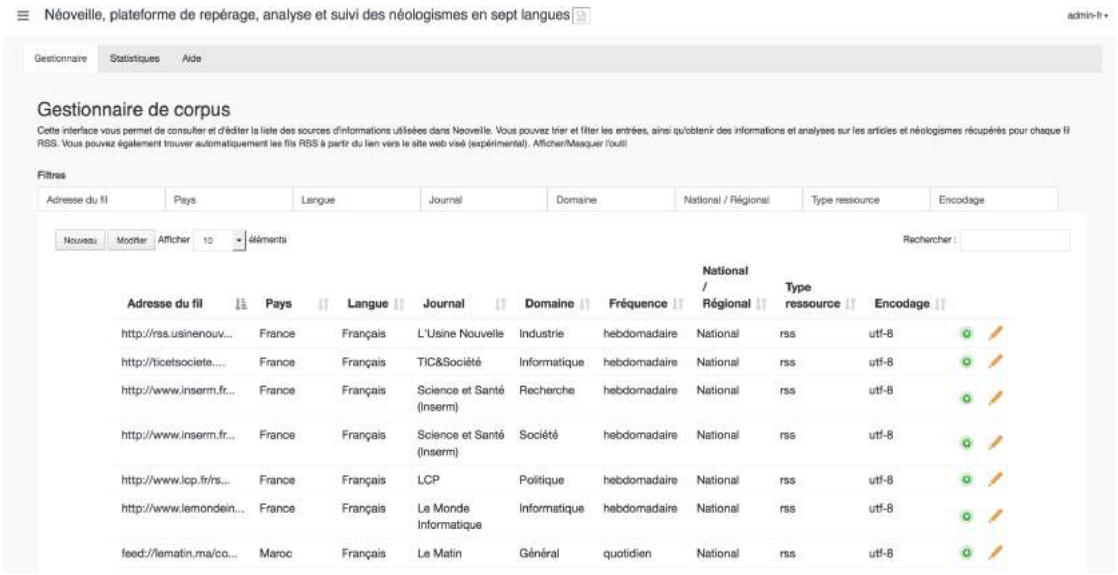


FIGURE 4.2 – Interface principale du gestionnaire de corpus, sous forme de tableau. Les différents boutons permettent d'ajouter des flux, de les modifier ou de les supprimer. La zone filtres permet de filtrer une sous-partie des flux disponibles.

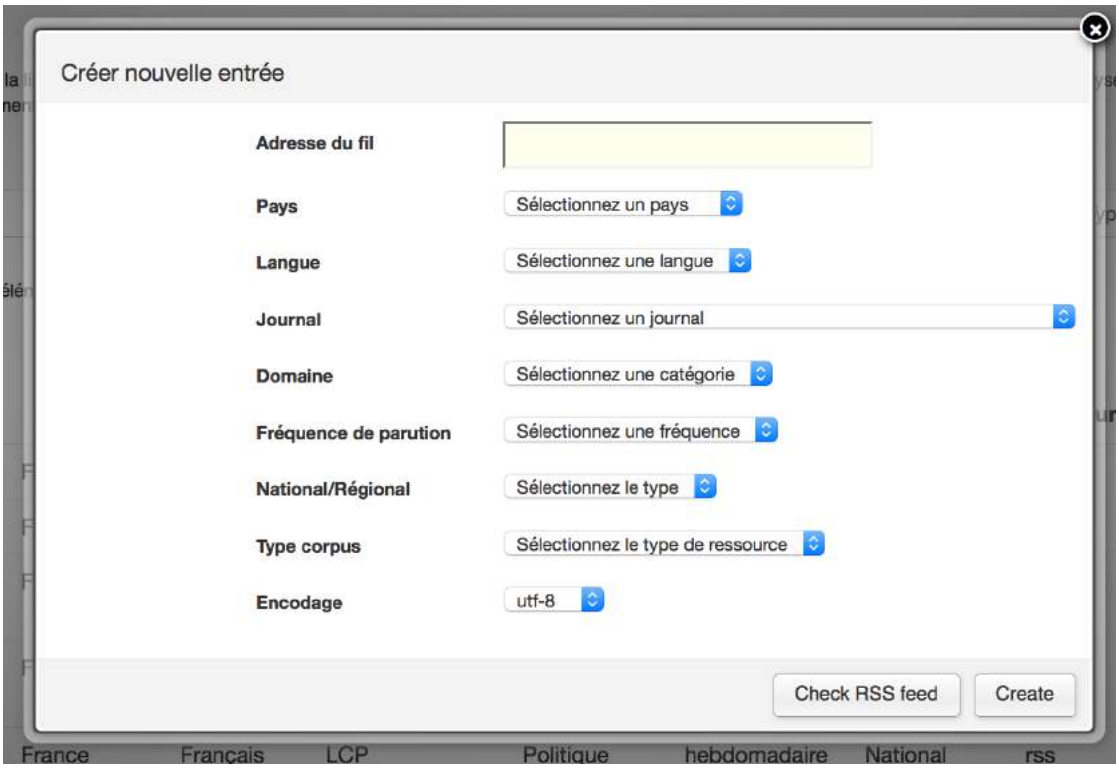


FIGURE 4.3 – Interface d'édition ou d'ajout d'un nouveau flux avec les différentes méta-informations à saisir.

4.3.2 Récupération et analyse linguistique des fils RSS

Les sources d'informations stockées sont ensuite récupérées automatiquement deux fois par jour. Des (méta-)informations complémentaires pourront être récupérées pour chaque fil d'information, dans le flux RSS : titre du document, auteur, date de publication, liste de mots-clés, domaines. Les fils RSS comprennent 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 [15]. Enfin, toutes ces informations sont stockées dans un moteur de recherche Apache Solr, pour exploitation ultérieure. Au final, chaque source d'information dispose ainsi de plusieurs informations complémentaires, résumées dans le tableau 4.1.

4.3.2.0.1 Perspectives L'inconvénient majeur des fils RSS ressortit à la non-systématicité de l'utilisation 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. L'accès à des sites web dans un empan temporel plus large a par ailleurs été entamé via un projet avec la Bibliothèque Nationale de France, via une collection appelée "Actualités", couvrant la presse en ligne généraliste depuis 2010-2017 (Cartier *et al.*, 2018a; Cartier *et al.*, 2018b). D'autre part, un module complémentaire de détection des thématiques de chaque texte est en cours de développement, en combinant les approches à base de ressources linguistiques et les méthodes non-supervisées (Cartier *et al.*, 2018a; Cartier *et al.*, 2018b). 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.

4.3.3 Détection automatique des néologismes de forme

Voir chapitre 5.

4.3.4 Détection automatique des néologismes sémantiques

Voir chapitre 6.

4.3.5 Gestionnaire de néologismes candidats

Le module de gestion des néologismes candidats propose aux experts linguistes le résultat des détections automatiques de néologismes formels et sémantiques. Il s'agit ici

15. Nous renvoyons à (Cartier, 2016b) pour une présentation détaillée de ces traitements : actuellement, nous utilisons JusText [16] pour effectuer le zonage, et Treetagger [17] pour l'analyse morphosyntaxique.

Type général d'information	Type d'informations	Informations complémentaires
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 [18]
	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.
	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 enrichi (analyse morpho-syntaxique)	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 [19] et le lemme.

TABLE 4.1 – Liste des informations disponibles pour chaque item d'information textuelle récupéré dans Néoveille

de valider ou d'invalider le processus automatique, en associant à chaque "néologisme candidat" une catégorie, se ramenant à "néologisme" (forme ou usage nouveau) ou "non-néologisme" (lexie en usage). les résultats de la catégorisation peuvent ensuite être validés afin d'alimenter le dictionnaire des néologismes, ou bien le dictionnaire des lexies ou des constructions en usage. Nous présentons ci-après : les interfaces pour effectuer ces opérations, les catégories de non-néologismes formels et les non-néologismes sémantiques (renvoyant au détail dans les chapitres 3 et 4), ainsi que la méthodologie de validation des néologismes candidats implémentée jusqu'ici.

4.3.5.1 Présentation des interfaces du module

Les résultats de la détection automatique des néologismes candidats (NC) formels et sémantiques sont présentés aux experts linguistes sous forme d'un tableau (voir figure

4.4).

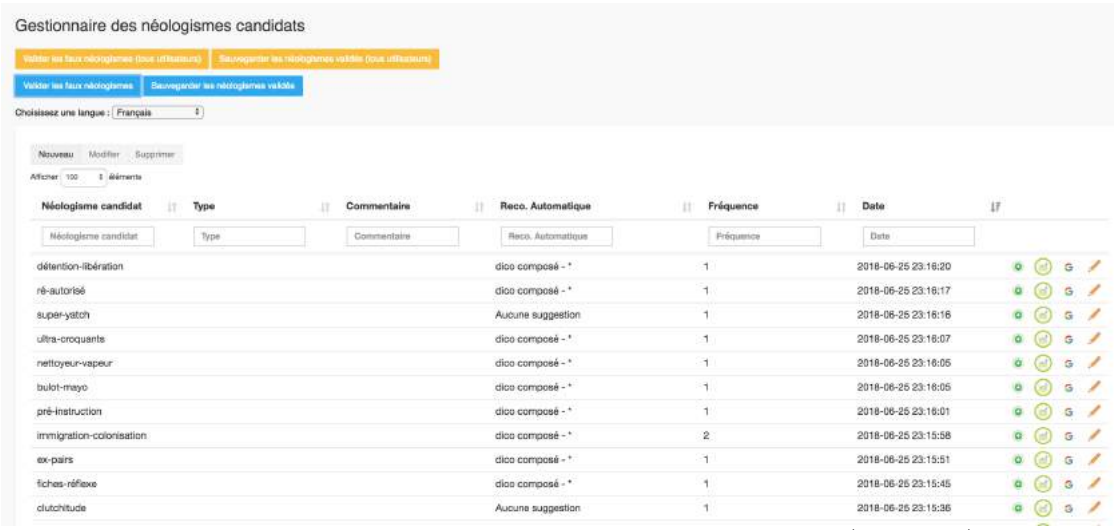


FIGURE 4.4 – Interface de validation-invalidation des néologismes (de forme) automatiquement détectés.

Cette interface présente un certain nombre d'informations pour chaque NC, de gauche à droite : forme exacte reconnue, type (non renseigné), commentaire(non renseigné), informations sur la reconnaissance automatique, suggérant à l'utilisateur la catégorie du NC, information sur la fréquence constatée dans le corpus du NC, information sur la date de la première occurrence rencontrée. Les deux champs type et commentaire sont à renseigner par les experts linguistiques, d'une part pour indiquer le type du NC (voir plus loin), d'autre part, pour saisir tout commentaire souhaité. La saisie de ces informations peut se faire soit en cliquant sur la cellule concernée, soit en sélectionnant la ligne puis en cliquant sur le bouton modifier (à droite). Les informations du tableau peuvent être filtrées, triées et il est possible de modifier le nombre de lignes affichées sur chaque page.

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 4.5).
— la visualisation enrichie d'informations sur les caractéristiques socio-pragmatiques des contextes 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 4.6).
— 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).

Une fois modifiée l'information de type, il est possible de valider les néologismes et les

FIGURE 4.5 – Exemple de visualisation de contextes pour le candidat néologisme *clutchitude*.

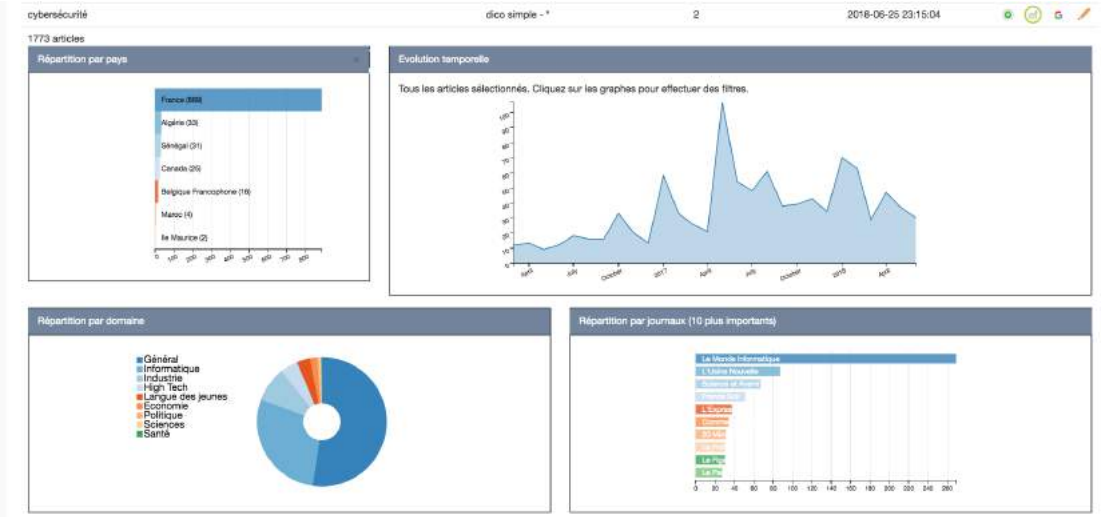


FIGURE 4.6 – Exemple de visualisation enrichie (les contextes sont omis) pour le candidat néologisme *cybersécurité*.

non-néologismes en cliquant sur les boutons situés au-dessus du tableau (Valider les faux-néologismes, sauvegarder les néologismes validés) : cela aboutit à retirer de cette liste les NC concernés et de les reverser vers l'un des dictionnaires de référence ou d'exclusion, soit de les reverser vers la base de données des néologismes validés, pour description linguistique et suivi de l'évolution.

4.3.5.2 Critères définitoires des néologismes et typologies des non-néologismes

Nous renvoyons au chapitre 2 pour une définition générale des néologismes et aux chapitre 5 (néologismes de forme) et au chapitre 6 (néologismes sémantiques) pour les différents types de néologismes. Nous évoquerons ici quelques critères objectifs que nous avons établi afin de délimiter le plus précisément ce que nous entendons par néologisme, notamment afin d'éviter le recours au *sentiment néologique* :le jugement de néologicité est évidemment lié à l'expérience linguistique (sous tous ses aspects) de chaque individu, d'une part, et la notion de langue standard étant extrêmement complexe à délier, d'autre part. Le présent travail n'échappe pas à cette difficulté. Nous avons opté pour un faisceau de critères à utiliser pour juger de la néologicité des lexies. Tout d'abord en utilisant à la fois des dictionnaires et des corpus de référence :

— absence d'occurrences dans le corpus Google Ngrams et dans le corpus Google avant 2010 ;

— absence dans les dictionnaires de référence (Le Larousse et Le Robert essentiellement) jusqu'en 2015.

En outre, nous avons mis au point une typologie des non-néologismes, présentée dans le tableau 4.2 :

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 ayant les propriétés d'un code-switching et dénotant une réalité locale	Lujo, furoshiki, rojigualda, tawakkul, etc.
gentilé	Lexie désignant un individu ou une caractéristique lié à un lieu ou une zone géographique spécifique	Amuesha, cubano-mexicaine, sino-russe, etc.
particularisme	Lexie entrée dans l'usage pour une aire socio-géographique spécifique	Xessal, tcha-tcho
Erreur typographique et autres erreurs	Erreurs diverses liées à l'orthographe	Spect, terroristea, berbatov, jijadiste, acceuille, traditionel, endless, etc.

TABLE 4.2 – 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 du système automatique (voir chapitre 6). Les lexies terminologiques représentent une autre catégorie qui montre la porosité entre le vocabulaire spécialisé et le vocabulaire général : des passages sont effectués régulièrement de l'un à l'autre par le biais des articles de vulgarisation de la presse. Les gentilés sont une autre catégorie de lexies exclue : il s'agit de formations à partir de noms propres géographiques ou socio-ethniques, construites régulièrement (notamment par suffixation : *les nzebis*, et par composition : *sino-russe, anglo-néo-zélandais*). Les particularismes sont plus rares.

Un cas complexe concerne le continuum entre les xénismes et les emprunts proprement dits. En effet, le passage de l'un à l'autre est toujours possible, comme en attestent les *sushis, pizzas, cookies* pour nous limiter au domaine culinaire. Cela explique la catégorie particulière (non-néologisme) des xénismes, qui est en quelque sorte, pour certaines formations, l'antichambre de l'emprunt, que pourra attester un suivi diachronique.

4.3.5.3 Méthodologie pour la validation/invalidation des néologismes candidats

Les néologismes sont d'abord détectés automatiquement par différentes méthodes (voir chapitres 5 et 6). En moyenne, pour les néologismes formels, pour le français, entre 100 et 200 néologismes candidats (NC) sont repérés chaque jour. La méthodologie de validation des néologismes suit le protocole suivant :

1. validation individuelle par les experts linguistes : chaque membre du groupe de travail [20] annote 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 section précédente). À ce stade, chaque membre du groupe travaille sur une liste distincte de néologismes candidats.
2. lors de réunions collectives mensuelles, une validation est effectuée pour les néologismes qui ont été typés individuellement lors de la première phase, après présentation par l'annotateur, 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 et un affinement des critères de validation des néologismes.

Au final, ce processus de validation a permis de constituer un jeu de référence (pour le français et le russe) et donc permis de vérifier le taux de précision du repérage automatique, qui est proche de 60 % pour le français. De juillet 2015 à décembre 2017, à partir d'environ 250 sources d'informations, 1 143 912 articles (92 millions de mots, 1 037 876 formes différentes) ont été récupérés. Parmi environ 35 000 néologismes formels candidats, 22 475 néologismes ont été validés, correspondant à 726 222 occurrences.

Il faut noter également que les non-néologismes sont ensuite reversés dans les différents dictionnaires de référence et d'exclusion, et sont utilisés ensuite par le détecteur automatique, permettant un apprentissage actif.

4.3.6 Gestionnaire des néologismes

Le gestionnaire des néologismes récupère les néologismes validés et permet de les décrire linguistiquement. La description linguistique des néologismes consiste tout d'abord à caractériser les mécanismes de formation. En utilisant une version simplifiée du modèle de (Sablayrolles et Pruvost, 2016). Néoveille propose par ailleurs une micro-structure générique pour l'ensemble des néologismes comprenant les champs détaillés dans le tableau 4.3:

A ces informations, permettant notamment par la suite d'effectuer des statistiques sur les procédés les plus couramment utilisés, les parties du discours les productives, etc. il faut ajouter d'autres informations concernant des néologismes spécifiques, par exemple,

20. Chaque langue dispose d'un groupe de travail. Pour le français, 7 personnes travaillent plus ou moins régulièrement sur le projet, ainsi que des étudiants de manière plus ponctuelle. Nous renvoyons au site internet pour la liste des personnes impliquées dans ce travail. Lorsque des étudiants sont impliqués dans le travail de validation, ils sont sous le contrôle d'un référent qui valide les annotations effectuées.

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)	
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 Etats-Unis.
Procédé(s) néologique(s) impliqués	Le ou les mécanismes néologiques impliqués dans l'innovation lexicale	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
Configuration morphologique	Décomposition morphologique de l'innovation, au moyen des notions de radical, d'affixe et de formant.	RAD RAD
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

TABLE 4.3 – Informations linguistiques de base pour les néologismes

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. De plus, un champ supplémentaire permettra de décrire le mécanisme néologique précis selon le formalisme présenté dans le chapitre 1 (section "Unités lexicales") et qui sera détaillé dans le chapitre 6.

À ces informations, il faut encore 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, les collostructions (Stefanovitsch et Gries, 2003) et les constructions lexico-syntaxiques les plus fréquentes ; le profil distributionnel, permettant d'accéder aux lexies sémantiquement similaires (et donc notamment (quasi-)synonymes, hyperonymes et hyponymes).

Nous illustrons les deux premières informations dans le tableau 4.4, la troisième nécessitant un corpus plus étendu pour obtenir des résultats fiables.

Nous reviendrons sur les méthodes utilisées pour obtenir ces données dans le chapitre 7.

4.3.7 Outils de suivi de l'évolution des néologismes

En dehors des informations linguistiques, la plateforme propose une visualisation interactive des contextes socio-pragmatiques des occurrences des néologismes. les méta-

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	Collocations : fast food (16), slow 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...

TABLE 4.4 – Informations combinatoires disponibles pour les néologismes

informations liées aux différentes articles (voir méta-informations dans le tableau 4.1) permettent en effet de se faire une idée précise des contextes d'occurrences : journal, public visé, type de texte, domaine(s), aire géographique, auteur(s) informent sur les caractéristiques diastratiques et diaphasiques des textes sources, et la date de publication permet d'obtenir une vision dynamique de l'évolution de ces caractéristiques. Ces visualisations concernent également les informations linguistiques décrites précédemment.

A titre d'illustration, nous présentons, pour l'emprunt *slow-food*, les distributions générales (à gauche) et leur évolution temporelle (à droite) concernant les formes et les dérivés morphologiques 4.7, les domaines 4.8 et les journaux 4.9.

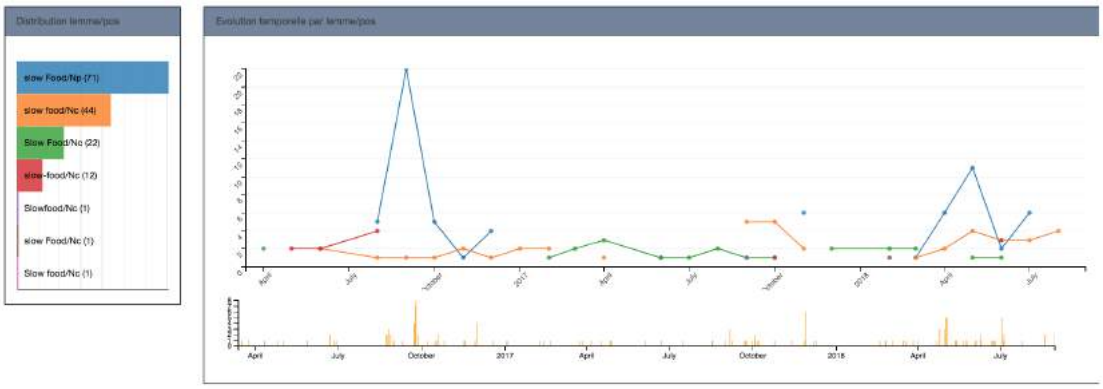


FIGURE 4.7 – Distribution générale et évolutive des formes et dérivés morphologiques de *slow-food*

On constate ainsi que l'emprunt a dès le départ connu trois variantes orthographiques (*slow food, slow-food, slowfood*), avec une légère domination du premier, qu'il est utilisé dans la presse généraliste et dans la presse féminine (dans une bien moindre mesure), et qu'il n'est pas attesté dans près d'une dizaine de journaux, montrant ainsi sa diffusion

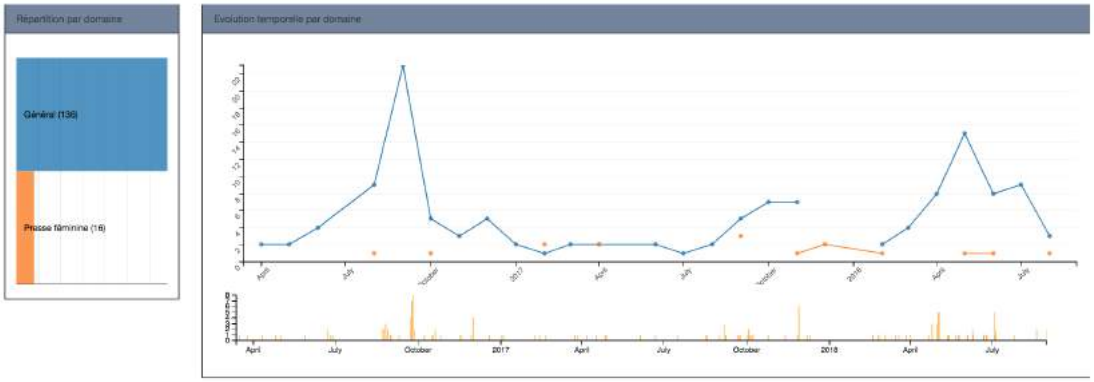


FIGURE 4.8 – Distribution générale et évolutive des domaines pour *slow-food*

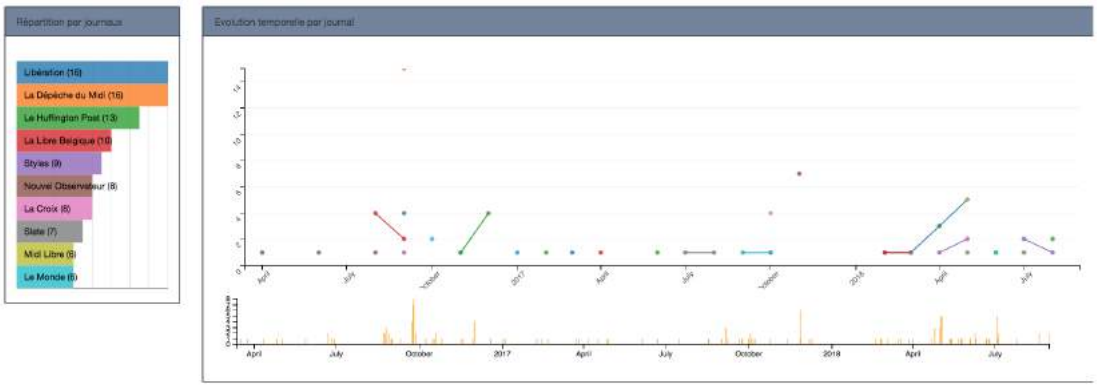


FIGURE 4.9 – Distribution générale et évolutive des journaux pour *slow-food*

rapide. La figure 4.10 montre finalement l'évolution temporelle de ses occurrences, indiquant un néologisme qui ressurgit sur la période selon les événements, mais qui reste d'une fréquence très faible.

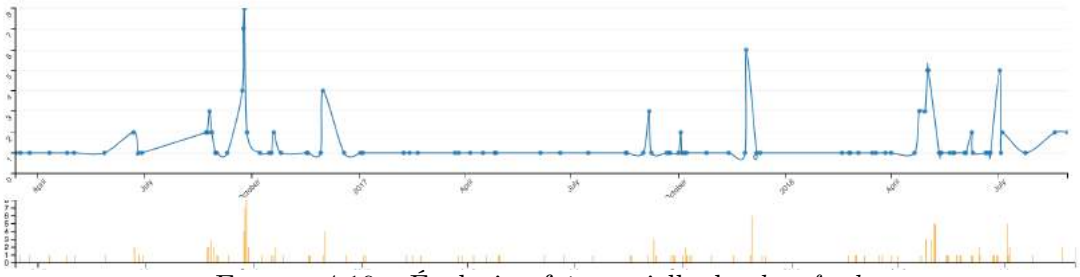


FIGURE 4.10 – Évolution fréquentielle de *slow-food*

4.4 Conclusion

Ce chapitre s'est focalisé sur l'architecture idéale d'une plateforme pour l'étude des néologismes en corpus. Le propos s'est organisé en deux sections, la première évoquant les différents travaux réalisés jusqu'à présent, qu'il s'agisse des outils généralistes de recherche ou de moteurs de recherche plus spécialisés dans la fouille de corpus, ou encore des outils spécifiquement dédiés à l'étude des néologismes.

De cette étude, nous avons extrait un certain nombre de composants et de fonctionnalités nécessaires pour une plateforme "idéale" : un gestionnaire des corpus, comprenant la liste des sources d'information qui seront récupérées régulièrement, et permettant l'ajout de méta-informations permettant de les caractériser socio-pragmatiquement ; un entrepôt de mémorisation de ces corpus une fois analysés, c'est-à-dire un moteur de recherche ; des ressources linguistiques permettant de stocker les informations glanées par les processus automatiques et/ou saisies par les experts linguistes, c'est-à-dire des bases de données lexicographiques, sans prendre partie pour telle ou telle structuration de l'information.

A partir de ces composants, trois processus automatiques principaux peuvent être mis en œuvre de manière continue:

— la récupération des sources d'informations, l'ajout de méta-informations, manuellement définies ou semi-automatiquement détectées, informant sur les spécificités socio-pragmatiques de chacune de ces sources d'informations, l'analyse linguistique des contenus textuels puis le stockage des résultats dans un moteur de recherche pour traitements ultérieurs et/ou interrogation enrichie ;
— la détection automatique des néologismes formels dans ces corpus enrichis et le stockage des résultats dans une base de néologismes candidats pour validation par les experts linguistes ;
— la détection automatique des néologismes sémantiques dans ces corpus enrichis et le stockage des résultats dans une base de néologismes candidats pour validation par les experts linguistes.

A ces trois processus, il faut ajouter les différentes interactions avec les experts linguistes :

— la gestion des corpus : édition complète (ajout, modification, suppression) , ajout de méta-informations à chaque source d'information ;
— la validation ou l'invalidation des néologismes automatiquement détectés ;
— la description linguistique des néologismes retenus ;
— l'interrogation enrichie des contenus textuels contenant les occurrences de néologismes, et des outils pour le suivi dynamique de l'évolution temporelle des propriétés linguistiques et socio-pragmatiques des néologismes.

Nous avons alors présenté la plateforme Néoveille, qui nous semble aujourd'hui la solution la plus complète pour détecter et suivre les néologismes en corpus : module de gestion de corpus, récupération continue des sources d'information et stockage dans un moteur de recherche, détection automatique des néologismes formels, validation/invalidation par les experts linguistes, module de description linguistique des néologismes

validés, outils de visualisation interactive de l'évolution temporelle des propriétés des néologismes.

Non seulement cette plateforme dispose de fonctionnalités qui nous paraissent incontournables pour l'étude des néologismes en corpus (et de manière plus générale, pour l'étude des lexies en corpus), mais elle répond au modèle théorique que nous avions fixé dans les premiers chapitres : prise en compte de l'interaction continue langue / discours et mémorisation (la langue), prise en compte des dimensions diatopiques, diastratiques et diasituationnels de la langue, prise en compte de la diachronie.

Le système présenté est disponible aux chercheurs intéressés : une interface publique propose les résultats des analyses effectuées dans les différentes langues du projet, et une interface privée sécurisée permet aux chercheurs qui le souhaitent d'effectuer différentes analyses.

Cependant, la réalisation actuelle n'est pas parfaite et nous mentionnerons pour terminer quelques défis à relever, concernant d'abord les sources d'informations [21] :

— gestion des corpus : actuellement, le système récupère les informations sur internet via des fils RSS ; une extension est en cours afin de pouvoir également récupérer des sources internet brutes ; il sera également nécessaire, à plus long terme, de prévoir l'ajout de corpus oraux et multimédia, qui sont sans doute plus productifs en néologismes ;

— assignation de propriétés socio-pragmatiques aux sources d'information : actuellement, ce type d'informations est donné de manière statique par les utilisateurs du système ; il conviendrait de mettre en place des méthodes pour la détection automatique d'informations plus précises sur les documents eux-mêmes : pour cela une modélisation approfondie des situations de communication serait nécessaire ; de même, une analyse automatique des parties des textes serait nécessaire : aujourd'hui, le système ne sait pas distinguer, dans les articles de presse, l'article lui-même, les commentaires, les renvois, les citations longues, etc. Ces zones ne sont pragmatiquement pas les mêmes, et les néologismes qui s'y trouvent n'ont du coup pas le même statut.

— assignation automatique de thématiques aux textes : une autre faiblesse du système actuel ressortit aux thématiques des textes : cette information est importante, car le ou les thèmes traités par les documents ont une incidence sur le domaine d'application des néologismes, qui peut par exemple être très limité si nous nous trouvons dans un document traitant d'informatique et qui utilise un jargon spécifique ; actuellement, l'assignation des thèmes est effectuée dans la source d'information elle-même, et devrait être plus dynamique ; nous avons effectué de premiers tests dans un autre projet (Cartier *et al.*, 2018a; Cartier *et al.*, 2018b) utilisant les méthodes classiques du *topic modeling* montrant l'intérêt d'une détection automatique de thèmes.

Un autre élément de recherche concerne le traitement des langues : actuellement, le système traite onze langues, en utilisant l'analyseur morphosyntaxique Treetagger.

21. Les améliorations des détecteurs automatiques de néologismes seront traitées dans les deux prochains chapitres

Celui-ci présente l'avantage de proposer de nombreux modèles de langues, mais présente l'inconvénient d'une qualité d'annotation plus faible que les systèmes les plus récents, d'une part, et de nécessiter la mise en place d'une chaîne de traitement spécifique lors de l'ajout d'une nouvelle langue. Nous travaillons actuellement, afin de faciliter l'intégration de nouvelles langues, à l'utilisation d'analyseurs plus récents. Nous avons, dans le projet Néonaute, utilisé l'outil Spacy [22], qui a donné satisfaction, et qui est théoriquement multilingue par développement de modèles de langue à partir de corpus annotés.

D'autres développements sont nécessaires, au point où nous en sommes, et qui concerne notamment la valorisation de cet outil : il s'agira très prochainement de mettre en *Open Source* le code. D'une part, afin que différents serveurs puissent être mis en place, sur des corpus autres que ceux qui nous ont intéressés - corpus spécialisés, corpus sur objectifs spécifiques, etc. Des collègues de Paris 7 (CLILLAC-ARP) sont actuellement en cours d'installation d'une version de Néoveille qui pourra être utilisée sur des corpus terminologiques. D'autre part, afin que s'organise autour de cette plateforme une communauté de développeurs et de chercheurs pour améliorer les fonctionnalités de la réalisation actuelle.

Enfin, plusieurs présentations du système dans différentes conférences internationales ont permis de constituer un réseau de chercheurs sur l'étude des changements lexicaux en corpus, et l'application Néoveille a favorablement été accueillie. Il faut envisager un projet plus vaste que celui qui a permis d'arriver jusqu'ici.

22. `spacy.io`

Chapitre 5

Repérage et suivi des changements lexicaux : néologie formelle

Sommaire

Dans ce chapitre, je focalise sur la néologie formelle. Il s'agit d'abord de modéliser précisèment le phénomène linguistique, et de proposer une catégorisation des procédés impliqués, en se basant à la fois sur les travaux qui ont pu être conduits en morphologie, ainsi que par les travaux des chercheurs intéressés par la néologie. À l'aide de cette modélisation, nous présentons les différentes méthodes de repérage automatique de la néologie formelle, puis la méthode que nous avons mise en œuvre dans l'outil Néoveille, avec une évaluation.

Le propos est donc organisé en trois sections : la première propose une caractérisation fine de la néologie formelle, à savoir une définition initiale, une discussion générale sur les procédés d'affixation et de composition, une typologie des procédés néologiques

impliqués et la notion de productivité. La seconde partie décrit les méthodes de repérage automatique des néologismes formels utilisées dans cette étude. Le propos se termine par une conclusion dans laquelle je propose une typologie des néologismes formels complétant la typologie proposée par (Sablayrolles et Pruvost, 2016) mais qui articule les éléments de cette typologie autour de deux opérations primaires : la réduction et la construction. Dans la partie applicative, les tendances morphologiques du français telles qu'elles ressortent de l'analyse d'environ 22 000 néologismes de forme collectés par Néoveille puis validés par l'équipe travaillant sur le français seront présentées.

5.1 Modélisation des néologismes de forme

Nous renvoyons aux chapitres 1 à 3 pour la modélisation globale des changements lexicaux. Ici, nous évoquons successivement : la délimitation du phénomène de néologie formelle, la typologie des néologismes formels, les types d'unités linguistiques touchées par le phénomène, la notion de productivité puis la description des mécanismes de formation des nouvelles unités lexicales.

5.1.1 Délimitation de la néologie formelle : mécanismes de formation des mots

Dans cette première section, nous tenterons d'établir le périmètre de la néologie formelle, en établissant les frontières entre la néologie formelle et d'autres champs. Il s'agira d'abord de rappeler la distinction néologie formelle / néologie sémantique, qui fonde de manière méthodologique le chapitrage utilisé ici. Nous tracerons également les frontières avec deux champs adjacents, celui de la flexion, d'une part, et celui des unités polylexicales, d'autre part.

5.1.1.1 Néologie formelle / néologie sémantique : rappel

Comme indiqué dans le chapitre 2, la distinction néologie formelle / néologie sémantique est méthodologiquement pratique, même si théoriquement les deux types se situent dans un continuum, notamment dans le cadre des grammaires de construction. La distinction repose sur une conception restreinte de la forme lexicale, réduite à sa forme morphologique, telle qu'on peut par exemple la rencontrer dans les dictionnaires : *surfer* est une lexie, même s'il est possible de décomposer cette unité formelle en deux composants, une base ou radical *surf* et un morphème *-er*. La néologie formelle est alors caractérisée par la création d'une nouvelle forme (ici par exemple, l'emprunt à l'anglais *surf*) ou encore d'un formant lexical (par exemple *cyber-*, *e-* ou encore *-thon*). Dans ce cadre, sans création d'une nouvelle forme lexicale (simple ou composée) ou infra-lexicale, tout autre néologisme sera dit sémantique puisqu'alors une forme existante est employée avec un sens nouveau, qu'il s'agisse d'une unité polylexicale qui devient non transparente (*action de groupe* dans le sens d'une action juridique menée par un ensemble de travailleurs pour faire reconnaître ses droits, ou encore *plan social* qui acquiert, à partir des années 2000 un sens spécifique en droit du travail), ou d'un emploi métonymique ou métaphorique

(par exemple *marathon* qui acquiert à partir des années 50 le sens figuré d'une activité longue et difficile exigeant de l'endurance, ou encore *souris* qui prend un sens spécifique dans le domaine informatique). Mais, dans une conception élargie dans la lignée des grammaires de construction, avec la notion de signe (construction) assimilée à une paire forme-sens, seul le critère du sens « nouveau »(et sa (quasi-)non-décompositionnabilité) permet d'établir le périmètre de l'innovation : il peut s'agir aussi bien d'un formant (affixes et morphèmes grammaticaux), d'un mot (*arriver, souris*), d'un syntagme plus ou moins figé (*effet de serre, ramasse-miettes*), d'un emploi lexico-syntaxique plus ou moins restreint (*SN arriver à VerbeInf, SN arriver à SN (lieu)*), et même d'une construction syntaxique non sémantiquement contrainte (*il pleuvoir*). La forme est alors conçue dans un sens étendu : toute séquence ayant un sens non-décomposable (ou seulement partiellement), et ayant une productivité lexicale plus ou moins importante (la lexie au sens strict ayant une productivité "formelle" nulle, les morphèmes liés et les constructions (lexico-)syntaxiques une productivité plus ou moins importante). Cette conception a l'avantage de toujours lier un sens à une forme, qui peut aller en-deçà comme au-delà du mot au sens traditionnel. Cependant, nous pouvons continuer à distinguer les formes, sur la base de leur taille (simple ou complexe), et ainsi conserver une frontière entre les formes à dominante morphologique (lexies, simples et composées, et morphèmes liés) et les formes à dominante syntaxique (unités polylexicales et constructions au sens plus traditionnel). Nous conserverons donc, pour des raisons pratiques et méthodologiques, la différence entre néologie formelle, qui se limite à une forme allant du morphème à la lexie simple ou composée et la néologie sémantique (les autres créations au-delà de la composition). Même avec ces définitions, il faut encore préciser les frontières, d'une part avec les flexions, d'autre part avec les unités polylexicales.

5.1.1.2 Flexion et dérivation

Une première frontière du périmètre de la formation des mots concerne la distinction dérivation / flexion. Nous partirons de (Riegel *et al.*, 2018) et de (Booij, 2006), ce dernier détaillant un certain nombre de critères pour établir la distinction, tout en insistant sur le continuum qui existe entre les deux champs.

(Riegel *et al.*, 2018, p.536-537) établissent un certain nombre de critères classiques pour distinguer les morphèmes grammaticaux et les morphèmes dérivationnels/lexicaux :

- **critère quantitatif** : « les morphèmes lexicaux appartiennent à des ensembles nombreux et ouverts qui se renouvellent constamment [...] les morphèmes grammaticaux constituent des ensembles clos et très restreints »(ibid, p.536). Cependant, les auteurs parlent ici de deux types de morphèmes lexicaux, les radicaux et les affixes lexicaux ; ces derniers sont en bien plus petit nombre ;
- **critère fonctionnel** : « les morphèmes grammaticaux contribuent de manière décisive à l'organisation grammaticale de la phrase »(ibid, p.536), à la fois par le biais des affixes flexionnels mais également par le biais des mots outils. Ils formeraient donc un paradigme fermé, dont l'évolution ne peut être qu'extrêmement lente, car une modification entraînerait une modification de pans entiers du système ;

— **critère sémantique** : les morphèmes grammaticaux dénotent des sens très généraux, souvent dépendant de la « situation d'énonciation (relations intersubjectives, temporalité, quantification, détermination, etc.) »(ibid, 536-537),alors que les morphèmes lexicaux couvrent l'ensemble des champs lexicaux. mais, en restreignant aux affixes dérivationnels, on constate que certains affixes ont également des sens très généraux (négation, opposition, quantification, évaluation, etc.) ;
— **critère formel** : une autre différence ressortirait à la longueur des formes, très courtes (souvent une syllabe) pour les désinences, alors que les morphèmes lexicaux seraient plus disparates de ce point de vue. Ce critère est également contestable, puisque les affixes lexicaux sont généralement également très courts (même s'il existe beaucoup d'affixes à deux syllabes).

On constate donc que, en dehors du second critère, la situation est plus floue qu'il n'y paraît au premier abord.

(Booij, 2006) reformule le critère fonctionnel : tandis que la dérivation sert à former de nouvelles lexies, la flexion sert à créer des variantes de la même lexie [1]. Cette différence dessine les deux champs principaux de la morphologie, « science de la forme des mots » : d'une part la morphologie grammaticale (ou flexionnelle) qui traite des marques (dites flexions ou désinences) liées aux différentes parties du discours et ajoutées aux radicaux des lexies, et la morphologie lexicale, qui traite des procédés de formation des mots.

Ce qui réunit les deux types de processus ressortit à l'utilisation, dans les deux cas, dans un grand nombre de langues, du procédé d'affixation (principalement suffixation pour la flexion), de l'adaptation phonologique et des processus de reduplication. Plusieurs auteurs (Bybee, 1985 ; Dressler, 1989 par exemple) pointent ainsi sur le continuum qui existerait entre les deux mécanismes, avec une tendance diachronique du passage de la dérivation (prototypique) vers la flexion (prototypique).

On peut cependant énoncer un faisceau de propriétés qui permettent de distinguer les deux procédés, tout en reconnaissant la continuité entre eux :

— **changement de partie du discours** : la dérivation peut générer une partie du discours différente de la partie du discours de la base lexématique, ce qui n'est jamais le cas de la flexion. (par exemple : *Macron (nom) > macroniser (verbe)*). Cependant, il existe aussi nombre de dérivés qui conservent la partie du discours initiale (*otage > ex-otage, science > scientifique, fille > fillette*). De l'autre côté, la flexion infinitive, en français par exemple, permet de convertir de manière quasi-systématique un verbe en nom (*parler > le parler*, mais *convertir > *le convertir*). Un mécanisme similaire permet de convertir les formes participiales du verbe en adjectifs (*dansé, dansant*). L'anglais et d'autres langues fonctionnent de la même manière (*snapchat (n) > to snapchat (v), > snapchating (n, v)*). De même, les formants créateurs d'adverbes (*-ment*, *-ly* en anglais), sont traditionnellement assimilés à des marques flexionnelles ;
— **Caractère obligatoire de la flexion versus caractère facultatif de la dé-**

1. « derivation (i.e. word-formation except compounding) is that kind of morphology that serves to create new lexemes, whereas inflection serves to create different forms of the same lexeme »(Booij, 2006, p.360)

rivation : cette propriété, qui semble établir clairement la distinction, connaît cependant un cas limite, celui des flexions non-marquées (singulier des noms par exemple), qui peut être interprété comme étant une absence de flexion ;

— **Paradigmes flexionnels** : le critère le plus utilisé concerne les paradigmes flexionnels, qui sont en nombre restreint et identifiables pour chacune des parties du discours. par exemple, les noms indiquent nécessairement le nombre et le genre, les verbes la personne, le nombre, le temps et le mode. Cette nécessité attachée aux parties du discours justifie alors la notion de flexion non-marquée, correspondant à une valeur par défaut (et non son absence). Ce critère pose le problème du statut des conversions, puisque dans ce cas, un verbe peut également, dans la même forme (et sans marquage particulier), être un nom. En allant au bout de l'argumentaire, le critère de l'existence d'un marqueur zéro sera définitoire, si nous considérons la conversion comme un procédé flexionnel. Le problème qui se pose alors est que ce procédé ne s'applique pas à tous les verbes ;

— **Généralité et productivité des paradigmes flexionnels** : une autre différence concerne le caractère général et totalement productif (pour la partie du discours considérée) des paradigmes flexionnels, ce qui n'est pas le cas des paradigmes dérivationnels, qui connaissent généralement des restrictions. En dehors de quelques exceptions (*pleuvoir, frire* en français, *courage, grace, march* en anglais), ce critère semble définitoire. Mais certaines langues agglutinatives, comme le turc, ont des processus dérivationnels extrêmement généraux ;

— **Transparence sémantique** : dans la très grande majorité des cas, les flexions ont un sens transparent, tandis que les dérivations sont parfois non-transparentes. Cependant, là encore, d'une part, un grand nombre de dérivés sont transparents, et certaines flexions ne le sont pas (par exemple différences entre singulier et pluriel : *la beauté, les beautés*) ; la contrepartie psycholinguistique de cette transparence est que les flexions sont des règles, tandis que les dérivations ne sont que partiellement des règles, car des contraintes liées aux morphèmes peuvent se produire ; mais là encore, des cas limites se présentent, notamment pour les verbes les plus courants, dont la flexion est irrégulière, et certains affixes sont tellement généraux qu'ils peuvent être assimilées à des règles (par exemple les formations en *non-*, ou encore les suffixations en *-iser/isation*) ;

— Récursivité des règles dérivationnelles : une autre différence concerne l'application récursive des règles de formation : tandis que la flexion s'applique une fois, la dérivation permet d'appliquer de manière récursive différentes formations (*anti-facebook-is-ation* en français, *anti-anti-merkel* par exemple) ;

— **Flexion et syntaxe, Dérivation et lexique**: un autre argument concerne le fait que la flexion appartiendrait à la fois au domaine morphologique et au domaine syntaxique, tandis que la dérivation appartiendrait aux domaines morphologique et lexical. En effet, les flexions des noms, des verbes et des adjectifs sont interdépendantes et doivent se répondre, de manière encore plus prégnante dans les langues casuelles, tandis qu'une formation dérivée n'a pas cette contrainte (dans le processus de sa formation). Cependant, là encore, les frontières sont floues,

puisque par exemple les flexions nominales en position sujet sont libres a priori, avant l'insertion en discours. À l'inverse, certaines préfixations imposent un choix syntaxique (exemple du néerlandais *be-* qui ne peut créer que des verbes transitifs) ;

— **Ordre des morphèmes** : généralement, les flexions sont périphériques aux constructions dérivées (*anti-dé-matéri-al-is-ation-s, dé-fais-ait*), ce qui tendrait à montrer que les dérivations portent exclusivement sur la base lexicale, tandis que les flexions ont un rôle syntaxique complémentaire qui ne peut s'appliquer qu'après construction de la (nouvelle) base lexicale.(Greenberg, 1963, p.93) a établi ce principe comme un universel des langues : « if both the derivation and the inflection follow the root, or they both precede the root, the derivation is always between the root and the inflection ». Cet ordre des éléments trouve une justification psycholinguistique : la construction se fait d'abord sur une base lexicale, puis les paradigmes ayant une valeur syntaxique sont mis en périphérie, avec un ordonnancement entre les paradigmes à valeur plus lexicales (nombre, genre) et ceux à valeur plus syntaxiques (les marqueurs de cas sont toujours en dernière position) ; (Booij, 1994, 1996) fait une distinction complémentaire en proposant les notions de flexion interne (ou inhérente) pour désigner les paradigmes flexionnels portant sur le sémantisme de la lexie (par exemple le nombre, le genre pour les noms), et de flexion externe (ou contextuelle) pour désigner les paradigmes qui sont liés au contexte syntaxique (le cas pour les noms, le genre et le nombre pour les adjectifs, la personne et le nombre pour les verbes). Dans ce cadre, les paradigmes dérivationnels s'appliquent d'abord, puis la flexion interne, puis la flexion externe ;

— **Les flexions concernent essentiellement les verbes, les dérivations les noms et les adjectifs** : une autre tendance prototypique concerne le lien entre flexion et verbe, dérivation et nom-adjectif. En effet, la flexion est surtout visible dans les verbes, alors que les dérivations concernent essentiellement les noms et, dans une moindre mesure, les adjectifs. Cette distinction n'est évidemment pas absolue, mais on peut rapprocher ce fait des liens flexion-syntaxe, dérivation-lexique.

Comme on le voit, il faut manipuler ces critères en ayant à l'esprit qu'ils permettent d'identifier des propriétés prototypiques, et ne permettent en aucun cas de tracer une distinction absolue entre les deux phénomènes. On retrouve une situation similaire à l'autre bout des mécanismes de formation des mots, avec la distinction entre composition et constructions syntaxiques formatrices d'unités polylexicales.

5.1.1.3 Composition et unités polylexicales

Parmi les procédés de formation lexicale disponibles dans les langues, deux procédés sont en continuité, la composition et la création d'unités polylexicales. La création d'unités polylexicales est un objet d'études depuis de nombreuses années, et constitue l'un

des procédés les plus productifs de création lexicale [2] Il s'agira ici pour nous de tracer les frontières et les points de contacts entre le procédé de formation par composition et celui par construction lexico-syntaxique. Les points de contacts sont évidents : très souvent, une formation composée peut être glosée par une unité polylexicale (exemples), et il existe de nombreux cas de concurrence entre un composé et une unité polylexicale (exemples). En diachronie, de très nombreux composés sont le résultat de la réduction d'une unité polylexicale (exemples). (Hermann, 1886) pointait déjà sur la difficulté à distinguer les deux procédés et (Hatzfeld et Darmesteter, 1890) considérait la composition comme un type de création d'unités polylexicales.

Plusieurs critères ont été avancés pour distinguer les deux procédés :

— **intégrité lexicale (Lexical Integrity Principle)** (Booij, 2009): cette notion correspond au fait que dans un composé, d'une part s'applique le principe de non-interruptabilité (*non-interruptibility*) et d'autre part, les règles syntaxiques ne peuvent pas manipuler les composants, mais seulement l'ensemble, ce qui n'est pas toujours le cas pour les unités polylexicales [3]. La non-interruptabilité se traduit par l'impossibilité de déplacer les constituants entre eux, ce qui permet de différencier les verbes à particule (des expressions composées) de formations composées à partir de verbes (*to undertake*). La seconde règle s'applique aux seules unités polylexicales nominales et adjectivales totalement figées ;

— **orthographe** : le second critère ressortit à l'unité orthographique (graphique) des composés, contrairement aux polylexicaux, dont les composants sont (au moins graphiquement) distingués. Cependant, ici les usages varient souvent, montrant la proximité des deux procédés (*arbre feuille > arbre-feuille, bien vieillir > bien-vieillir, mot dièse > mot-dièse*). D'un point de vue diachronique, il pourrait s'agir ici de deux phases vers la lexicalisation. Booij donne plusieurs exemples en néerlandais où l'usage veut que les composés ne comprennent aucun espace, mais où l'usage en ajoute pour la clarté du composé.

En dépit de leurs différences, les deux procédés sont donc relativement proches, et beaucoup d'auteurs considèrent qu'il existe une filiation diachronique entre les deux procédés, étant donné que dans beaucoup de cas, les composés sont le résultat d'une réduction des unités lexicales correspondantes. En français, les types de formations composés les plus fréquentes sont, en dehors de la formation N-N, toutes des formations syntaxiques valides (ADJ-N, N-ADJ, V-N ou ADJ) ce qui renforce l'idée d'une continuité entre les deux procédés de création lexicale. Même si la création par unité polylexicale ne fait pas partie de la morphologie, elle fait néanmoins partie de la néologie formelle, pour autant

2. De façon surprenante, aucune évaluation quantitative fiable n'existe sur le nombre d'unités polylexicales par langue. (Mel'čuk, 2011) estimait leur nombre à au moins deux fois le nombre d'unités lexicales simples.

3. les deux principes sont tirées d'une étude visant à établir des universaux dans le domaine morphologique : « We may distinguish the following two (related) types of formal universal constraints in the domain of morphology: (i) constraints on the kind of relations that are possible between syntax and word structure, and (ii) constraints on the accessibility of the internal structure of complex words for modules of the grammar such as the syntax and semantics »(Booij, 2009, p.2-3). Le principe a été d'abord énoncé par (Anderson 1992:84) : « The syntax neither manipulates nor has access to the internal structure of word »

qu'une forme nouvelle soit créée.

5.1.2 Typologie des néologismes formels

Nous évoquons maintenant, dans le champ ainsi circonscrit de la morphologie dérivationnelle et compositionnelle, les typologies proposées pour spécifier le champ de la néologie formelle. Comme nous venons de le voir, l'une des sources d'information vient des travaux en morphologie (formation des mots, morphologie constructionnelle, morphologie productive sont des dénominations régulièrement employées pour désigner ce chapitre de la morphologie). Nous y ferons donc d'abord référence, avant de présenter les typologies proposées par les chercheurs s'intéressant spécifiquement à la néologie.

5.1.2.1 Grammaires traditionnelles : dérivation et composition

Les grammaires traditionnelles ont très tôt établies une typologie des mécanismes de formation lexicale, distinguant la dérivation, construite sur la base d'affixes, la composition, construite sur la base de lexies, et d'autres phénomènes comme la troncation et les conversions. Les emprunts (lexicaux) ne sont évidemment pas traités, puisqu'ils proviennent d'un système externe.

Durant la période de domination des grammaires transformationnelles et génératives, un intérêt poussé a été porté à la morphologie et à la syntaxe, permettant l'émergence de théories linguistiques, qui seront ensuite renouvelées avec le développement des grammaires cognitives et de constructions [4]. Plus récemment, avec le développement d'études descriptives multilingues de grande envergure, on trouve des typologies des mécanismes de formation lexicale tendant à l'universalité ((Booij, 2010; Štekauer *et al.*, 2012; Schmid, 2015b; Hippisley et Stump, 2017).

dérivation et composition, morphèmes libres et liés : les quatre ouvrages reprennent la distinction entre morphème lié (affixes) et morphème libre (lexies) pour distinguer entre dérivation et composition. Ces deux mécanismes *de construction* s'opposent aux mécanismes ne faisant pas appel à des combinaisons de lexies et/ou morphèmes (conversion, accent tonique, transformation interne), ou faisant appel à des opérations de *soustraction* (*substractive word-formation*), notamment la troncation et les différents types d'abréviation. On trouve dans le troisième ouvrage un graphique rendant compte de cette hiérarchie de procédés de formation (voir figure 5.1), avec quelques aménagements (notamment la conversion est considérée comme un procédé "morphologique", mais pas la mot-valisation (*blending*).

Les auteurs spécifient assez précisèment les sous-types de dérivation et de composition. Ainsi, pour la dérivation, on peut distinguer la préfixation, la suffixation (les deux mécanismes les plus répandus dans toutes les langues), l'infixation, la circumfixation et d'autres combinaisons des trois procédés primaires. S'y ajoute la dérivation inverse (*backformation*). Parmi les compositions, une distinction est faite entre la composition simple (combinaison de deux lexies), l'amalgame (*blending, portemanteau*, consistant à

4. Il ne s'agit pas ici de faire un panorama de ces théories. On en trouvera une synthèse dans les ouvrages cités plus bas.

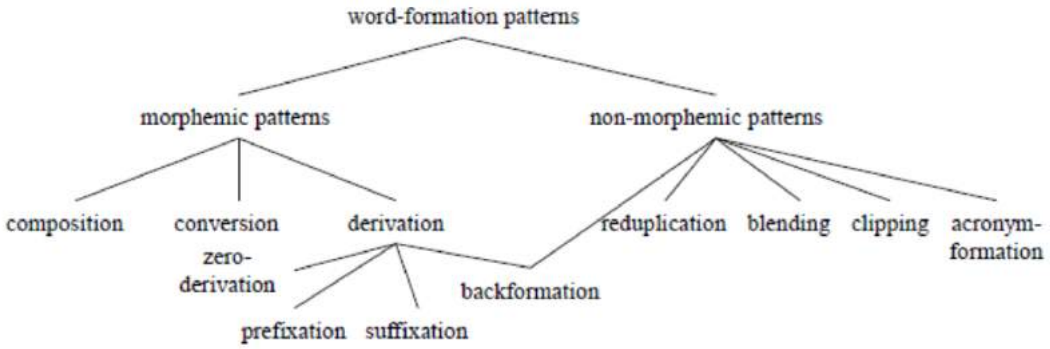


FIGURE 5.1 – Procédes de formation des mots (Schmid, 2015b)

combiner une partie d'une lexie avec une autre (avec ou sans partie commune) et la reduplication (consistant à dupliquer une même lexie : *hush-hush*).

(Booij, 2010), à la suite de (Bloomfield, 1926) distingue les composés endogènes et exogènes : dans les premiers, il existe une dépendance syntaxique entre les deux composants et l'un des composants constitue la tête donnant au composé ses caractéristiques morphosyntaxiques (en français, cela concerne les constructions N-ADJ et N-N, plus rarement ADJ-N : dans *vente-hommage*, le nom *vente* joue le rôle de tête, le second nom jouant le rôle de complément ; *vente* donne à l'ensemble nominal créé son nombre et son genre) ; dans les seconds, aucun des composants ne devient la tête syntaxique, qui doit être sémantiquement générée (cela correspond en français à des composés nominaux construits sur des bases V-N : dans *porte-drapeau* aucun des éléments n'est la tête, le sens correspond à la personne (ici) ou l'entité qui effectue l'action dénotée ; *pickpocket*, en anglais, a le même fonctionnement). Il faut ajouter également les composés coordonnées (en français schémas ADJ-ADJ, comme *socio-démocratique* et N-N, comme dans *film-documentaire*).

(Schmid, 2015b) considère en outre que les typologies peuvent encore être affinées en utilisant une série de six paramètres : les propriétés morphologiques des constituants (de la base et affixe), les propriétés sémantiques générées par la formation, les propriétés phonologiques et (ortho)graphiques générées par la formation (changement d'accent, adaptation phonologique, etc.), l'ordre des constituants, les restrictions liées aux constituants, et la productivité de la formation. Nous reviendrons sur ce dernier paramètre, qui constitue l'une des propriétés essentielles des formations par construction.

Les travaux en morphologie ne s'intéressent qu'aux procédés primaires, mais ne tiennent généralement pas compte de formations par combinaison de mécanismes. D'autre part, elles restreignent le champ d'étude aux mécanismes internes à la langue, sans tenir compte du mécanisme d'emprunt lexical, qui rentre pourtant aussi dans le domaine de la néologie formelle. Afin de traiter ces cas, il faut se tourner vers les propositions des chercheurs qui ont pris comme objet la néologie et pas seulement la morphologie.

5.1.2.2 Typologie des matrices néologiques de Jean-François Sablayrolles

Dans le domaine de la néologie, les premières typologies focalisant sur les « mots nouveaux »sont dûes à (Carnoy, 1927; Stern, 1931), qui articulent leur typologie autour de l'opposition onomasiolologique (des sens aux mots pour les exprimer) / sémasiologique (des mots aux sens qu'ils expriment). Les distinctions proposées par les auteurs reprennent différentes distinctions classiques, notamment concernant les mécanismes de 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 travail, nous partirons d'une typologie plus récente, dont l'objectif est de mettre au jour les *matrices lexicogéniques*, c'est-à-dire les procédés élémentaires de formation de lexies (Sablayrolles, 2000b; Sablayrolles et Pruvost, 2016). Ce dernier propose une typologie des procédés de formation des néologismes selon différents critères (voir figure 5.2).

Fortement inspiré par les travaux de Tournier (Tournier, 1985; Tournier, 1991), il s'agit de décrire les différentes dimensions permettant d'identifier les mécanismes de formation des néologismes. Il distingue tout d'abord les matrices internes à la langue d'une matrice externe, qui permet de rendre compte des phénomènes d'emprunts. Dans le premier groupe, il distingue plusieurs catégories, selon les effets sur la forme et le sens.

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 dans les grammaires ; par imitation et déformation, opérant principalement sur des éléments phonologiques.

Parmi les **procédés de dérivation**, on retrouve les mécanismes primaires de préfixation et suffixation ; l'auteur ajoute la dérivation inverse (*prester* est construit à partir de *prestation ou prestataire* par réduction à la racine, puis la formation verbale est générée par ajout de la flexion infinitive) et la formation par parasynthèse, dans laquelle s'appliquent simultanément une préfixation et une suffixation.

Pour ce qui concerne la **composition**, une première subdivision oppose les compositions simples de lexies, de celles qui opèrent également une réduction de l'une ou des deux formes (amalgame). Parmi les compositions simples, (Sablayrolles, 2000b) distingue les compositions proprement dites des synapsies, qui sont des unités polylexicales, ainsi que la composition savante (sur la base de formants savants ici du grec et du latin) et la composition hybride (mêlant formants savants et lexies). Parmi les compositions par amalgame, il faut distinguer quatre cas : la fractocomposition consiste à utiliser non pas deux lexies simples, une forme tronquée, correspondant grosso modo au radical de la lexie, de l'une ou de l'autre (ici *télé*, pour *télévision*) ; la compocation est une forme plus complexe, puisque dans ce cas, la réduction est simple ou double, et l'un et/ou l'autre des éléments est tronqué, mais non pas réduit à une forme de son radical (*mobylette*

MATRICE INTERNE	morpho-sémantiques	construction	Affixation	**préfixation**	détatouer
				suffixation	statuesque
				dérivation inverse	prester
				parasynthétique ?	désidéologisé ?
			flexion		ils closirent, la représaille
			Compo-sition	**composition** **synapsie** **composition savante** **hybride**	voiture-bélier lanceur d'alerte batracianophile e-commerce, aquacinéaste
			Compo-sition par amalgame	**fracto-composition**	téléspectateur
				compocation	mobinaute, dircab
				factorisation	optipessimiste
				mot-valise	peopolitique
		imitation et déformation		**onomatopée** **f coupe ou paronymie**	dzoing la nesthésie, infractus
	syntactico-sémantiques	changement de fonction		**conversion**	la glisse, la gagne
				conversion verticale	de rejuvénation, un ex
				déflexivation	le boire, le manger
				combinatoire syntax° / lexicale	ironiser un texte encourir la liberté
		changement de sens		**métaphore**	souris (inform.)
				métonymie	sac à dos 'touriste'
				autres figures	escorteuse 'call girl'
	morpho-logiques	réduction de la forme		**troncation**	blème, petit déj
				siglaison /acronyme	LMD, ECUE
	phraséologique	pragmatico-sémantique		**détournement**	faire marcher la planche à promesses
		création		**création**	ne pas faire du huit megabits
matrice externe				**emprunt**	break, cool fioul, redingote

FIGURE 5.2 – Matrices lexicogéniques (Sablayrolles et Pruvost, 2016)

+ internaute > mobi- + -naute) ; la factorisation consiste à prendre une (ou plusieurs) des syllabes communes aux deux lexies (la partie factorisée), et à construire le composé (*optimiste > opti- + pessimiste > pessimiste » optipessimiste*, où -miste est commun aux deux lexies) ; enfin, le mot-valise consiste à joindre deux lexies dont la fin de la première est commune avec le début de la seconde, de sorte que la première lexie est phonologiquement complète (mais pas nécessairement la forme orthographique : *people + politique > peopolitique*).

En ce qui concerne les mécanismes d'imitation et de déformation, l'auteur distingue les onomatopées et les paronymies.

les néologismes phraséologiques : ils opèrent au niveau d'une séquence de lexies, emportent également une création de sens, et comprennent deux sous-types, la création, auquel cas l'unité polylexicale est nouvelle (à rapprocher de la synapsie), et le détourne-

ment, auquel cas l'unité polylexicale se base sur une autre unité polylexicale pré-existante (*faire marcher la planche à billets* > *faire marcher la planche à promesses*) ;

les mécanismes syntactico-sémantiques : ils opèrent une modification au niveau morphosyntaxique ou syntaxique, et s'accompagnent soit d'un changement de fonction soit d'un changement de sens. Sabalyrolles regroupe dans les changements de fonction les phénomènes de conversion (on part d'une forme attestée dans une certaine partie du discours, par exemple une forme verbale, *gagne(r)* pour générer une forme dans une autre partie du discours (*la gagne*), la conversion verticale (auquel cas aucune modification de forme ne se produit, seulement un changement de partie du discours), la déflexivation (sur la base du verbe infinitif, nominalisé en français par adjonction d'un déterminant) et la combinatoire lexicale (ou syntaxique) consistant, pour un verbe, à utiliser un argument nominal ou adjectival non prototypique, générant un changement de sens. Pour ce qui concerne les changements de sens, il s'agit du domaine de la néologie sémantique proprement dite, avec les phénomènes de métaphore, de métonymie, et d'autres figures, l'auteur excluant les extensions et restrictions de sens qui selon lui ne rentrent pas dans le domaine de la néologie proprement dite, qui implique une rupture par rapport à un état de langue précédent. Dans ces derniers cas, il y aurait une modification insensible. Ces deux procédés étaient inclus dans (Sablayrolles, 2000b).

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

5.1.2.3 Remarques sur la typologie de Jean-François Sablayrolles

Dans cette typologie, un certain nombre de points peuvent être discutés :

— **existence de mécanismes purement morphologiques** : on peut en effet se demander si troncation et abréviation emportent ou non, en plus d'une création de forme, une création de sens nouveau. En effet, en considérant que le sens ne se réduit pas à la dénotation, mais inclut également la connotation, la perspectivation ainsi que d'autres paramètres (registre de langue notamment, mais également restrictions diverses d'emploi), il y a bien une différence entre *ado* et *adolescent*, la connotation étant plus marquée pour le premier, et restreinte à un registre de langue. Pour les abréviations, il est vrai que la différence de sens est beaucoup moins évidente ;

— Une des conséquences de cette prise de position amène l'auteur à inclure un certain nombre de procédés (déformation et imitation) dans une catégorie distincte (celle correspondant aux procédés morpho-sémantiques), alors qu'il semble exister une parenté entre la paronymie et les phénomènes de troncation, toutes deux jouant sur la forme et impliquant un changement d'une partie du sens. À l'évidence, le phénomène d'onomatopée doit par contre être isolé, puisqu'il consiste

à imiter un élément référentiel par transcription linguistique du son produit. Il n'y a donc aucune modification d'une forme préexistante. Par contre, les phénomènes de transformation réglée (comme le verlan, le javanais, le loucherbem) qui consistent là encore à transformer la forme, le processus aboutissant à un effet de sens non négligeable, sont bien placés dans les procédés morpho-sémantiques. La reduplication, même si elle est rarissime en français, devrait y être également incluse ;

— Le procédé de changement de fonction est à la limite entre la morphologie et la syntaxe, et peut-être serait-il judicieux de créer un procédé de type morphosyntaxico-sémantique, en laissant à part la combinatoire lexicale, qui, elle, est effectivement purement syntactico-sémantique.

— un dernier point qui mérite un développement concerne les distinctions entre formants savants et formants populaires (composition savante versus composition hybride), ainsi que les notions de fractolexèmes (fractocompositions). Nous y reviendrons dans une section à suivre sur les unités linguistiques dans la néologie formelle.

Au final, la typologie proposée permet d'identifier les différents cas de néologie formelle, c'est-à-dire, dans une perspective de traitement automatique, les néologismes qui emportent la création d'une forme nouvelle lexicale : les procédés morphosémantiques et les procédés morphologiques sont les seuls concernés. Mais il faut également y ajouter un certain nombre de cas couverts par la matrice externe.

5.1.2.4 Typologie de la matrice externe : les emprunts

Pour ce qui concerne la matrice externe, sa première caractéristique est de constituer un groupe à part parmi les mécanismes néologiques, puisqu'il provient d'un système linguistique externe à la langue dans laquelle il se produit. Il s'agit d'un des effets du contact et des échanges entre les langues. Ces contacts aboutissent à plusieurs phénomènes que nous pouvons analyser selon au moins deux points de vue : point de vue des mécanismes formateurs d'emprunts, permettant de proposer une typologie des emprunts ; point de vue du cycle de vie des emprunts, qui permettent de préciser le périmètre des emprunts proprement dits.

Du point de vue des formes linguistiques de l'emprunt, (Sablayrolles et Pruvost, 2016) distingue trois familles de phénomènes induits par les contacts de langue : les emprunts lexicaux véritables, les créations d'équivalents (généralement proposés par les organismes institutionnels de régulation linguistique) et d'autres créations subissant une influence étrangère mais où l'une des matrices internes prédomine. Les emprunts lexicaux proprement dit comprennent les lexies dont le signifiant (avec ou non adaptation phonique et/ou graphique : *staff, lobby, mildiou* < *mildew, paquebot* < *packet boat*) et/ou le signifié (*réaliser* dans le sens de l'anglais *to realize*, /comprendre/) sont directement importés dans la langue réceptrice. Ces vrais emprunts s'opposent à toute une série de cas où une matrice interne prédomine, avec une influence étrangère pouvant prendre différentes formes :

— **la traduction**, lorsque un formant existant dans la langue réceptrice est utilisé

en place du formant étranger, l'influence ressortissant alors au sens, qui est généralement intégralement repris (par exemple *souris* < *mouse*, où c'est l'apparition d'une nouvelle acception de souris, par la matrice sémantique par métaphore, qui prédomine) ;

— **le calque morphologique** : lorsque la traduction opère sur un mot complexe, on parle alors de calque morphologique (exemples : *gratte-ciel* < *skyscraper, cheval de Troie* < *Trojan Horse*) ;

— **l'allogénisme ou faux emprunt** : « Sous les dénominations faux emprunts et allogénismes sont rangées des lexies qui n'existent pas dans la langue censée être la source, mais qui sont fabriquées dans la langue cible avec des formants issus d'une autre langue (tennisman par exemple), en particulier les hybrides : *serial menteur* ou les influences de structure : *royale attitude.* » (Sablayrolles 2016 : 5) ;

— **la synthèse néologique (ou trou comblé)**: il s'agit des cas où la création lexicale ne peut être expliquée que par l'influence d'un sens provenant d'une autre langue, sans que l'on retrouve de lien entre les formes (exemple de la synapsie *lanceur d'alerte* < *whistleblower*).

Nous renvoyons à (Sablayrolles et Pruvost, 2016) pour une présentation détaillée de cette typologie.

Dans le monde anglo-saxon et germanique, la typologie de (Haugen, 1950) reste la base des typologies (voir (Backus *et al.*, 2009; Winter-Froemel, 2009) pour une revue). Nous présentons ici (tableau 5.1) une typologie inspirée de (Loubier, 2011), croisée avec celle de (Furiassi *et al.*, 2012, p.5-12).

Tout d'abord, selon la couche linguistique concernée, on distingue les emprunts phonologiques, lexicaux [5], syntaxiques (Loubier, 2011, p.11-16). Etant donné le moindre impact des emprunts lexicaux, ils sont généralement les plus fréquents, lorsque les deux langues sont dans une situation d'autonomie réciproque (situation d'adstrat). Parmi les emprunts lexicaux, on distingue trois catégories : les emprunts intégraux (importation, ou direct borrowings) : emprunts de forme et de sens, adaptés ou non : *staff, lobby, mildiou (<mildew)* ; les emprunts partiels (substitution ou indirect borrowing) : emprunts de forme ou de sens ; les hybrides (hybrids) (*top-niveau* < *top-level*) et les faux emprunts (emprunts de forme non attestée dans la langue d'origine : *tennisman, brushing* ; emprunts de forme existante mais sens différent : *slip, pin's*).

Parmi les emprunts partiels, ou calques, il faut distinguer : le calque phonologique (forme *ing* avec nasale vélaire [ŋ]) ; le calque morphologique (loan translation et loan rendition) : emprunts de sens, avec traduction littérale de la forme : *supermarket* > *supermarché, sky-scraper* > *gratte-ciel, e-mail* > *courrier électronique* ; le calque syntaxique, lorsque l'emprunt de sens s'accompagne de la traduction littérale d'une expression figurée (*to travel light* > *voyager léger, against the watch* > *contre la montre*) ou encore de l'import d'une structure syntaxique (*introduce someone* > *introduire qn*) ; enfin le calque sémantique (semantic loan), emprunt de sens, avec réutilisation d'une forme existante : *butterfly* > *papillon.*

Au final, nous pourrions proposer une définition générique de l'emprunt (ou trans-

5. Il faut entendre lexical dans un sens large, couvrant les éléments phraséologiques.

Catégorie	**Sous-catégorie**	**Définition succincte**	**Exemples**	
Emprunt intégral		emprunt de forme (adaptée ou non) et de sens	staff, lobby, mildiou (<mildew)	
Emprunt partiel (ou calque)		Emprunt de forme ou de sens		
	Calque phonologique	Emprunt d'un segment phonologique adapté	-ing	
	Calque morphologique	Emprunt de sens, avec traduction littérale de la forme	supermarket >supermarché ; sky-scraper >gratte-ciel ; e-mail >courrier électronique	
	Calque syntaxique	Emprunt de sens, avec reproduction de la construction syntaxique	introduce someone >introduire qn ; against the watch >contre la montre ; to travel light >voyager léger	
	Calque sémantique	Emprunt de sens, avec réutilisation d'une forme existante	butterfly >papillon ; hashtag >mot-dièse	
Emprunt hybride		Emprunt d'un des composants formels, l'autre appartenant à la langue emprunteuse	top-niveau <top-level	
Faux-emprunt		Emprunt de forme non attestée dans la langue d'origine	tennisman, brushing	
		emprunt de forme avec sens différent	slip, pin's	

TABLE 5.1 – Proposition de typologie des emprunts

fert ?), en adaptant la définition d'anglicisme proposée par (Gottlieb, 2005, p.163) : « any individual or systemic language feature adapted or adopted from English [a donor language], or inspired or boosted by English [donor language] models, used in intralingual communication in a language other than English [the recipient language]. » Soit : « tout transfert, réel ou imaginé, d'une propriété linguistique, ou d'une combinaison de propriétés constitutives d'une unité linguistique, provenant d'une langue donneuse et utilisée dans une langue récipiendaire ». Notons dès à présent que pour notre étude applicative sur la néologie formelle, ni les calques syntaxiques, ni les calques sémantiques ne seront évoqués, puisqu'ils utilisent un matériau formel existant de la langue récipiendaire.

Du point de vue du cycle de vie des emprunts, plusieurs notions ont été proposées, tentant de couvrir les différentes phases saillantes de ce type d'innovation lexicale :

— le **xénisme** 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 soit une paraphrase descriptive, soit une note explicative en bas de page quand il s'agit d'un texte écrit » (Guilbert, 1971, p.92). Il

est très proche de l'alternance codique (*code switching*), puisqu'il consiste à citer une réalité propre à une langue L1 dans la langue L2. Restreint aux éléments lexicaux, la mention de la lexie est assumée comme appartenant à la mémoire linguistique de L1 : quelques lexies dorénavant ressenties comme empruntées (ou plus ressenties comme telles) sont passées par l'état de xénisme : *sushi, glasnost, jazz*, etc. Cependant, généralement, le xénisme désigne une réalité culturelle propre à L1, ce qui bloque son intégration plus avant dans la langue L2. Par exemple, un très grand nombre de plats locaux sont désignés par leur nom d'origine, adapté ou non. Le *feiao tropeiro*, plat typique brésilien à base de haricot, de lardons et de saucisses est un exemple de terme qui sans doute restera à l'état de xénisme, pour des raisons à la fois culturelles et linguistiques ;

— le **pérégrinisme** désigne l'état d'un xénisme qui diffuse dans plusieurs couches sociales sans toutefois se fixer ; xénismes et pérégrinismes constitueraient dans ce cadre des états du phénomène d'emprunt [6] ;

— l'emprunt proprement dit, dans ce cadre, désignerait un état d'intégration encore plus avancé : « Il y a emprunt linguistique quand un parler A utilise et finit par intégrer une unité ou un trait linguistique qui existait précédemment dans un parler B (dit langue source) et que A ne possédait pas. L'unité ou le trait emprunté sont eux-mêmes appelés emprunts » (Dubois, 1962, p.177). On notera à ce propos que les anglo-saxons utilisent deux termes distincts pour désigner le processus (*borrowing*) et le résultat « lexical » (*loanword*). On remarque ici que la définition inclut un critère d'intégration avancée de la lexie en L2, ce qui fournit un nouveau critère de distinction entre emprunt et xénisme ;

le cycle de vie d'une lexie empruntée ne s'arrête pas au moment de son émergence, ce qui nous fait diverger de la définition précédente, qui limite l'emprunt aux lexies arrivées au stade de l'intégration. Or il faut distinguer les phases d'émergence, de diffusion et de lexicalisation. D'autres paramètres permettent d'identifier les phases d'intégration des emprunts et notamment, d'un point de vue exclusivement linguistique : intégration phonologique (*riding coat* > *redingote*, mais *baby-sitting* > *baby-sitting*), morphologique, intégration dans le système de morphologie productive (*facebook* > *facebookisation, facebookeur, se défacebooker*...).

Enfin, un dernier point concerne l'éventuelle politique linguistique de la langue réceptrice, qui peut chercher à amoindrir une tendance à l'emprunt jugée trop invasive, notamment lorsque le processus est massif ou ressenti comme tel. Il s'agit d'un élément important dans le domaine français qui dispose depuis longtemps d'organismes régula-

6. « 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 » (McMahon, 1994, p.209). Voir aussi (Dubois, 1962, p.512) : « le pérégrinisme renvoie encore à la réalité étrangère [celle du xénisme] mais la connaissance de son sens est supposée partagée par l'interlocuteur. ». (Deroy, 1956, chap.IX) assimile xénisme et pérégrinisme. Concernant le rapport entre le xénisme et le pérégrinisme, (Chadelat et Pergnier, 2000) affirme que «Les pérégrinismes ne sont après tout que des mots voyageurs ou migrateurs considérés du point de vue linguistique, en fonction d'une place hypothétique au sein du système susceptible de les adopter, tandis que les xénismes sont ces mots étrangers considérés du point de vue des locuteurs en fonction de leur forme exotique».

teurs.

5.1.3 Unités linguistiques à la base des opérations de dérivation et composition : affixe, fractolexème, troncats, lexie

Revenons un moment sur ce qui distingue dérivation et composition : tandis que la composition fonctionne à partir de lexèmes, la dérivation fonctionne à partir d'un lexème et d'un ou plusieurs affixes, définis comme des morphèmes liés (ou affixes), par opposition aux morphèmes libres. La distinction entre morphème lié (ou non-autonome) et morphème libre (autonome , c'est-à-dire les lexies) provient de (Bloomfield, 1926) [7] et ce critère prévaut pour distinguer composition et dérivation dans la plupart des travaux depuis une trentaine d'années (Corbin, 1992, p.328), (Booij, 2005a), (Haspelmath et Sims, 2002, p.85), (Fradin, 2003, p.195). Rappelons cependant que (Darmesteter, 1874; Guilbert, 1971) considéraient que la composition appartenait à la syntaxe [8].

Dans certains cas, la frontière n'est en effet pas évidente:

— **Cas de mots grammaticaux à emploi affixal** : la première difficulté ressortit à l'existence de formations incluant des prépositions et/ou adverbes, ayant à la fois la valeur de lexèmes et qui, dans certains emplois, s'apparentent à des préfixes : *avant-guerre, après-ski, surexposition, etc.*. Ces formations sont généralement considérées comme des dérivés, et la préposition/adverbe un affixe, sans doute parce que qu'il s'agit de mots-outils. Mais étant donné leur caractère de lexème, on pourrait également considérer qu'ils génèrent des composés plutôt que des dérivés. Une explication de ce cas limite (qui n'est pas propre au français, voir (Booij, 2005b)) consiste à considérer que l'emploi affixal emporte des propriétés distinctives (accent tonique absent, valeur sémantique distincte, pouvoir de déterminer la catégorie de la production dérivée) (Amiot, 2004b). Le linguiste français insiste par ailleurs sur la notion de "préfixisation" (grammaticalisation), car de nombreux préfixes du français ont été à l'origine des prépositions employées en latin ou grec également comme des prépositions (*anté-, anti-, co-, extra-, hyper-, hypo-, infra-, inter-, pré-, post-, sub-, super-, supra-, trans- et ultra-*), et les plus récents (*arrière(-), avant(-), après(-), contre(-), entre(-), sans(-), sous(-), sur(-)*, etc.) conservent encore leur valeur prépositionnelle/adverbiale, mais seraient

7. « 9. Def. A minimum form is a morpheme ; its meaning a sememe.

Thus a morpheme is a recurrent (meaningful) form which cannot in turn be analyzed into smaller recurrent (meaningful) forms. Hence any unanalyzable word or formative is a morpheme.

10. Def. A form which may be an utterance is free. A form which is not free is bound.

Thus, *book, the man* are free forms ; *-ing* (as in *writing*), *-er* (as in *writer*) are bound forms »(p.155)

8. La composition « forme [. . .] une expression synthétique qui éveille dans l'esprit plus d'idées que n'en présentent les éléments composants pris chacun à part : timbre-poste ne veut pas dire simplement timbre et poste, mais timbre de la poste, timbre pour la poste, et se résout en une périphrase qui met en lumière l'ellipse fondamentale du composé » et « groupe dans une unité simple des idées qui se présenteraient naturellement séparées [. . .]. » (Hatzfeld et Darmesteter, 1890, p.72). Et un peu plus loin : « le mot composé est [...] une proposition en raccourci ». Darmesteter considère qu'il existe trois types de composés : la juxtaposition, où les règles de la syntaxe sont respectées (*pomme de terre, arc-en-ciel, gendarme*), la composition elliptique (*arrière-cour, porte-feuille, timbre-poste*) où la syntaxe n'est plus explicitée, et la composition par particules, c'est-à-dire la préfixation.

dans une phase de "préfixisation" ;

— **Existence d'affixes devenant des morphèmes libres** : de manière inverse, un certain nombre de préfixes sont utilisés comme lexèmes (*ex-, anti-*). Le critère définitoire morphème lié, pour identifier les affixes, ne semble pas donc, à tout le moins, suffisant, ou alors il faut considérer qu'il s'agit d'homonymes, ce qui clairement n'est pas le cas ;

— **Existence de lexèmes employés comme préfixes** : un troisième argument contre une distinction nette entre dérivation et composition ressortit à l'existence de formations construites à partir de lexèmes, qui s'apparentent là encore à des préfixes, étant donné leur productivité. Par exemple : anglais : *-like, -way, -wise, -worthy, -ware, -monger, -wright, -able*,français : *-thon, -phare, -clé, -gate, agro-, socio-, télé-, éco-, cyber-, e- etc.*. Ces morphèmes sont généralement appelés *affixoïdes*, ou bien encore *fractolexèmes*, pour ceux qui sont réduits à leur radical, ou tronqués d'une manière ou d'une autre. Là encore, une perspective diachronique permet de considérer que ces formants sont en phase de grammaticalisation, et qu'il existe donc bien un continuum entre lexème et morphème affixal. Ce qui distinguerait les deux ressortit à leur productivité, d'une part, et à d'autres propriétés généralement associées aux affixes. Un cas protypique est celui de *-thon*, issu de *marathon*, nom commun, pour désigner l'épreuve sportive, apparu peu après les premiers jeux olympiques modernes en 1892, par antonomase de la ville d'où était parti le soldat pour avertir les athéniens de la victoire obtenue contre les perses, qui ensuite a acquis un sens abstrait (*un véritable marathon*) puis a été employé comme avec le fameux *telethon* aux Etats-Unis dans les années 50, donnant ensuite lieu à toutes sortes de formations (*vidéothon, radiothon, blogathon, hackathon*, etc.) ;

— Enfin, si l'on considère les différents types de langue, il faut noter que les langues isolantes, comme le vietnamien ou le chinois, ne disposent pas d'affixes, puisque les unités lexicales sont toutes libres. La formation de mots se produit donc uniquement par composition et, tout en reconnaissant les spécificités prototypiques des lexèmes et des affixes, reconnaître qu'il existe des possibilités de passage des premiers vers les seconds, et peut-être également, en bout de chaîne, des seconds vers les premiers (*anti- et ex-* par exemple) pour les langues synthétiques et flexionnelles.

Pour rendre compte de ces cas, et continuer à adopter la distinction entre dérivation et composition, (Darmesteter, 1874, p.5) considère que les formations composées sont intrinsèquement syntaxiques : "Un mot composé est une proposition en raccourci". (Anderson, 1992), dans une conception appelée *Item-and-Process*, considère que dans le cas de la dérivation, ce sont des règles de formation des mots (*Word Formation Rule*) qui s'appliquent et permettent de générer, à partir d'un lexème, un autre lexème ; au contraire, dans la composition, nous quittons le domaine de la morphologie pour aller dans celui de la syntaxe (*Word Structure Rule*), la composition étant une mise en relation syntagmatique de lexèmes. Cette vision permet de rapprocher la dérivation d'autres procédés de création lexicale, qui appliquent également des règles de formation de mots (tronca-

tion, reduplication, conversion). Chez Anderson, cette conception a pour conséquence que les mots dérivés n'ont plus aucune structure interne (hypothèse de l'A-morphie). Au contraire, les composés conservent une structure interne, qui est atteinte par les règles syntaxiques : dans la composition, il subsiste toujours une relation syntaxique entre les composants, avec une tête et des dépendants (*attrape-nigaud == V - N, chou-fleur == N(tête) - N (modifieur)*).

Ces arguments et ces faits amènent - encore une fois - à envisager un continuum entre les lexèmes et les affixes, l'espace intermédiaire étant occupé par les fractolexèmes (ou affixoïdes). Il existerait, en diachronie, un processus de grammaticalisation qui mènerait pour certains lexèmes à se suffixiser. Un processus inverse de dégrammaticalisation permettrait également de rendre autonomes certains affixes. Il existerait un continuum, par conséquence, entre la composition et la dérivation, qui passerait notamment par les processus de fractocomposition, de compocation et de troncation, puisque ces différentes opérations comprennent un processus de réduction des lexèmes, dans différentes situations. Il reste néanmoins qu'il existe des affixes prototypiques, et donc une opération de dérivation distincte de la composition. Nous pouvons essayer, à partir des cas prototypiques, de définir un faisceau de critères définitoires complétant le critère de non-autonomie :

— **la forme phonologique est généralement réduite** : tout d'abord, on notera qu'il n'existe pas d'affixes de plus de deux syllabes ; dans le cas des lexèmes, évidemment, cette contrainte ne tient pas, même si la majorité des lexèmes ne va pas au-delà de trois syllabes ;
— **le sens des affixes est généralement générique** : de manière parallèle, le sens des affixes est généralement générique, souvent rapproché d'un sens procédural, correspondant à des expressions sémantiques qu'il serait d'ailleurs intéressant de répertorier afin de voir quels sont les champs sémantiques couverts, avec des comparaisons entre langues, et une étude des évolutions diachroniques ; cela rapproche évidemment les affixes des flexions, et une étude plus poussée des domaines sémantiques couverts respectifs pourrait permettre de tracer des frontières nettes ou minimalement de mettre au jour des catégories sémantiques générales pouvant s'appliquer aux lexies ;
— **dans la dérivation, le sens produit est syntaxiquement gouverné par la règle affixale portée par l'affixe ; dans la composition, le sens produit provient de la relation sémantico-syntaxique entre les deux composants** : ce fonctionnement différent entre les deux procédés a pour premier corrolaire l'endocentricité des lexèmes dérivés (Scalise, 1992), puisque le sens du dérivé est lié au sens de la base lexématique (souvent sous forme d'hyperonymes ou hyponymes : *ultrason, superproduction, antimatière* ; un second corrolaire est que le genre du dérivé est toujours déterminé par la base ;
— **Dans la dérivation, l'affixe peut généralement s'appliquer à plusieurs catégories de mots, et peut générer plusieurs catégories de mots** (Amiot, 2004b) ;
— **Lorsque l'affixe a aussi un emploi autonome, l'emploi affixal a généra-**

lement un sens plus générique (Amiot, 2004b) : ce cas couvre les préfixations ayant également une forme prépositionnelle autonome ;

— **La productivité d'un dérivé est lié à la règle portée par l'affixe ; la productivité d'un composé dépend de la règle syntaxique sous-jacente ; la productivité d'un composé avec une lexie spécifique est un cas intermédiaire**. Nous reviendrons sur cette notion dans la prochaine section.

Il paraît donc raisonnable de considérer qu'il existe un continuum entre les procédés de dérivation et ceux de composition, comme d'ailleurs un continuum entre dérivation et flexion, et entre composition et création d'unités polylexicales. Néanmoins, il existe des pôles qui permettent, via un faisceau d'indices convergents, de définir ces différents procédés de formation lexicale. Une étude diachronique montre bien, via les processus de grammaticalisation et plus rarement de dégrammaticalisation, la dynamique qui existe entre ces différents pôles. D'autre part, et de manière parallèle, il existe un continuum entre les lexies et :

— du côté de la réduction des formes et de la généralisation sémantique, les troncats (qui restent libres) les fractolexèmes (ou affixoïdes, qui sont liés), les affixes et même les morphèmes flexionnels ;

— du côté de la construction des formes et de la spécification sémantique, les composés simples, les unités polylexicales et les constructions, au sens des grammaires de construction.

Nous résumons cette analyse dans la figure 5.3 et donnons les propriétés typiques et des exemples de chacun des pôles dans le tableau 5.2 .

5.1.4 Notion de productivité

Un concept important concerne la productivité des règles de formation. Ce concept est important, car finalement la morphologie lexicale traite des mots potentiels et des règles qui permettent de les générer. Chacune de ces règles a donc une productivité. Le concept doit permettre de rendre compte de la capacité de certains lexies et de certains formants à générer d'autres unités par préfixation, suffixation mais aussi par composition (les schèmes *-clé, -phare* par exemple) et par fracto-composition (*e-, cyber, eco-, bio-*). Prenons *e-* et *cyber-*. Ils ne font pas classiquement partie des affixes, mais leur comportement s'en approche de par leur capacité à permettre la génération d'un grand nombre de formations lexicales : *e(-)* est le résultat de la troncation de *electronic*, et a été importé d'abord dans *email*, puis, depuis quelques années, se rencontre dans un nombre grandissant de lexies (*e-gouvernance, e-addictif, e-recruter, e-scooter, e-manif*, etc.), au point de devenir, en français comme dans d'autres langues, un formant emprunté productif.

Le calcul de la productivité peut porter non seulement sur des règles de formation liées à un formant (affixe, fracto-lexème, lexème) comme dans les exemples donnés plus haut, mais peut également s'appliquer de manière plus générale à des schémas syntaxiques ou lexico-syntaxiques, notamment pour ce qui concerne les procédés de composition.

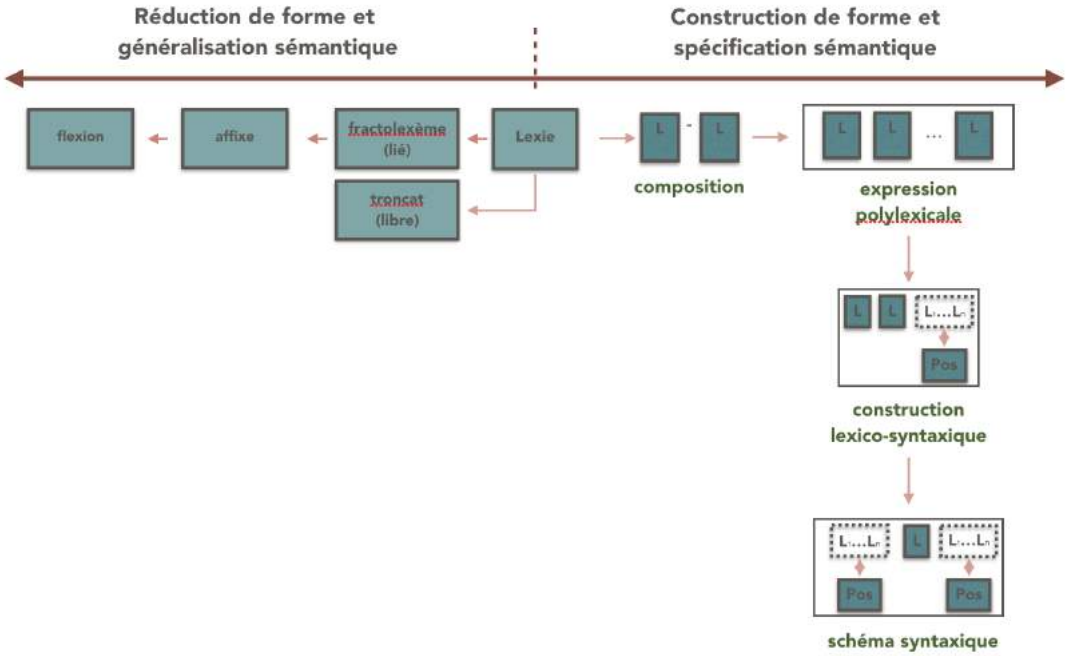


FIGURE 5.3 – Continuum entre lexies et affixes d'un côté, et lexies et constructions syntaxiques, de l'autre

5.1.4.1 Définition générale et mesures

La notion de productivité était présente dès les premiers travaux de morphologie (on en trouve des traces dans la grammaire de Panini, puisqu'il détaille les règles de la morphologie et de la syntaxe), mais a été introduite dans la période moderne par (Schultink, 1961) puis reprise dans tous les travaux en morphologie, allant jusqu'à la création de la dénomination "morphologie productive" comme synonyme de "formation des mots". Pour définir la notion, des approches qualitatives et quantitatives ont été proposées (voir (Dal, 2003; Bauer, 2001) pour une revue complète). Par exemple, (Corbin, 1987, p.177) indique : « 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. » Cette définition, qui sera reprise notamment par (Bauer, 2001), explicite deux aspects de la productivité : tout d'abord, la disponibilité (*availability*), à savoir la capacité à créer de nouvelles formations lexicales à partir d'une règle : lorsqu'un affixe ou un affixoïde ne produit plus de formations lexicales, il n'est plus disponible, et il pourrait s'agit d'un critère définitoire de cette unité ; d'autre part, la rentabilité (*profitability*), à savoir l'extension maximale que peut avoir une règle morphologique, qui peut être définie par les contraintes intrinsèques portées par l'affixe.

Très vite, des mesures ont été proposées pour quantifier la productivité. Les plus

Type d'unité	**Propriétés**	**Exemples**
morphème flexionnel	Formelle : séquence indécomposable courte	-ions, -s, etc.
	Fonctionnelle : pas de catégorie morphosyntaxique, morphème lié	
	Sémantique : dénotation très générale	
	Productivité : générique pour l'ensemble d'une catégorie fonctionnelle	
affixe	Formelle : séquence indécomposable courte	anti-, -isme, etc.
	Fonctionnelle : pas de catégorie morphosyntaxique, morphème lié	
	Sémantique : dénotation très générale	
	Productivité : pour la formation de mots nouveaux par dérivation	
affixoïde (fracto-lexème)	Formelle : séquence indécomposable courte	e-, cyber-, bio-, -gate, -street, etc.
	Fonctionnelle : pas de catégorie morphosyntaxique, morphème lié	
	Sémantique : dénotation spécifique	
	Productivité : pour la formation de mots nouveaux par dérivation (ou composition, selon typologie)	
lexie	Formelle : séquence décomposable en racine et flexion (pour les noms, adjectifs, verbes)	marchons, vents, grandes, bobos, ados, app, etc.
	Fonctionnelle : catégorie morphosyntaxique, morphème libre	
	Sémantique : dénotation spécifique	
	Productivité : non	
lexie composée	formelle : séquence décomposable en racine(s) et flexion	ramasse-miettes, aoutléthisme, peignoir-couteau, etc
	Fonctionnelle : catégorie morphosyntaxique	
	Sémantique : dénotation spécifique	
	Productivité : partiellement, pour certains composés	
unité polylexicale	Formelle : séquence décomposable en racines et flexion(s)	pomme de terre, robe de chambre, ciel menaçant, et
	Fonctionnelle : catégorie morphosyntaxique	
	Sémantique : dénotation spécifique (non totalement compositionnelle)	
	Productivité : non	
construction lexico-syntaxique	Formelle : séquence syntaxique décomposable en lexies (racine+flexion) et paradigmes lexicaux	perdre la boule, perdre la tête, perdre son sang-foid,
	Fonctionnelle : schéma syntaxique	
	Sémantique : dénotation spécifique	
	Productivité : partielle, sur certains éléments	
construction syntaxique	Formelle : schéma syntaxique générique	DET N, DET N ADJ, etc.
	Fonctionnelle : schéma syntaxique	
	sémantique : dénotation générique	
	Productivité : totale pour les parties du discours concernées	

TABLE 5.2 – Propriétés et exemples des différentes unités linguistiques intraproposionnelles

répandues et les mieux acceptées sont celles proposée par (Baayen, 2009; Baayen, 1992; Baayen, 1993) qui distingue trois acceptions de la notion, et trois mesures associées :

— **la productivité réalisée(realized productivity)** désigne les réalisations déjà constatées dans le passé à partir du formant ou de la lexie, et qui peuvent être comptabilisées dans un dictionnaire [9]. La mesure est donc un simple comptage

9. "A first measure of productivity focuses on the size of the morphological category. A category with

des attestations, dans un dictionnaire ou un corpus. Cette mesure était nommée *extent of use* dans (Baayen, 1993). Elle se calcule par la formule

$$V(C, N)$$

où C représente le nombre d'éléments distrincts de la catégorie dont on mesure la productivité (par exemple les formations lexicales en -able), et N le nombre total de types de mots différents dans le corpus ;

— **la productivité en expansion (expanded productivity)** quantifie le nombre d'hapax nouveaux (de néologismes) créés par la règle dans un corpus. La mesure associée est plus difficile à établir, car elle doit se baser sur les nouveaux termes, qui a priori ne sont pas disponibles, et doivent donc être évalués sur un corpus déjà constitué [10]. Baayen propose une mesure en deux étapes : tout d'abord, dans un corpus C, on calcule le nombre total d'hapax legomena (ci-après simplement "hapax") de la catégorie étudiée C dans un corpus comprenant N types de mots différents (V(1,C,N)). On divise alors ce nombre par le nombre total d'hapax dans le corpus (V(1,N)) ; la mesure est alors calculée comme suit

$$P* = V(1, C, N)/V(1, N)$$

. Cette mesure donne alors une idée de la contribution de la catégorie à l'expansion du lexique (également appelée *hapax-conditioned degree of productivity* dans (Baayen, 1992)) ;

— **la productivité potentielle (potential productivity)**, enfin, mesure l'étendue maximale possible de cette productivité [11]. Elle est mesurée en divisant le nombre d'hapax de la règle par le nombre de formations de la même catégorie syntaxique attesté dans le corpus. Cette productivité était nommée *category-based-conditioned degree of productivity* dans (Baayen, 1993). La formule s'exprime ainsi :

$$P = V(1, C, N)/N(C)$$

Cette mesure peut être utilisée pour distinguer différentes règles morphologiques, ou encore pour comparer la productivité d'une règle dans différents corpus.

Ces mesures posent un certain nombre de problèmes méthodologiques et pratiques :

---

many members is more productive in the sense that it has produced many complex words that are useful to the language community. A rule that is highly productive in this sense is like a successful company selling a product that has a large share of the market." (Baayen, 2009, p.6)

10. "A second measure of productivity assesses the rate at which a morphological category is expanding and attracting new members. A category that is expanding at a higher rate is more productive than a category that is expanding at a lower rate, or that is not expanding at all. A rule that is highly productive in this sense is like a company that is expanding on the market (independently of whether that company has or does not have a large share of the market)."(Baayen, 2009, p.7)

11. "A company may have a large share of the market, but if there are hardly any prospective buyers left because the market is saturated, it is nevertheless in danger of going out of business. A third measure of productivity gauges the extent to which the market for a category is saturated. A rule with a low risk of saturation has greater potential for expansion, and hence a greater POTENTIAL PRODUCTIVITY."(Baayen, 2009, p.7)

— d'une part, la première mesure n'est pas fiable dans la pratique, puisqu'il faudrait avoir accès à l'ensemble des corpus effectivement produits par les locuteurs, ce qui n'est pas envisageable ;
— lla mesure de la potentialité n'est pas possible (voir (Bauer, 2001, p.331)), car une mesure doit s'appuyer sur des données déjà produites, et ne peut donc, en tout cas avec la méthode statistique proposée, évaluer la potentialité ; de plus, la limitation à une catégorie morphosyntaxique pour évaluer la potentialité maximum n'est pas valide, car une règle peut-être transcatégorielle, d'une part, et beaucoup d'affixes ont une productivité potentielle limitée également au niveau lexical (EXEMPLES) ;
— la mesure de productivité potentielle doit être revue pour éviter la surestimation des règles à faible fréquence, à règles contraintes.
— la mesure de productivité en expansion : difficulté pour compter l'ensemble des hapax d'un corpus, car il faudrait restreindre aux mêmes catégories. On peut faire une approximation en comptant juste le nombre d'hapax pour un affixe ou affixoïde donné, et, afin de de ne pas pénaliser les règles à faible fréquence, en divisant le nombre d'hapax par le nombre total de formants déjà produits et attestés.

Dans les études applicatives, nous utiliserons généralement la productivité en expansion, mais en modifiant la notion d'hapax, car cette notion n'est plus valide avec les modes de communication actuels : comme nous le verrons (voir chapitre 4), dans la très grande majorité des cas, les formations nouvelles seront répétées un certain nombre de fois sur une période temporelle courte, puis disparaîtront : c'est ainsi que nous définirons l'émergence d'une nouvelle forme, en place de la notion d'hapax. Concernant la délimitation de la productivité potentielle, il nous semble plus adéquat de définir la règle de manière qualitative, en décrivant ses contraintes le plus finement possible.

A noter enfin que la notion de productivité peut s'appliquer également aux constructions lexico-syntaxiques et aux schémas syntaxiques, qui connaissent très souvent des contraintes (par exemple : *arriver à N(lieu) versus arriver à N(désignation d'un processus arrivé à son terme)*).

Il est enfin nécessaire de lier productivité et fréquence : en effet, lorsqu'un procédé est productif, il est productif en exapansion lorsqu'il génère de nouvelles lexies. Mais il arrive très fréquemment, le plus souvent même, que parmi les occurrences produites, certaines aient une grande fréquence. ces dernières occurrences "fixent" des pôles lexicaux qui deviennent alors soit des prototypes pour le procédé, soit des unités indépendants avec un sens particulier, notamment si le sens devient non transparent. Par exemple, *anti-nucléaire* a une grande fréquence, et dans le même temps il fait partie de la dérivation en anti, auquel il répond aux règles. Il devient donc lexicalement indépendant, de par sa fréquence, même si il est totalement dépendant de la règle. En revanche, *antidote*, qui a été formé dès le grec (*antidotos* > *lat. antidotum*) est devenu complétement indépendant de la règle, même si on peut encore inférer de sa forme son origine. De même, *antibiotique*, créé dès 1871 par dérivation de *anti-* sur *gr. biôtikos*, avec au départ un sens large 'opposé à la vie', puis un sens restrictif avec ll'invention du médicament dans les années 40, est

devenu indépendant, mais conserve un lien avec la règle, là très perceptible. Mais le glissement de sens qui n'est plus totalement déductible de la règle de dérivation la rendu (partiellement) indépendant de celle-ci. La présence dans les dictionnaires est l'un des signes de cette indépendance de certaines formations dérivées.

5.1.5 Description formalisée des mécanismes de formation

En dehors de la productivité, il conviendrait de décrire les procédés de formation de manière plus détaillée et formalisée. De ce point de vue, selon les théories morphologiques, différentes approches ont vu le jour : la plus classique, issue des travaux de Bloomfield et repris ensuite par les linguistiques transformationnelles et génératives, considèrent que l'on peut décomposer la formation dérivée en ces constituants (formule (5.1)):

$$anti - macronisme \Rightarrow \underbrace{anti}_{\text{préfixe}} - \underbrace{macron}_{\text{radical}} - \underbrace{isme}_{\text{suffixe}} \tag{5.1}$$

Cette approche (appelée *item-and-arrangement*) permet d'isoler les différents morphèmes, libres et liés, mais ne rend pas compte de l'ordre d'application des règles. D'où une seconde approche (dite *item-and-process*) consistant à décrire le processus itératif de construction des lexèmes dérivés par des règles concaténatives qui, partant de la base lexicale (radical), appliquent successivement les règles morphologiques pour générer d'autres lexèmes, dits dérivés. Ainsi, pour le dérivé *anti-macronismes*, en partant du lexème *Macron, macron*, on peut construire la série de règles de réécriture suivante (formule (5.2)) :

$$\begin{aligned}
&0. \quad Macron \\
&1. \quad macron + isme \Rightarrow macronisme \\
&2. \quad anti + macronisme \Rightarrow anti - macronisme \\
&3. \quad anti - macronisme + s \Rightarrow anti - macronismes
\end{aligned} \tag{5.2}$$

Cette approche, dite aussi guidée par des règles, est très liée au développement de la grammaire générative, et peut encore être plus détaillée. Par exemple, pour décrire les préfixations adjectivales en un-, en anglais, (Aronoff, 1976, p.63) propose le formalisme suivant (formule (5.3)):

$$\begin{aligned}
&[X]_{Adj} \rightarrow [un - [X]_{Adj}]_{Adj} \\
&un - X = notX
\end{aligned} \tag{5.3}$$

où X représente n'importe quelle forme adjectivale : la première ligne décrit la transformation par adjonction du préfixe *un-*, générant une nouvelle lexie également de type adjectival ; la seconde ligne décrit le sens de cette formation.

(Booij, 2010), dans une perspective constructionnelle, propose une représentation des règles de formation encore plus complète (formule(5.4)):

$$\begin{aligned}
&[[X]_{Adj}ness]_{Nom} \quad \text{/the property/state of X/} \\
&[[aware]_{Adj}ness]_{Nom} \quad \text{/the property/state of being aware/}
\end{aligned} \tag{5.4}$$

Le schéma (ligne 1) détaille à la fois les contraintes de la règle (ici les catégories syntaxiques des composants lexématiques), la ou les variables pouvant être instanciées (X) et le sens réglé attaché à la règle de formation. La ligne 2 explicite un exemple, ou encore une instanciation de cette règle. Dans la mémoire collective, les deux informations sont stockées, en gardant à l'esprit que, dans l'esprit des grammaires de construction, c'est la multiplication des instances qui génèrent la règle, et qu'il s'agit toujours d'un état évolutif, tant que la langue est vivante. On voit bien ici par ailleurs que la forme d'une lexie forme porte deux types d'information : une information phonologique et phonographique, d'une part, et une information morphosyntaxique. La troisième information est l'information sémantique. Dans une vision constructionnelle, en outre, les mots complexes, comme les constructions syntaxiques, sont des instantiations de schémas constructionnels (Croft, 2001, Goldberg, 2006:5). Les règles de formation de mots répondent donc parfaitement à ces définitions. On peut définir non seulement les règles de dérivation, mais également les règles de composition, selon le même formalisme (formule(5.5)):

$$\begin{aligned}&[[a]_{Xi}[b]_{Nj}]_{Nk} \Leftrightarrow [SEM_i \text{ avec une relation R avec } SEM_j]_k \\ &[[homme]_N[trompette]_N]_{Nom} \Leftrightarrow [un/homme/quijouedela/trompette/]\end{aligned} \tag{5.5}$$

Dans le présent travail, nous laisserons la description formalisée systématique des différents procédés de formation pour un travail ultérieur.

5.2 Méthodes de repérage automatique de la néologie formelle

Dans cette section, j'évoque les travaux menés pour modéliser et repérer automatiquement en corpus les néologismes de forme.

5.2.1 Méthodes de repérage automatique des néologismes formels

Les méthodes de repérage automatique des néologismes de forme se distinguent par les caractéristiques à prendre en compte pour leur repérage (apparition d'une forme nouvelle, forme interne, cotexte, contexte, etc.), ainsi que par les méthodes utilisées, soit symboliques, soit statistiques/probabilistes et parmi ces dernières on peut distinguer les méthodes supervisées, semi-supervisées et non-supervisées.

L'objectif est ici restreint au repérage de formes nouvelles, non existantes dans un état synchronique donné de la langue.

5.2.1.1 Apparition d'une forme nouvelle

La première approche qui a été mise en place consiste à utiliser une ressource lexicographique de référence pour repérer dans un corpus tous les mots inconnus, puis met en œuvre différents filtres afin d'identifier des candidats néologismes (Cabré et De Yzaguirre, 1995; Cabré *et al.*, 2003; Janssen, 2008; Ollinger et Valette, 2008; Kerremans

et al., 2012; Sagot *et al.*, 2013; Gérard *et al.*, 2014). Cette méthode nécessite un dictionnaire de référence suffisamment couvrant, des ressources liées aux entités nommées, et des algorithmes efficaces de repérage des erreurs typographiques. Elle est donc beaucoup plus complexe à mettre en œuvre qu'il n'y paraît, tout d'abord pour ce qui concerne le dictionnaire d'exclusion (ou dictionnaire de référence) :

— il est **très difficile de disposer d'une ressource lexicographique à jour, quelle que soit la langue considérée** ; pour le français par exemple, plusieurs ressources lexicographiques ont été développées (voir (Grezka *et al.*, 2015) pour une revue complète). Les différentes ressources ont des mérites respectifs : tandis que Morfetik (Grezka *et al.*, 2015) est le dictionnaire actuellement le plus couvrant du point de vue des formes simples, mais il ne dispose d'aucune forme composée ; Le Lefff (Sagot, 2010) dispose d'une couverture moindre, mais intègre un grand nombre de noms propres ; GLAWY (Sajous et Hathout, 2015), enfin, est une ressource issue du Wiktionnaire converti dans un format XML exploitable, qui comprend la plus large couverture lexicographique. Le problème ici réside dans une trop grande couverture, puisque ce dictionnaire comprend un certain nombre de néologismes, non marqués comme tels, ce qui rend difficilement exploitable en l'état la ressource. On peut citer également d'autres ressources, notamment les dictionnaires développés pour les correcteurs orthographiques, qui sont généralement très couvrants (par exemple Hunspell [12]) ;

— Lorsqu'ils sont disponibles, les **analyseurs morphosyntaxiques** par apprentissage supervisé sur corpus de référence (les plus efficaces aujourd'hui) peuvent être utilisés pour détecter des formes nouvelles. C'est ainsi que Treetagger (Schmid, 1995) permet d'annoter spécifiquement les mots inconnus (de lui) dans les textes, pour différentes langues. Cette méthode présente l'avantage de ne pas avoir à construire de ressource linguistique de référence mais dépend de la qualité et de la couverture du corpus d'apprentissage utilisé. Un analyseur spécifique multilingue a par ailleurs été mis en place par (Janssen, 2012a).

— pour les **langues peu dotées en dictionnaires**, une telle ressource lexicographique prise comme corpus d'exclusion n'est pas disponible. Il est alors nécessaire de recourir à d'autres méthodes pour construire une telle ressource. Une solution consiste à récupérer les formes attestées dans un corpus suffisamment large, et à ne prendre en compte que les formes ayant une fréquence minimale supérieure à celle de l'hapax. La difficulté réside là dans le choix du corpus, puisqu'il faut alors en disposer, d'une part, et établir la couverture linguistique visée. D'autre part, cette technique s'avère complexe à mettre en place pour les langues morphologiquement riches, puisqu'aucune lemmatisation n'est disponible. Cependant, avec l'avènement d'analyseurs morphosyntaxiques non ou faiblement supervisés de bonne qualité, et le développement de dictionnaires collaboratifs de type Wiktionnaire, cette méthode de **construction dynamique de la ressource linguistique** paraît aujourd'hui une voix prometteuse.

12. `http://hunspell.github.io`

La méthode par dictionnaire de référence (désormais DREF), très souvent complétée par une analyse morphosyntaxique, reste la méthode privilégiée de tous les systèmes existants (NeoCrawler, Logoscope, Obneo et, nous le verrons, Néoveille).

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

— **les noms propres** : ceux-ci peuvent être repérés : sur des bases purement typographiques (mots capitalisés), mais il faut alors pouvoir distinguer ces capitales de la capitalisation du premier mot de phrase, ainsi que prendre en compte la possibilité de parties de noms propres non capitalisées (exemple : *Pierre du Verger, Saône et Loire, etc.*) ; sur la base d'un dictionnaire de référence, mais cette solution ne peut fonctionner que pour les noms propres les plus usuels, la catégorie des noms propres étant la plus productive dans les langues ; au moyen de règles de formations internes (exemple : Prénom + Mot capitalisé éventuellement répété), ou de règles contextuelles (*la mairie de Bordeaux, le port de Aigues-Mortes, le directeur de Renault, etc.*) . Le repérage des entités nommées a donné lieu à de nombreux travaux en TAL [13], ce qui rend cette tâche accessible ;

— **les erreurs typographiques** : les coquilles, les erreurs d'orthographe sont une source de repérage de mauvais néologismes ; dans ce cadre, il convient d'utiliser les correcteurs orthographiques disponibles (type HUNSPELL [14]), même si les algorithmes de correction orthographique peuvent très souvent aussi tenter de corriger de vrais néologismes (exemple : *biotiful => beautiful*) ;

— **les passages en langue étrangère** : les textes d'entrée peuvent comprendre des citations en langue étrangère, qu'il faut donc repérer et éliminer. Actuellement, les systèmes de détection de langue ne fonctionnent que globalement sur un texte, ce qui rend compte la détection de ces zones ;

— **Le niveau de langue et les normes de discours**: un système de détection devrait pouvoir fonctionner avec des textes répondant à des normes diverses, mais cela nécessite une adaptation de la ressource de référence (d'exclusion),par exemple si l'on souhaite traiter des tweets ; un autre problème concerne les pages internet, qui peuvent contenir des zones dans un langage moins normé (par exemple les commentaires), et qu'il conviendrait d'exclure ou tout au moins de marquer comme tel ;

— **les erreurs issues des étapes précédentes du traitement automatique** : lorsque le corpus provient notamment du web, l'extraction des zones de texte peut être fautive, comme l'étape préalable de segmentation des textes en mots.

5.2.1.2 Forme interne des néologismes de forme

Une autre approche, éventuellement complémentaire de la précédente, cherche à identifier les caractéristiques spécifiques formelles des néologismes de forme. L'hypothèse est que les formes nouvelles ont des caractéristiques spécifiques qui les distinguent des formes existantes. Cela semble par exemple évident pour les emprunts, les troncations ou encore

13. Notamment, le Stanford NER Recognizer (Finkel *et al.*, 2005), `https://nlp.stanford.edu/software/CRF-NER.html`

14. `http://hunspell.github.io/`

les néologismes par compocation. Quelques travaux ont été menés dans ce sens pour le repérage des emprunts : (Kang et Choi, 2002; Alex, 2008; Jacquet-Pfau, 2003) détaillent les consécutions de lettres permettant de repérer dans les textes dans une langue L des « mots » en langue étrangère (spécifiquement anglicismes). Un travail similaire est mené actuellement pour repérer les mots-valises et les troncations (thèse Vinogradova, en cours).

À notre connaissance, aucun système actuel n'effectue aujourd'hui ce type de traitement de manière systématique.

5.2.1.3 Contextes spécifiques des néologismes formels

L'analyse des contextes des néologismes de forme (comme de tous les néologismes) peut avoir des spécificités discriminantes. Notamment, au moment de l'apparition d'une forme nouvelle s'accompagne dans la très grande majorité des cas d'une apparition en mention, ainsi que d'une glose définitoire-explicative (Cartier, 2011b). Cette approche est un moyen complémentaire de repérage des néologismes formels et sémantiques, offrant de plus l'avantage de proposer une définition initiale de la nouvelle forme ou du nouveau sens. A notre connaissance, aucun système n'a été développé dans cette direction.

5.2.2 Méthode(s) de repérage utilisée(s)

Parmi les trois approches présentées ci-dessus, la première est celle qui est actuellement la plus utilisée, et c'est celle que nous avons mis en place dans Néoveille, dans un premier temps. Cependant, nous avons mis en place une architecture permettant d'améliorer au fur et à mesure le système. La figure 5.4 montre que le système automatique de reconnaissance des formes délivre à l'expert linguiste une liste de candidats néologismes, ce dernier devant valider ou invalider les propositions : en cas d'invalidation du candidat, l'expert doit choisir dans quel dictionnaire d'exclusion la lexie devra être versée [15] ; en cas de validation, le candidat sera reversé dans la base des néologismes à étudier plus avant. Le système interactif de validation permet donc d'améliorer les reconnaissances futures, et de construire des jeux de référence pour la mise en oeuvre d'outils d'apprentissage automatique ou Deep Learning ((Lejeune et Cartier, 2017; Cartier, 2018b)).

15. A mi 2017, cinq dictionnaires sont proposés : dictionnaire des formes simples, des formes composées, terminologique, des xénismes et d'exclusion proprement dit.

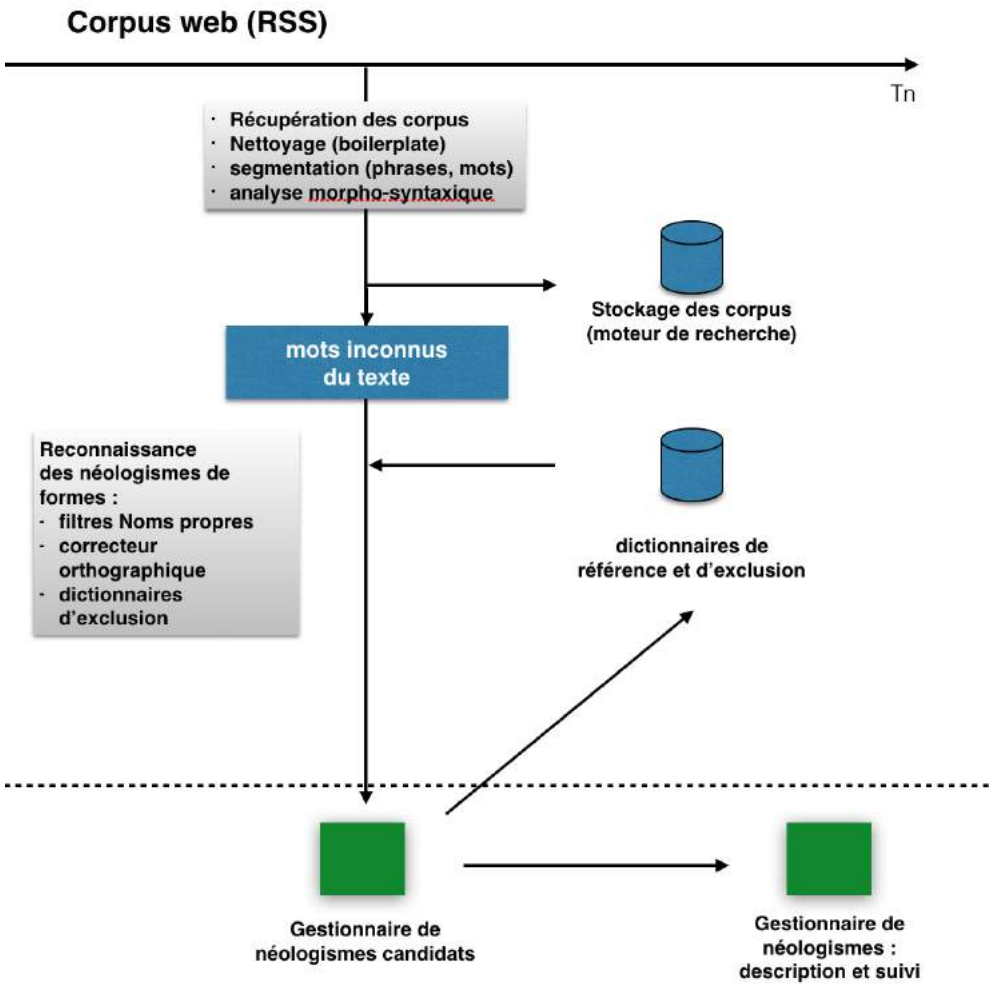


FIGURE 5.4 – Néoveille : reconnaissance des néologismes formels

5.2.3 Évaluation du système de repérage des néologismes de forme

Nous avons mené une évaluation de la qualité de la détection automatique en français, et du temps moyen pour arriver à une F-mesure acceptable pour commencer à travailler sur les néologismes. Concernant la F-mesure, nous l'avons calculé en utilisant les décisions effectuées sur près de 3000 néologismes candidats entre janvier et mars 2016. Les résultats de cette expérimentation sont une F-mesure de 64,7 %, une précision de 54 % et un rappel de 73 %.

La seconde évaluation permettait d'évaluer le temps de mise au point des dictionnaires d'exclusion, dans le cadre de la méthode utilisée. Elle a été menée sur la français et sur l'italien. Dans l'un comme dans l'autre cas, un travail initial sur environ 2000 néologismes candidats permet d'éliminer la très grande majorité des lexies manifestement à exclure. Cela représente environ trois semaines des travail pour deux experts linguistes

à plein temps.

5.2.4 Analyse et perspectives

La méthode par dictionnaire d'exclusion que nous utilisons actuellement a été reprise des travaux antérieurs. Mais elle comporte au moins trois limites :

— Elle nécessite une mise en place initiale d'un dictionnaire d'exclusion avec l'ensemble des formes existantes dans une période temporelle antérieure, qui n'est pas si simple à obtenir. Par exemple, pour le russe, la ressource récupérée (le dictionnaire des formes de Hunspell) n'était pas suffisamment couvrant, ce qui a occasionné beaucoup de travail de validation/invalidation de la part de responsable de cette langue ; pour le chinois, nous avons du nous restreindre aux résultats de l'analyse morphosyntaxique du Treetagger, qui n'a pas une F-mesure suffisante. Il est évidemment possible de créer une telle ressource à partir de corpus, qui sont maintenant disponible, mais cela occasionne un coût non négligeable ;

— même avec un dictionnaire de référence suffisamment couvrant, cette méthode nécessite une phase initiale d'apprentissage itératif dont le coût humain n'est pas négligeable, qui a été évalué à 3 semaines sur le français et l'italien (environ 2000 lexies), soit deux personnes sur dix jours à raison d'environ 100 néologismes par jour ;

— cette méthode n'empêche pas que, même après une phase d'apprentissage itératif à l'aide des validations/invalidations humaines, il reste en moyenne 40% de néologismes candidats encore à exclure.

De ce fait, et étant donné que nous disposions notamment pour le français d'une donnée de référence, nous avons expérimenté l'utilisation des propriétés formelles internes et contextuelles des néologismes pour leur détection automatique par des méthodes d'apprentissage automatique. Les résultats (sur 20 000 néologismes candidats, avec trois catégories : dictionnaire, lexies exclues (souvent coquilles fréquentes) et néologismes), permettant d'obtenir une précision de 80% pour le français (méthode SVM : (Lejeune et Cartier, 2017)). Nous obtenons même une précision supérieure à 90% avec un réseau de neurones ANN à deux couches, en n'utilisant que les propriétés internes des lexies (Cartier, 2018b) (voir matrice de confusion figure 5.5).

Cette voie semble donc prometteuse, et c'est l'un des travaux qui seront menés à court terme.

5.3 Conclusion prospective

Ce long chapitre est divisé en deux parties : l'une qui cherche à modéliser la néologie formelle, spécialement les procédés de dérivation et de composition, et l'autre qui s'intéresse aux méthodes de détection automatique des néologismes formels.

Dans la première partie, nous avons entrepris de modéliser les procédés propres à la néologie formelle. Cela concerne la dérivation, la composition, les procédés de troncation et de transformation et les emprunts lexicaux. Nous avons tenté de détailler les différents

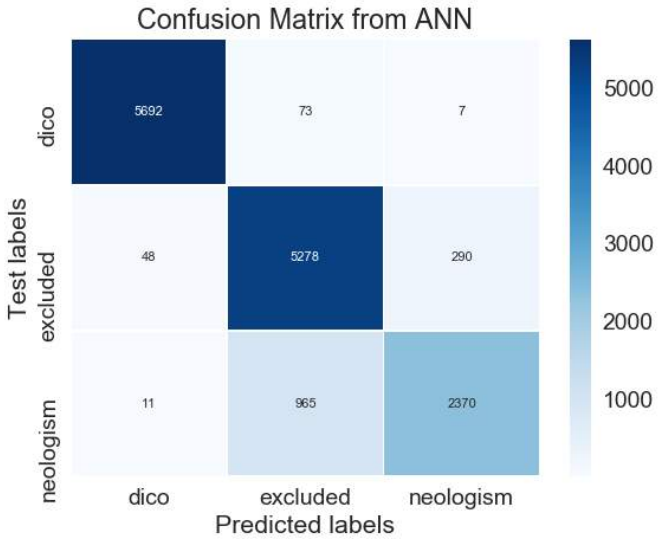


FIGURE 5.5 – Matrice de confusion avec un ANN (Cartier, 2018b)

procédés et d'en établir les propriétés. Dans une première section, nous avons tracé les limites entre flexion, dérivation, composition et synapsie à l'aide de propriétés typiques, et explicité la porosité des frontières entre ces procédés. Dans une seconde section, nous avons abordé les typologies des néologismes formels, en passant en revue les propositions des morphologues (pour ce qui concerne la dérivation et la composition) et des néologues (pour ce qui concerne l'ensemble des procédés). Concernant la typologie proposée par (Sablayrolles et Pruvost, 2016), nous avons indiqué quelques aménagements possibles. Nous avons également présenté de manière détaillée la matrice externe, dont les emprunts lexicaux constituent le seul cas de néologie formelle. Dans une troisième section, nous avons analysé les unités linguistiques à la base des opérations de dérivation et de composition (affixe, fractolexème, troncat, lexie) : classiquement, la dérivation et la composition sont distinguées sur la base des formants (dans le premier cas les affixes, dans l'autre les lexies), mais nous avons montré que les fractolexèmes créent un passage entre les deux types de procédés. Nous avons tout de même établi une série de propriétés typiques de l'un et l'autre cas, en établissant un continuum jusqu'aux constructions lexico-syntaxiques et syntaxiques. Dans une quatrième section, nous avons abordé le concept de productivité, qui est central pour l'analyse de la formation de nouvelles lexies, car, que ce soit les morphèmes liés (affixes et fractolexèmes), les morphèmes libres (lexies et troncats), ils peuvent être caractérisés par cette notion selon trois modalités: une productivité réalisée, une productivité en expansion et une productivité potentielle. Cette productivité se vérifie à la fois pour la formation de dérivés, mais aussi pour la formation de composés et pour les constructions à composante syntaxique. Cette productivité est l'un des signes du passage du fractolexème vers l'affixe. Elle est également notable dans les composés et les constructions. Enfin, une cinquième et dernière section évoque quelques pistes pour décrire plus en détail et de manière systématique le fonc-

tionnement des procédés de formation de mots, qu'il s'agisse des procédés constructifs (dérivation, composition), des procédés réductifs (troncations) ou des procédés mixtes (amalgames). Il s'agit là d'une des pistes futures de travail sur ce sujet, avec une étude plus multilingue que celle menée ici.

Dans la seconde partie, il s'agissait de mettre au point une méthode de repérage automatique des néologismes formels précédemment délimités. Nous évoquons les méthodes précédemment testées qui se ramènent toutes, peu ou prou, à la méthode dite du dictionnaire d'exclusion ou de référence : on établit une liste de formes attestées de la langue considérée pour une période p, et, sur cette base, dans un corpus nouveau d'une période $p + 1$, on cherche les lexies non attestées qui deviennent des néologismes candidats. Nous soulignons l'inconvénient majeur de cette méthode, qui, *intrinsèquement*, ne peut pas filtrer suffisamment pour obtenir des listes propres de néologismes (noms propres, erreurs typographiques, passages en langue étrangère, non-prise en compte des niveaux de langue, des styles, de la terminologie, erreurs typographiques, difficulté à constituer la ressource de référence). Cependant, en l'absence d'autres méthodes disponibles, nous avons, dans la plateforme Néoveille, mis en place cette technique en ajoutant un certain nombre de modules pour améliorer les résultats. Pour chacune des langues du projet, nous avons effectué une analyse morphosyntaxique automatique qui nous fournit en sortie la liste des mots qu'il ne reconnaît pas. Ensuite, nous appliquons plusieurs filtres : correcteur orthographique, élimination de séquences de lettres non valides dans la langue, élimination des noms propres. Une innovation consiste ensuite à faire valider ou faire invalider par des experts linguistes les résultats : les lexies non validées par les experts sont ensuite réinjectées dans le système comme un dictionnaire d'exclusion. Avec cette méthode nous obtenons une précision de sortie automatique d'environ 60%, après un cycle de validation d'environ 2000 néologismes. Ce système a été mis en place pour les onze langues du projet et donne satisfaction. Cependant, nous présentons de premières expérimentations à l'aide de méthodes d'apprentissage automatique qui montrent que, à partir d'un jeu de référence d'environ 20 000 lexies (néologismes et non-néologismes), le modèle appris permet d'obtenir une précision de près de 90% sur un jeu de référence. À l'évidence, il s'agit là d'une piste de travail que nous aborderons dans un prochain travail.

Troisième partie

Applications

Résumé

Dans cette partie nous présentons le travail effectué à partir de la plateforme *Néoveille* sur l'innovation lexicale en français contemporain à partir de corpus dynamique. Il s'agit d'une présentation qui étend et détaille le travail effectué par un groupe de travail composé de linguistes et de linguistes informaticiens qui ont travaillé sur le sujet depuis septembre 2015. Dans le premier chapitre nous présentons les tendances générales. Dans les chapitres 2, 3 et 4, nous nous intéressons successivement aux procédés de dérivation, de composition, et d'emprunts.

Chapitre 6

Tendances néologiques du français contemporain (2015-2018)

Sommaire

Dans ce chapitre, nous analysons les résultats du travail mené sur la détection, la description et le suivi des néologismes formels du français dans la plateforme Néoveille, de 2015 à juin 2018. Le travail de validation des néologismes candidats, ainsi que la description linguistique des néologismes validés a été effectué par le groupe de travail pour le français [1]. Une version synthétique de ce travail est disponible dans (Cartier *et al.*, 2018c). L'ensemble des données ayant servi à l'analyse qui suit est disponible sur le site Néoveille (`http://www.neoveille.org`).

Cette étude est basée sur un corpus conséquent : Néoveille a récupéré, pour le français, deux fois par jour, depuis juillet 2015, les articles d'environ 250 organes de presse. Au 1er juin 2018, cela représentait 1 143 912 articles pour un total de plus de 176 millions

1. Les personnes suivantes ont effectué le travail de validation et de description des néologismes : Emmanuel Cartier, Najet Boutmgharine, Massimo Bertocci. Des séances collectives de travail ont également eu lieu pendant toute la période, impliquant en outre : Jean-François Sablayrolles, John Humbley, Natalie Kubler, Christine Jacquet-Pfau et Giovanni Tallarico.

de mots [2].

Nous présentons dans ce qui suit tout d'abord la méthodologie suivie, ensuite les tendances générales identifiées. Les chapitres suivants présenteront les résultats pour trois procédés : dérivation, composition, emprunt.

6.1 Méthodologie

6.1.1 Corpus pour l'étude

Le corpus d'étude Néoveille est constitué d'une base d'articles issus de 257 sites de presse en ligne, automatiquement récupérés depuis mars 2016 deux fois par jour à partir des fils RSS publiés par les éditeurs. Au total, au 1er juillet 2018, cela représente plus de 1,5 millions d'articles, soit près de 400 millions de mots. À chaque source d'information sont associées des méta-informations, soit définies manuellement par les linguistes lors de la définition de la source d'information : nom du journal, public visé (presse généraliste, de vulgarisation, spécialisée, féminine, des jeunes), type de texte (actuellement, exclusivement article de presse), domaine (général, économie, industrie, etc [3]), aire géographique (pays et régional/national), soit récupérées automatiquement pour chaque item d'information dans le fil RSS ou dans la page web contenant l'article : auteur(s), date de publication, mots-clés, thématique(s). Les distributions diatopique (tableau 6.1 [4]) et par domaine (tableau 6.2) sont présentées ci-dessous. On trouvera dans l'annexe 1 les tableaux complets pour le français et les autres langues.

Pays/région	**Nbre sources**	**Nbre d'articles**
France	157 (139, 18)	1 209 824
Algérie	54 (44,10)	115 902
Sénégal	25 (23,2)	56 167
Liban	8 (8,0)	34 154
Canada	4(1,3)	94 321
Belgique	4 (4,0)	52 270
Maroc	4 (4,0)	11 763
Totaux	**256**	**1 574 401**

TABLE 6.1 – Synthèse sur la répartition des articles par pays

On notera que la presse féminine, correspondant à un type de public cible, regroupe différents domaines : mode, beauté, achats, cuisine, art de vivre ; la presse des jeunes

2. On trouvera en annexe 1 la liste complète des sites scannés depuis cette date, ainsi que le nombre d'articles récupérés

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

4. Dans la colonne Nbre sources, entre parenthèses sont donnés respectivement les chiffres pour la presse nationale et la presse régionale.

Domaine	Nbre sources	Nbre d'articles
Général	51	962 036
Sport	14	129 569
Presse féminine	55	34 800
Informatique	4	17 066
Politique	3	16 089
Industrie	1	13 599
Economie	6	9 956
Nature	1	6 417
Santé	1	6 015
Sciences	2	5 741
Presse des jeunes	14	4 124
High Tech	1	2 619
Société	3	1 581
Recherche	1	213
Totaux	157	1 209 825

TABLE 6.2 – Synthèse sur la répartition des articles par domaine

n'est actuellement pas récupérée, sauf la rubrique *ado* du Monde. Les différences dans la distribution par pays tiennent évidemment au nombre de titres de presse, d'une part, et au démarrage de la récupération, plus tardive (mars 2016) pour les pays francophones hors France. On remarquera également que le domaine *Général* correspond aux fils RSS annotés par les éditeurs comme correspondant à l'ensemble des informations (spécifié comme *tous les fils de presse, toutes les nouvelles, etc.*, ce qui rend cette catégorie peu utilisable. On notera enfin qu'en dehors du corpus Néoveille, les experts linguistes avaient la possibilité d'ajouter dans le système des néologismes repérés par collecte manuelle dans les sites de presse, ou via le moteur de recherche *Google*. Cependant, l'ensemble des occurrences dans la présente étude est limitée au corpus Néoveille.

6.1.2 Validation des néologismes collectés automatiquement

Comme indiqué dans le chapitre 5, les néologismes sont d'abord détectés automatiquement par la méthode itérative du dictionnaire de référence. En moyenne, pour le français, entre 100 et 200 néologismes candidats (NC) sont ainsi 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 type de néologismes détaillés dans le chapitre 3, soit, pour la seconde, un type particulier de non-néologisme parmi une liste qui a été mise au point progressivement, selon les cas rencontrés : erreurs typogra-

phiques [5], autres erreurs [6], gentilé, particularisme [7], xénisme [8], dictionnaire de référence (mot simple ou composé), dictionnaire terminologique). Le processus de validation des néologismes suit le protocole suivant : chaque membre du groupe de travail annote 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 l'ajout de : 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.

6.1.3 Description des néologismes validés

Une fois validés, les néologismes passent, sur la plateforme, dans le gestionnaire de néologismes, qui permet de décrire linguistiquement les innovations, et d'obtenir un certain nombre d'informations. Comme indiqué dans le chapitre 1, deux niveaux d'information sont disponibles à chaque état de langue : un niveau linguistique et un niveau socio-pragmatique. En diachronie, des informations sur l'évolution des différentes propriétés sont disponibles. Du point de vue de la description linguistique, nous avons utilisé une version simplifiée de la microstructure développée dans (Cartier et Sablayrolles, 2009; Sablayrolles, 2010). Cela aboutit à une microstructure comprenant les champs détaillés dans le tableau 6.3, les deux derniers champs étant propres aux emprunts.

À ces informations, il faut 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 (Stefanowitsch et Gries,

5. il s'agit généralement de coquilles.

6. il s'agit généralement d'erreurs liées au traitement automatique, par exemple un mauvais découpage en mots (*cesondagepour*).

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. Voir (Reutner, 2017) pour une synthèse récente concernant le français.

8. Nous utilisons la définition de (Guilbert, 1971). 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 (McMahon, 1994, p.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. La conservation de ces lexies permet de suivre leur cycle de vie qui peut éventuellement aboutir à les intégrer comme emprunts.

Informations	Définition succincte	Exemple (solférinodactyle)
Partie du discours	Catégorie morphosyntaxique parmi : nom, verbe, adjectif, etc.	Nom
Classe sémantique	Classe sémantique générique. Comprend huit valeurs inspirées de (Le Pesant et Mathieu-Colas 1998)	Humain
Définition		désigne un partisan du parti socialiste, de manière ironique
Procédé(s) néologique(s) impliqués	Le ou les mécanismes néologiques impliqués dans l'innovation lexicale, en partant de la typologie de (Sablayrolles et Pruvost, 2016)	Fracto-composition
Configuration syllabique	Description générique et détaillée de la configuration syllabique de la lexie, au moyen des notions de syllabes ouverte et fermée.	FOOOFF
Configuration morphologique	Décomposition morphologique de l'innovation, au moyen des notions de radical, d'affixe et de formant.	Solférino-dactyle (RAD-RAD)
Lexie base	Identification de la ou des lexies ayant servi de base au néologisme	Soférino, dactyle
Partie du discours lexie base	Identification de la partie du discours de la lexie base, ou de la racine.	Nom, Adjectif
Influence langue	Indication de la langue origine de l'emprunt ou des influences présentes dans la création d'un néologisme, formé par une / des matrices internes	Aucune
Mode influence	Type de l'influence, parmi l'une des catégories suivantes : traduction (ou calque sémantique), calque morphologique, formant emprunté, sens emprunté, structure empruntée, synthèse néologique	NA

TABLE 6.3 – Description linguistique manuelle dans Néoveille

2003) et les constructions lexico-syntaxiques les plus fréquentes ; le *profil distributionnel* (Harris, 1954; Baroni et Lenci, 2010; Mikolov *et al.*, 2013), permettant d'accéder aux lexies les plus similaires dans leurs distributions et donc sémantiquement similaires, n'est actuellement pas implémenté, car il nécessite un corpus plus conséquent. Nous illustrons les deux premières informations dans le tableau 6.4.

6.2 Tendances générales du français contemporain

Nous analysons maintenant les résultats du travail collaboratif de validation et de description des néologismes, de juin 2015 à juin 2018. 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 est disponible sur la plateforme du projet.

6.2.1 Répartition par mécanismes

De juin 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

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	**Collocations :** *fast food (16), slow food (16), street food (11), raw food (9), junk food (7), food market (7)* **Collostructions :** *tendance food (10) => ADJ phénomène food (9) => ADJ projet food (5) = > ADJ Det (masc) food (10) => NOM* **Constructions lexico-syntaxiques :** *food + verbe : aller, débarquer, arriver, consister, cartonner...*

TABLE 6.4 – Description linguistique automatique dans Néoveille

été récupérés. Parmi environ 35 000 néologismes formels candidats, 22 475 néologismes 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 %.

Dans le tableau 6.5, 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. On constate que :

	formes uniques		**occurrences**		**Moy. d'occ.**
Mécanisme néologique principal	**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
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 / moyenne	**22 475**	**100,00%**	**726 222**	**100,00%**	**26**

TABLE 6.5 – Synthèse sur les matrices néologiques en français contemporain

— le procédé largement 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. Cette disparité entre la préfixation et les autres procédés provient notamment de la très grande productivité d'une vingtaine de préfixes (voir section sur les préfixes) ;
— 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. Le fort nombre d'occurrences pour les emprunts est notamment dû à un certain nombre d'emprunts dont nous souhaitions analyser l'implantation et non simplement l'émergence (notamment toutes les noms des réseaux sociaux, et leurs dérivés : *Facebook, Twitter, Instagram*, etc.) ;

— 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).

6.2.2 Répartition par journaux, domaine, pays

Les chiffres précédents doivent être ramenés aux paramètres diatopiques et diastratiques disponibles sur la plateforme. On constate alors que (voir figure 6.1) :

— En valeur absolue d'occurrences, les **journaux les plus productifs de néologismes** sont : *L'Express*, *France Soir*, *La Croix*, *Le Monde* puis *Libération*. En valeur relative (en divisant le nombre d'occurrences par le nombre total d'articles par journal), on obtient un résultat qui rend mieux compte de la "néologicité" de chaque journal : on constate alors qu'un magazine de sport (*So Foot*), les magazines *Slate* et *Le Nouvel Observateur* sont les plus grands pourvoyeurs de néologismes, loin devant l'ensemble des autres sources, avec *Libération*, le *JDD* et *Le Monde* en tête de ce second wagon. Notons également que si l'on effectue cette distribution par type de néologismes, les répartitions sont *grosso modo* les mêmes, en dehors des emprunts pour lesquels les magazines féminins (*Elle, Grazia*, etc.), la presse à base anglo-saxonne (*Slate, Le Huffington Post*) et la presse *people* sont en tête. Du point de vue des innovateurs, en se basant sur les seules premières occurrences de néologismes, les répartitions sont les mêmes.

— La **répartition par domaine** révèle la prédominance de trois domaines[9] : le sport (10 %), la presse féminine (10 %) et l'informatique (5 %). Étant donné la distribution de notre corpus, correspondant à cette répartition, il est difficile d'en tirer une quelconque conclusion. Cependant, les domaines informatique, de la presse féminine, et dans une moindre mesure du sport, 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.

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

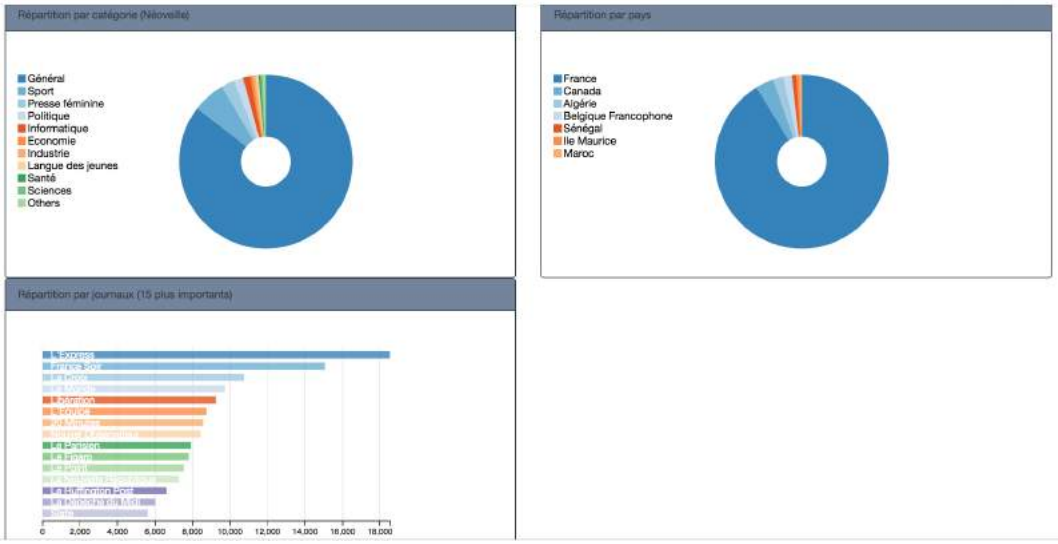


FIGURE 6.1 – Distribution des néologismes par domaine, par pays et par journaux (au 31/08/2018)

6.2.3 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 complet [10] montre que la néologie génère principalement des noms et des adjectifs, verbes et adverbes étant sous-représentés. Étant donné l'absence de chiffres de référence antérieurs, il est difficile d'en induire des particularités contemporaines. Cette distribution répond aux répartitions constatées pour les mécanismes de formation de mots étudiés dans un cadre multilingue (par exemple la partie II de (Lieber et Štekauer, 2014)), ainsi que pour certains procédés (par exemple la troncation, voir (Antoine, 2000)).

6.2.4 Cycle de vie des néologismes

Pour étudier le cycle de vie des néologismes, nous pouvons partir de la notion d'hapax, ainsi que des trois phases généralement identifiées pour désigner les phases saillantes de ce cycle de vie : émergence, diffusion, lexicalisation (voir chapitre 2 pour une discussion).

6.2.4.1 Émergence des néologismes

Comme nous l'avons vu, la moyenne d'occurrences est relativement faible par néologisme. La déviation standard [11] est importante (111 par forme néologique, 237 par

10. 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.

11. 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

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 6.6 détaille ces répartitions (colonne 1 : nombre d'occurrences entre deux seuils, seuil supérieur non inclus ; colonnes 2 et 3 : nombre de néologismes par fréquence calculée par le total d'occurrences ; colonnes 4 et 5 : nombre de néologismes par fréquence calculée comme le nombre de documents couverts (pour ne pas prendre en compte la répétition au sein d'un même document). On constate que seulement 27% des néologismes

Nb d'occurrences	**Par nbre d'occurrences total**	**%**	**Par nbre de documents différents**	**%**
1	6138	27,37%	7336	32,72%
2	4244	18,93%	3768	16,80%
3	1232	5,49%	1978	8,82%
4	768	3,43%	754	3,36%
[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%

TABLE 6.6 – Distribution des néologismes par fréquence

sont des hapax (fréquence 1). Mais si nous regardons les néologismes ayant une faible fréquence (entre 2 et 5), nous trouvons des exemples comme *digérateur, serial-buteuse, éco-orgasme, droitdelhommiste,workaholisme, sur-coûter, tablatiste* qui sont à l'évidence des néologismes récents et qui ont toute chance de ne pas diffuser et de rester des hapax. Il est donc nécessaire de réviser la notion d'hapax généralement associée à celle d'occasionnalisme (*nonce-word*), pour plusieurs raisons :

- il existe un continuum, matérialisé par une courbe de Zipf ou bien un graphe à points (voir figure 6.2), de la distribution de fréquence) entre les fréquences les plus basses et les plus hautes fréquences ; il existe bien une zone "hapaxique" comprenant le plus grand nombre de néologismes, mais cette zone couvre plus que les seuls hapax au sens strict ;
- la communication via internet accélère la transmission et la reprise, parfois tel quel, des informations, ce qui ne présume pas du sort des néologismes, mais favorisant une répétition pendant la période d'émergence ;
- de nombreux groupes de presse détiennent plusieurs titres, et il est fréquent de voir des répétitions survenant quasiment au même moment, dans des journaux différents, souvent 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

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.

titres différents, le 18 mars 2016 (trois compocations basées sur trois emprunts anciens : *cookie, doughnut, brownie*, début du XIXe):

> « 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)

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%

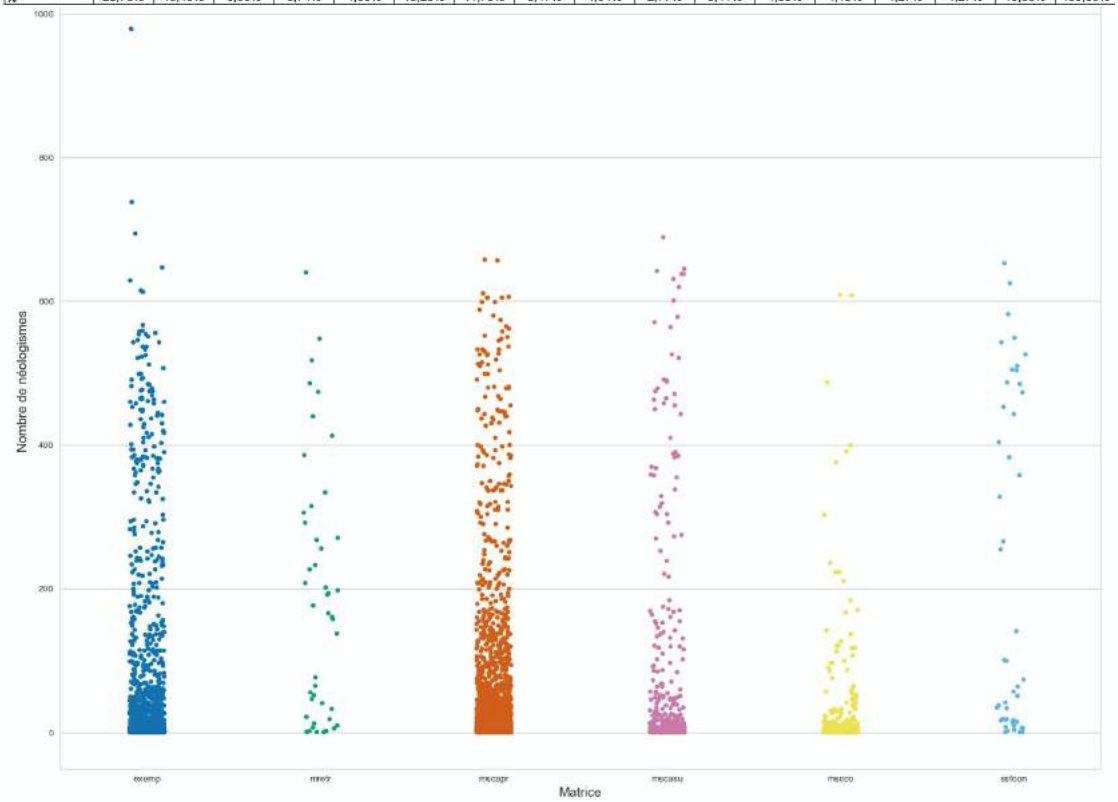


FIGURE 6.2 – Distribution des néologismes par fréquence pour les emprunts (tableau) et distribution des néologismes par fréquence (en abscisse), pour les six types de néologismes les plus fréquents (en ordonnée)

Il ne semble donc pas possible de tracer une frontière linguistique pertinente entre les hapax et les autres néologismes (notamment ceux ayant une fréquence (très) « faible ») et, si nous souhaitons pouvoir expliciter des critères pour identifier l'émergence, la seule fréquence n'est pas suffisante. 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 via la notion d'hapax. Le critère socio-pragmatique permet encore d'affiner la définition de l'émergence. En effet, cette

dernière se produira généralement dans un type de textes donnés. 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 6.3 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).

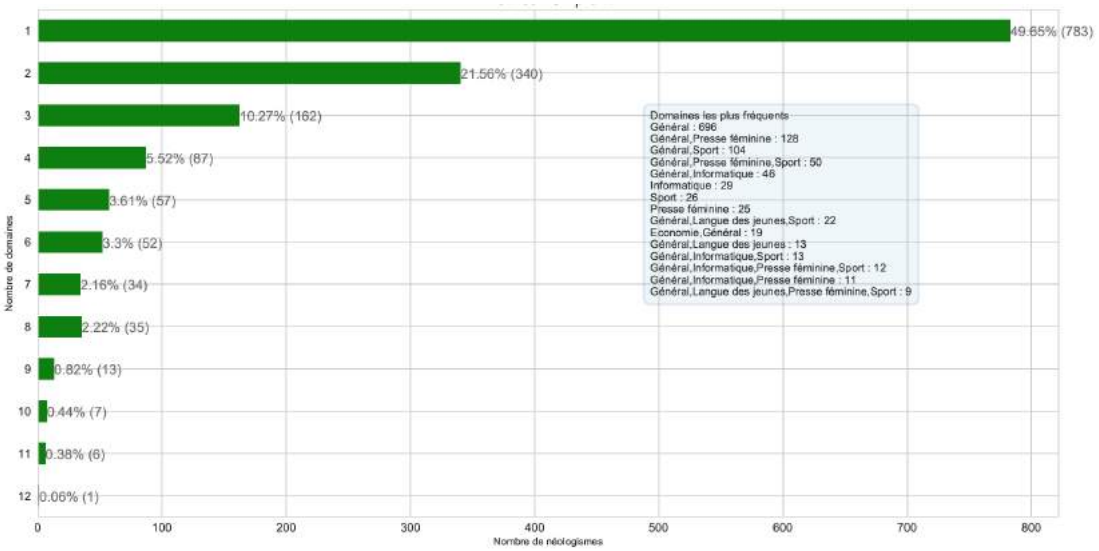


FIGURE 6.3 – Distribution des emprunts par nombre de domaines représentés. Le tableau interne détaille les combinaisons de domaine les plus fréquentes

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 compositionnalité : « ...Réponse jeudi où se tiendra juste après le défilé, une grande after-party... » (Elle, 21/02/2017). L'absence de glose est également très fréquente pour les procédés d'affixation, dont le sens est généralement compositionnel et donc transparent.

Nous pouvons cependant proposer une caractérisation générale de l'émergence :

> « **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 (et plus précisèment dans les mêmes contextes socio-pragmatiques) que celui de la première apparition. Lorsque le sens n'est pas compositionnel, l'émergence s'accompagne généralement d'une glose.** »

Une caractéristique des quasi-hapax (au sens de *nonce-word*) 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*).

Il est également utile de distinguer la distribution des fréquences par matrice néologique. En effet, on constate notamment que les préfixations et les dérivations sont moins marqués par les quasi-hapax que les autres mécanismes.

6.2.4.2 Diffusion des néologismes

Pour étudier la diffusion des néologismes, nous prendrons comme exemple les emprunts. Dans ce groupe, près de 15% des néologismes ont une fréquence supérieure à 50, d'une part, ce qui permet d'étudier la diffusion. D'autre part, les emprunts présentent une particularité : en tant que matériau provenant d'un autre système, il se produit une phase spécifique d'adaptation phonologico-morphologique, qui est l'un des premiers signes d'une diffusion.

6.2.4.2.1 Modèles fréquentiels de diffusion En partant du modèle de la courbe sigmoïde de diffusion des innovations (Rogers, 2010), la phase de diffusion doit être caractérisée par une augmentation rapide des occurrences, correspondant à l'adoption de la lexie par les adopteurs précoces. Cependant, dans notre corpus, pour les néologismes à forte fréquence (supérieurs à 100 occurrences) nous n'obtenons pas de courbe significative répondant à cette définition, et il semble difficile de trouver un ou des modèles de diffusion, peut-être parce que l'empan temporel n'est pas encore suffisant pour ce corpus. Prenons l'exemple de *cyberattaquant*, dont les premières mentions datent de 2014, malgré l'apparition bien avant de *cyberattaque*[12]. Les figures 6.4 et 6.5 montrent respectivement, sur la période 2016-2018, l'évolution fréquentielle et la distribution fréquentielle par domaine des sources. Pour l'évolution fréquentielle, nous avons indiqué la fréquence brute (en noir), la moyenne roulante (en prenant les deux valeurs précédentes et en calculant la moyenne pour chaque point), la moyenne par expansion (en tenant compte de l'ensemble des valeurs passées et en calculant la moyenne), et la moyenne par

12. Sous la forme *cyber-attaque*, les premières attestations sont vers 1988, mais la forme soudée n'est pas attestée avant 2008

expansion exponentielle (même méthode que précédemment mais en affectant un coefficient décroissant aux valeurs passées selon leur distance au point courant). Ces courbes montrent, pour un néologisme ayant 148 occurrences, en dehors d'un pic d'emploi durant l'été 2017 (cyberattaques aux Etats-Unis et dans toute l'Europe), une fréquence faible, sans tendance croissante ou décroissante claire. Il ne semble donc pas que le néologisme diffuse, même si à regarder les contextes, aucune glose n'est faite, peut-être du fait de la transparence de la formation. À regarder la seconde figure, qui explicite la distribution temporelle par domaine, on s'aperçoit que le néologisme est bien implanté dans les nouvelles généralistes mais également dans de nombreux domaines, spécifiquement dans les domaines principalement concernés (high tech, industrie, informatique) mais également, de manière plus ponctuelle, dans d'autres domaines. Qu'en conclure ? Pour ce qui concerne *cyberattaquant*, on peut émettre l'hypothèse qu'il s'agit, d'une part, d'une lexie appartenant à un domaine spécifique et qui, de ce fait, ne pourra pas diffuser au-delà d'un seuil de fréquence. Mais sa diffusion dans beaucoup de domaines laisse à penser qu'il a déjà diffusé, voire est adopté par la communauté linguistique pour désigner les acteurs de cyberattaques. Cependant, on voit bien que la courbe de fréquence n'est pas d'une grande aide, même si un empan temporel élargi permettrait peut-être d'obtenir une vision différente. Il faut donc utiliser d'autres paramètres pour émettre une hypothèse sur la diffusion.

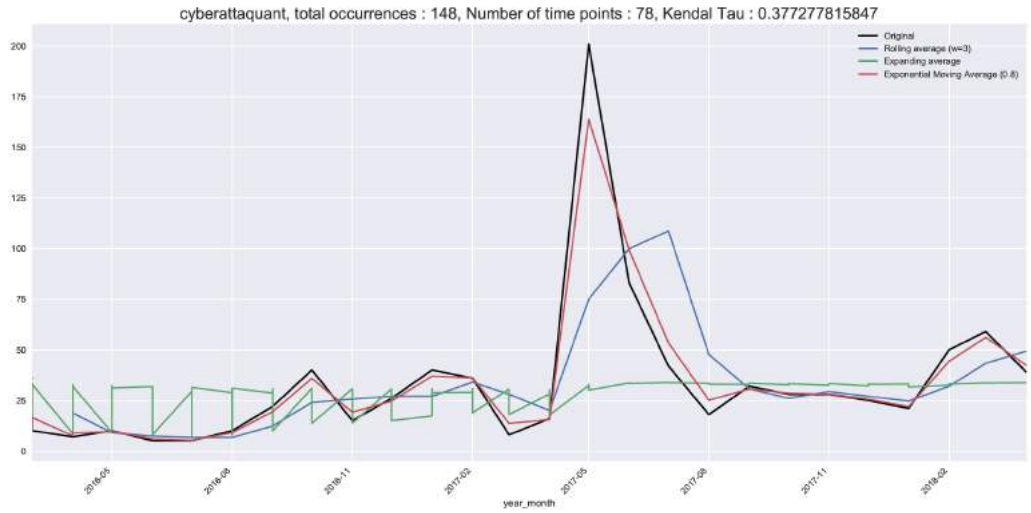


FIGURE 6.4 – Évolution fréquentielle des occurences de *cyberattaquant*

6.2.4.2.2 Adaptations phonologique, orthographique et morphosyntaxique On peut faire l'hypothèse que les emprunts obéissent aux différentes phases communes d'intégration. Il se produit également, étant donné que le matériau emprunté provient d'un autre système, une adaptation au système du français aux niveaux phonologique, orthographique et morphosyntaxique.

Concernant l'adaptation orthographique, le corpus Néoveille ne présente pas de par-

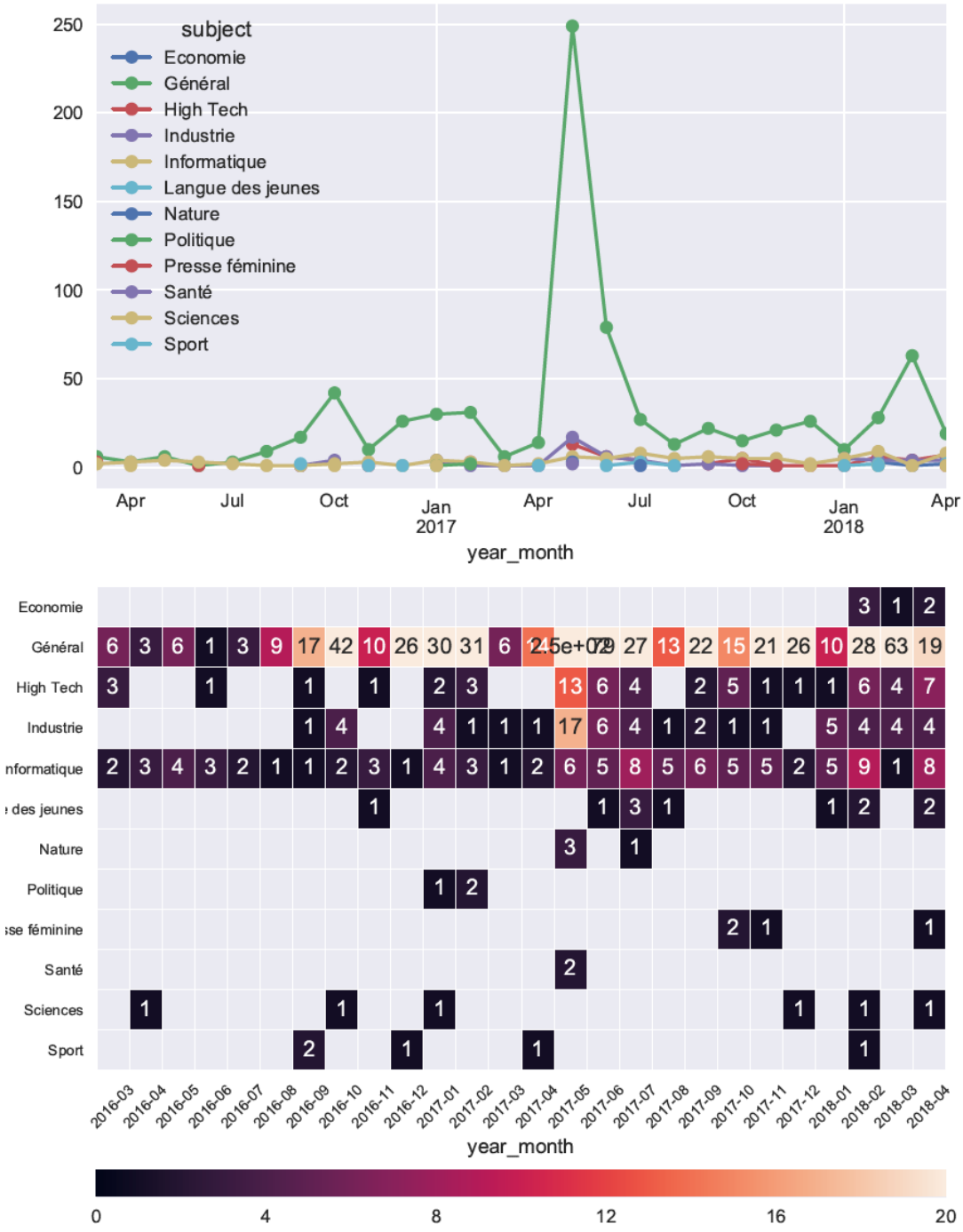


FIGURE 6.5 – Évolution temporelle par domaine pour *cyberattaquant*

ticularité concernant les emprunts à des langues alphabétiques, dont les lexèmes sont généralement rendus tels quels [13]. 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, p.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.

6.2.4.2.3 Intégration à la morphologie productive Un autre signe de diffusion concerne l'intégration à la morphologie productive. La grande majorité des emprunts à fréquence moyenne ou forte relevés dans notre corpus subissent ce « passage » à la dérivation. Les exemples les plus typiques sont liés aux bases nom propre issus des réseaux sociaux (tableau 6.7)). 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 (voir plus loin et (Cartier *et al.*, 2018c). 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 dérivé français pourtant très bien implanté 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*).

Le même phénomène se produit pour les emprunts à base nom commun les plus populaires (plus ou moins récents) : *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-

13. 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*).

	facebook	**twitter**	**instagram**	**snapchat**	**youtube**
Intégration morphologique (morphème flexionnel)	facebooker (v)	twit(s) (n), tweeter (v), twitto(s) (n)	instagram(m)er (v, n)	snapchater (v), snapchat(s) (n) > snap(s)	youtuber (v)
Intégration morphologie productive (affixes, fracto-lexèmes et formants savants)	facebookeur (-euse), facebookien (-ne), facebooking, anti-facebook, facebookisme	twitteur (-euse), tweeteur (-euse), re(-) tweeter, tweeterisation, tweeting, anti-tweet, demi-tweet, non-tweet, auto-tweet, pseudo-tweet, tweetesque, tweetable, tweetonade	instagram(m)eur (-euse)	snapchat(t)eur (-euse), snapchat(t)ien (-ne)	youtubeur (-euse), youtubing

TABLE 6.7 – Intégration à la morphologie productive des noms de cinq réseaux sociaux

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).

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 : *tweet-wall, acrostweet, live-tweet, tweetdeck, tweetstorm, fake-tweet, tweetbot, commander-in-tweet*, etc.

6.2.4.2.4 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, caractérisant 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 [14]. Très rapidement le terme se répand aussi en 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 6.8), 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* ;

14. 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 » de l'article ci-après : `https://jezebel.com/charlize-theron-broke-up-with-sean-penn-by-ghosting-him-1712760688`

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.) fixent très rapidement l'usage syntactico-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...

TABLE 6.8 – Exemples d'emplois de *ghosting* dans Néoveille

La fixation d'un riche profil combinatoire est également illustrée pour *food* dans le tableau 6.9.

6.2.4.2.5 Évolution des contextes socio-pragmatiques des innovations 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 6.6 et 6.7, 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 sociolinguistique spécifique, tandis que la seconde, dont on peut constater

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	**Collocations :** *fast food (16), slow food (16), street food (11), raw food (9), junk food (7), food market (7)...* **Collostructions :** *tendance food (10, ADJ), phénomène food (9, ADJ), projet food (5, ADJ), Det(masc) food (10,N)...* **Constructions lexico-syntaxiques :** *food + verbe : aller, débarquer, arriver, consister, cartonner...*

TABLE 6.9 – Profil combinatoire de *food* dans Néoveille

l'émergence dans les domaines *hightech* et informatique, se diffuse maintenant dans la presse généraliste.

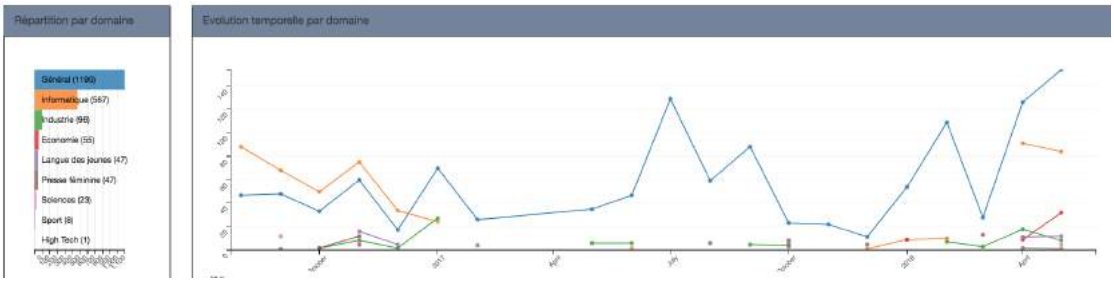


FIGURE 6.6 – Distribution temporelle par domaine pour *smartphone*

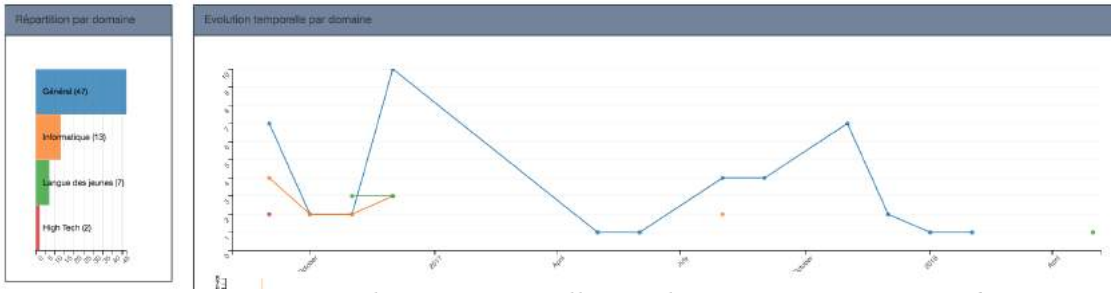


FIGURE 6.7 – Distribution temporelle par domaine pour *smartwatch*

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), les groupes sociaux innovateurs et diffuseurs (adopteurs), grâce au suivi temporel de la distribution domaniale.

6.3 Conclusion prospective

Dans ce chapitre, nous avons présenté les tendances néologiques du français contemporain (2015-2018) telles qu'elles ont pu être constatées grâce à la plateforme Néoveille et au travail de validation et de description d'un groupe de travail sur le français. Après avoir présenté quelques éléments méthodologiques pour effectuer ce travail de validation et de description, nous avons détaillé les caractéristiques principales des néologismes formels : domination quantitative de la préfixation, suivie par la composition simple, les emprunts, la suffixation et la fracto-composition. On peut y voir la domination du procédé de création lexicale le plus immédiatement accessible et le moins ressenti comme néologique, puisqu'il est aussi analysable comme faisant partie de la morphologie productive. Concernant la répartition par domaine, on constate que le sport, la presse féminine et l'informatique sont particulièrement productifs, ce que confirme également la répartition par journaux. Par contre, à ce stade, il n'est pas possible de tirer des conclusions sur la distribution diatopique, étant donné la distribution du corpus qui comprend beaucoup plus de titres de presse métropolitains. La répartition par parties du discours confirme la prédominance des noms, suivis des adjectifs et des verbes. Cela suit la distribution habituellement constatée sur corpus entre ces catégories. Nous avons également étudié les spécificités du cycle de vie des néologismes : la très grande majorité des néologismes sont des hapax, mais nous avons proposé de définir plus précisèment l'émergence des néologismes comme une période courte où se produit une faible diffusion dans des contextes socio-pragmatiques déterminés et similaires. Concernant la diffusion, nous avons, à partir de l'exemple des emprunts, établi un certain nombre de paramètres pour la caractériser : évolution fréquentielle, baisse des mentions et des contextes métalinguistique, adaptation et stabilisation phonologique, orthographique et morphosyntaxique (propre aux emprunts), diffusion dans d'autres contextes socio-pragmatiques, intégration à la morphologie productive, enfin stabilisation du profil combinatoire. La validité de ces paramètres doit être confirmé par des études plus systématiques, qui pourront être effectuées dans des études ultérieures. De même, il sera intéressant d'étudier les évolutions des distributions selon les paramètres fréquentiels, diastratiques et diatopiques.

Chapitre 7

Dérivation en français contemporain (2015-2018)

Sommaire

7.1 Définitions

La dérivation regroupe tous les phénomènes d'affixation : préfixation, suffixation et parasynthèse, résultat de l'adjonction simultanée d'un préfixe et d'un suffixe à une base [1].

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 formants [2].

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 [3]. Ce procédé induit généralement un processus de transcatégorisation (Sablayrolles, 2000 : 264-265).

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

2. 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 numéraux (*déca, déci, hepta, tétra*, etc.). On pourra consulter la liste des préfixes et suffixes du français selon Le *Trésor de la Langue Française Informatisé* à cette adresse : `https://hugonlp.wordpress.com/2015/10/22/articles-sur-les-prefixes-et-les-suffixes-du-tresor-de-la-langue-francaise-informatise/`

3. `https://pr.bvdep.com/demo/AidePR/Pages/SuffixesA.HTML`

Nous rappelons également que le fonctionnement de ces deux éléments de formation emportent deux informations principales : d'une part, une information sur la partie du discours du dérivé ; d'autre part, une information sémantique spécifique, généralement générique. Nous avons donc le schéma de fonctionnement suivant pour les dérivés à base de préfixes :

$$X_{Pref} + Y_{Pos} \Rightarrow \left\{ \begin{array}{l} \text{forme : X(-)Y} \\ \text{syntaxe: catégorie du dérivé donnée par Y} \\ \text{sémantique : sens de Y modifié par X} \end{array} \right\} \tag{7.1}$$

Ou, dans un formalisme plus ramassé, mettant en équivalence la forme de surface et syntaxique avec la valeur sémantique ((Booij, 2010)):

$$[[X]_{Pref} - [Y]_{Pos}]_{Pos} \Leftrightarrow \text{['Sens de Y modifié par X']} \tag{7.2}$$

Sur un exemple :

$$\text{pré}_{Pref} + \text{adolescent}_{Adj} \Rightarrow \left\{ \begin{array}{l} \text{forme : pré(-)adolescent} \\ \text{synt.: Adj} \\ \text{sém. : état précédent celui d'adolescent} \end{array} \right\} \tag{7.3}$$

Et pour les suffixes [4] :

$$X_{Pos} + Y_{Suff} \Rightarrow \left\{ \begin{array}{l} \text{forme : XY} \\ \text{synt.: catégorie du dérivé donnée par Y} \\ \text{sém. : relation R dénotée par Y appliquée à X} \end{array} \right\} \tag{7.4}$$

Dans le formalisme de Booij:

$$[[X]_{Pos1} - [Y]_{Suff}]_{Pos2} \Leftrightarrow \text{['relation R dénotée par Y appliquée à X']} \tag{7.5}$$

Sur un exemple :

4. Nous considérons ici que c'est le suffixe qui donne l'instruction de catégorie morphosyntaxique (ou catégorielle) au dérivé. Ce qui ne signifie pas nécessairement une transcatégorisation : *maison => maisonnette, brillant => brillantissime.* Ce serait l'une des propriétés distinctives entre préfixes et suffixes. (Corbin, 1999) réfute cette analyse, considérant que les deux formants auraient une capacité catégorisatrice, qui serait liée en premier lieu à l'instruction sémantique qu'ils portent. Les contre-exemples qu'elles donnent ne sont pas convaincants : d'une part, les formations créant un adjectif à partir d'une base nom *anti-gang, apétale, transalaska* peuvent être interprétés comme des emplois épithètes relationnels, d'autre part, les créations verbales à partir de bases nominales (*apaiser, dépuceler, transvaser* ou adjectivales (*aplatir, élargir, enlaidir*) peuvent être interprétés comme le résultat d'abord d'une transformation (générant des verbes non attestés) des adjectifs ou des noms en verbe, par les suffixes verbaux.

$$\text{Youtube}_{Np} + \text{eur}_{Suff} \Rightarrow \left\{ \begin{array}{l} \text{forme : youtubeur} \\ \text{synt.: Nom Commun} \\ \text{sém. : (agent) producteur de contenu} \\ \text{publié sur Youtube} \end{array} \right\} \quad (7.6)$$

7.2 Préfixation

Dans le corpus Néoveille, nous avons repéré 14 875 néologismes, correspondant à 69 formes préfixales, pour 197 244 occurrences. La distribution est explicitée dans la figure 7.1 [5]. On retrouve une distribution de Zipf classique, mais avec une particularité par rapport aux autre procédés : si l'on analyse la boîte à moustaches, il y a un grand nombre de cas à fréquence exceptionnelle (*outliers* d'une part, et le dernier quartile (75%) est bien à l'extérieur des valeurs médianes (la boîte colorée, elle-même assez large, limitée par la premier et le troisième quartiles, et indiquant la médiane), montrant qu'il existe beaucoup de néologismes à forte fréquence. Cette situation a une explication principale : nous avons affaire avec la préfixation à l'un des procédés les plus productifs, dont les formations ne sont pas répertoriés par les dictionnaires, étant donné la régularité des règles de ces formations. Il est donc peu évident de dire si un néologisme préfixal est un néologisme ou un mot potentiel inscrit dans la langue par la règle qui permet de le générer. De plus, nous avons opté pour la validation de ces néologismes sur l'utilisation de *Google N-grams*, dont le corpus va jusqu'en 2008. Nous retrouvons cependant pour les préfixes une moyenne d'occurrence normale de 12,68 occurrences par néologisme et une médiane de 3, montrant la grande dispersion de la distribution.

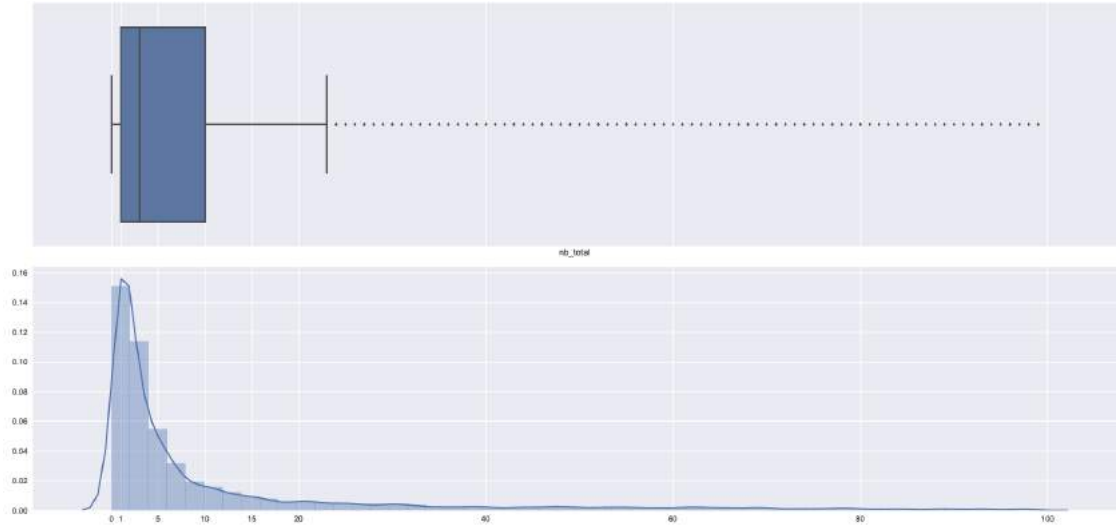


FIGURE 7.1 – Distribution des préfixes dans Néoveille

5. Pour une explication des données contenues dans la *boîte à moustaches*, nous renvoyons par exemple à `https://www.parfenoff.org/pdf/1re_S/stat_proba/1re_S_Diagramme_en_boite.pdf`

Le tableau 7.1 explicite les 69 formes préfixales repérées dans Néoveille, ainsi que : le nombre de formes uniques correspondantes, le nombre total d'occurrences, le nombre d'hapax au sens strict (fréquence = 1), le nombre d'hapax au sens large (fréquence < 6), et la productivité en expansion (division du nombre d'hapax au sens large par nombre de formes uniques). Le tableau est trié par le nombre total d'occurrences.

Ces données appellent plusieurs commentaires:

— Parmi la liste pré-établie de préfixes, 12 sont absents (*ambi-, ana-, apo-, cata-, circo-/circum-, dia-, ecto-, endo-, eu-/ev-, juxta-, pén(é)-, per-*), devenus non-productifs ou en tout cas non attestés dans le corpus Néoveille ;

— Les autres préfixes sont déjà mentionnés par les études antérieures (Dubois, 1962; Corbin, 1987) ;

— On note, pour un certain nombre de préfixes, une ou des formes fléchies (*supers, micros, pseudos, demie, ultras, minis, posts*). En dehors de *demi*, ces formes suffixales fléchies sont nouvelles. Elles ne sont cependant pas en grand nombre, avec exclusivement des hapax. Il pourrait s'agir, pour un cas, d'un emploi fautif (*posts-matchs*). Par contre, tout en étant fautifs, les autres cas, de par leur répétition dans des contextes différents, peuvent laisser penser qu'une évolution se dessine (*supers-productions, supers-génies, micros-partis, micros-robots, pseudos-prescriptions, pseudos-amis, ultras-riches, ultras-conservateurs,minis-jobs, minis-poussins*). Ces cas concernent les préfixes qui disposent également d'une valeur adjectivale ou nominale par ailleurs (*pseudo, ultra, mini*). On peut voir dans ce phénomène un retour à l'autonomisation des préfixes, qui s'est produit pour d'autres formants (*ex, anti,extra*). Il s'agit là encore d'un signe du dynamisme lexical. On notera également que cela concerne en majorité des préfixes évaluatifs et quantitatifs ;

— On constate une continuité entre les différentes fréquences (que l'on prenne le nombre de formes uniques ou le nombre total d'occurrences), même si les cinq premiers se détachent *non, ex, anti, quasi, ultra*, et les 15 derniers (à partir de *proto*) sont en situation de quasi-non-productivité. Si l'on compare ces chiffres avec un comptage antérieur ((Cartier *et al.*, 2018c)) qui couvrait une période inférieure de six mois, on constate des évolutions qui dénotent une variation indiquant le dynamisme lexical et sans doute des effets de mode.

— du point de vue de la productivité en expansion, d'autres formants sont en tête (supérieure à 0.7) : *mini, post, co, multi, mi, micro, auto, semi, extra, pseudo, ré, dé, archi, infra, tri*. Il est cependant difficile d'en tirer des conclusions, car il pourrait s'agir de tendances ponctuelles ;

— Les préfixes s'appliquant aux verbes sont beaucoup moins représentés dans cette liste (*post, auto, sous, re, co, pré, sur, contre*, etc.) que les préfixes produisant des noms, adjectifs et adverbes.

Il serait nécessaire de faire une analyse des préfixes par champs sémantiques couverts. À notre connaissance, aucune catégorisation sémantique exhaustive des préfixes n'a été faite. En partant de la liste partielle de champs proposée par (Corbin, 1987) et reprise dans (Corbin, 1999), on peut regrouper les préfixes selon les champs suivants :

préfixe	formes uniques	occurrences	Hapax (1)	Hapax (<6)	productivité
non	2048	46806	344	916	0.447
ex	1693	37067	392	935	0.552
anti	1767	20348	617	1210	0.685
quasi	772	12509	200	436	0.565
ultra	964	10713	254	596	0.618
super	585	7425	140	346	0.591
mini	902	5470	342	690	0.765
pro	429	4903	134	263	0.613
post	543	4431	207	393	0.724
sous	183	3983	64	116	0.634
hyper	405	3595	110	264	0.652
pré	435	3585	145	286	0.657
co	145	3534	30	66	0.455
multi	361	3335	129	259	0.717
mi	602	3119	338	516	0.857
micro	400	3020	155	301	0.752
demi	378	2808	157	306	0.81
auto	247	2386	94	177	0.717
semi	318	2293	125	239	0.752
inter	172	1909	54	112	0.651
re	85	1866	35	56	0.659
sur	239	1790	74	160	0.669
extra	127	1420	57	97	0.764
contre	57	1106	12	34	0.596
pseudo	264	856	129	236	0.894
hors	8	808	0	1	0.125
méga	93	806	28	64	0.688
tout	46	780	4	17	0.37
ré	186	738	75	151	0.812
sans	12	706	2	4	0.333
bi	26	434	5	12	0.462
dé	55	329	29	46	0.836
mono	20	155	6	11	0.55
archi	13	143	3	10	0.769
trans	5	134	0	1	0.2
maxi	10	119	5	7	0.7
intra	12	90	6	8	0.667
pluri	10	80	2	4	0.4
in	4	68	2	3	0.75
infra	6	57	2	5	0.833
tri	16	54	12	15	0.938
sub	5	44	2	4	0.8
après	6	43	2	3	0.5
poly	8	31	4	5	0.625
supers	10	31	5	9	0.9
méta	1	26	0	0	0.0
micros	8	25	5	7	0.875
supra	5	19	1	4	0.8
primo	2	19	1	1	0.5
avant	5	16	2	4	0.8
demie	3	11	1	2	0.667
pseudos	4	10	2	3	0.75
ultras	4	10	2	4	1.0
anté	1	10	0	0	0.0
proto	5	9	2	5	1.0
omni	3	7	1	3	1.0
alter	2	6	1	2	1.0
dés	4	5	3	4	1.0
épi	1	5	0	1	1.0
minis	4	5	3	4	1.0
a	2	4	0	2	1.0
meta	2	3	1	2	1.0
para	2	2	2	2	1.0
dis	1	2	0	1	1.0
giga	1	1	1	1	1.0
hypo	1	1	1	1	1.0
macro	1	1	1	1	1.0
pan	1	1	1	1	1.0
posts	1	1	1	1	1.0

TABLE 7.1 – Liste des préfixes repérés dans Néoveille (31/08/2018)

— **évaluatifs** : éléments exprimant prototypiquement une évaluation : *giga, mini, ultra, super, maxi, archi, méga, extra, hyper*, éléments pouvant avoir une composante évaluative : *sous, micro, sur, sans, sub, supra, proto, omni, para, hypo*,

suffixe *issime*.

— **quantitatifs** : *macro, micro, quasi, mini, hyper, hypo, multi, mi, demi, semi, tout, mono, archi, maxi, pluri, tri, bi, poly, supra, omni, primo, épi* ;
— **localisation temporelle** : *ex, post, pré, re, après, avant, primo ?, anté, pan ?, trans, co, proto ?* ;
— **localisation spatiale** : *pan, para, infra, intra, trans, contre, sur, inter, sous* ;
— **négation, privation, opposition** : *para, contre, dis, dé, para, a, in, sans, contre, anti, pro, non* ;
— **autres** : *méta, pro, co, auto, pseudo, quasi, alter*

Nous nous cantonnerons à faire ici une analyse des évaluatifs, en partant des analyses proposées par (Amiot, 2004a). Le linguiste explicite un certain nombre de caractéristiques des évaluatifs exprimant le haut degré : la majorité sont issus de prépositions grecques puis latines, sauf *archi, méga et maxi* qui sont issus d'adjectifs, pour les deux derniers, et le premier proviendrait également d'une expression adjectivale liée à la supériorité hiérarchique ((Amiot, 2004a, p.5)). Tous les évaluatifs sont polysémiques (voir figure 7.2), avec, dans le domaine de l'évaluatif, trois pôles : évaluation quantitative (par rapport à une norme : supériorité (*superbombardier, surdoué*) et excès (*hypertension, surexposé*) et le haut degré [6]. Pour le sens haut degré, ils se construisent quasi exclusivement avec des adjectifs (et beaucoup plus rarement avec des noms dont il est difficile d'évaluer la norme : /textitsuperforme), et semblent interchangeables (*ultrafin, extrafin, superfin, etc.*). Un test pour déterminer l'interprétation 'évaluation quantitative' est de substituer le préfixe par un préfixe d'infériorité (*sous-doué*, mais pas **sousfin*).

On peut compléter cette analyse avec les données proposées dans Néoveille. On constate en effet la domination de trois préfixes : *ultra, super et hyper*, les autres (*extra, méga, archi, maxi, giga*) et *sur, sous* (qui se partagent entre un sens de localisation et d'évaluation) étant en nombre bien moindres. La concurrence entre les trois premiers est-elle réelle, ou bien se répartissent-ils le champ ? La prédominance d'*ultra* avait déjà été notée par (Corbin, 1987) avec une application majoritairement aux adjectifs (468) mais également aux noms (14 occurrences : *ultra-bike, ultra-centre, ultra-communication, ultra-connexion*, etc.). Comme *super, hyper*, il se construit sur des bases adjectivales ou nominales, pour construire des dérivés de la même catégorie que la base. Mais *ultra* et *super* n'ont pas tout à fait le même sens, l'intensité étant plus grande pour le premier (*ultra-disponible versus super-disponible*). On notera que l'un comme l'autre ont également un emploi autonome, comme nom (*un ultra*) ou comme adjectif (*super*). Cela a un effet pour le dernier, car dans un certain nombre de cas, une interprétation adjectivale ou adverbiale pourrait être faite (*super-chaud, super-tendre*) amenant presque à

6. « Les sens évaluatifs – la supériorité, l'excès et le haut degré – semblent aussi résulter d'une opération de repérage et de localisation mais le repère n'est plus comme précédemment une entité concrète (localisation spatiale), un cadre institutionnel ou un jalon sur une échelle de référence (localisation non spatiale et supériorité hiérarchique) mais une norme implicite et cette norme est, dans tous les cas, dépassée. Les interprétations en termes de supériorité et d'excès sont très proches l'une de l'autre (elles mettent en jeu les mêmes catégories lexicales) mais, dans la première le dépassement est conçu comme positif : c'est « un plus » vu comme un mieux alors que dans la seconde elle est conçue comme négative : le plus est un trop. »(Amiot, 2004a, p.9)

considérer les formations comme des compositions. Les deux préfixes ne sont donc pas complètement en concurrence et leur productivité en expansion est comparable. La véritable concurrence apparaît donc plutôt entre les deux premiers et *hyper*. Ce dernier a une règle plus large : les bases ne sont pas limitées aux noms et adjectifs, mais peuvent également être des adverbes (*hyper-tard, hyper-tôt, hyper-bien, etc.*), alors que les mêmes emplois sont moins acceptables avec *ultra*. Il y aurait donc un couple *super-hyper* pour les adverbes, et une concurrence *ultra-hyper* pour les noms et les adjectifs. Notons cependant une petite différence de structure sémantique entre les deux formants : *ultra* est monosémique, tandis que *hyper* a également un sens d'évaluation quantitative (*hypermagasin, hypermarché* > *hyper*). On notera également que *méga*, qui étaient limités à des expressions techniques ou au registre relâché (langue des jeunes) semble étendre son périmètre d'application, puisqu'on recense 93 dérivés dans le corpus, non restreints à ce registre de langue. Il se combine principalement à des noms, mais aussi à des adjectifs, au contraire de *maxi*, qui se construit exclusivement avec des noms. Les deux ont toujours leur valeur adjectivale d'origine.

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.). *Sous* est d'un emploi plus restreint, de par sa valeur dominante spatiale, mais est le seul à pouvoir s'appliquer à des verbes.

	loc. spatiale	loc. non spatiale	sup. hiérar	supériorité	excès	haut degré
archi-			*archidiacre* *archiduc*			***archi-nul*** ***archi-sévère***
extra-	***extra-territorial*** ***extra-corporel***	***extra-universitaire*** ***extra-conjugal***				*extra-fin* *extra-souple*
hyper-				*hypermolécule* *hyperespace*	***hypertension*** ***hyperchlorydrie***	*hypernerveux* *hyperraffiné*
super-	*superstructure*	*supersonique* *super-léger*		***superbombardier*** ***superovulation***		*superlong* *superfin*
sur-	*surveste* *surnappe*		*surintendant* *surarbitre*		***suralimentation*** ***surévaluer***	*surexcité* *suraigu*
ultra-		*ultraviolet* *ultrason*				***ultra-libéral*** ***ultra-court***

FIGURE 7.2 – Polysémie des préfixes exprimant le haut degré (Amiot, 2004a)

La préfixation n'est pas propre à un domaine ou à un autre. La distribution des occurrences par domaine correspond à la distribution par domaine du corpus.

Un autre élément intéressant concerne le cycle de vie de ces lexies : plus gros contingent de néologismes, la préfixation propose une distribution classique des fréquences. Par contre, un trait disparaît quasiment complètement au moment d'émergence : la glose métalinguistique. Il s'agit là d'un trait partagé avec les suffixes, le sens du dérivé pour ces deux procédés restant largement compositionnel et réglé.

7.3 Suffixation

Dans le corpus Néoveille, nous avons repéré 1 639 néologismes, correspondant à 47 formes suffixales, pour 25 894 occurrences. La distribution est explicitée dans la figure 7.3. On retrouve une distribution de Zipf classique, avec la même particularité que les préfixes, avec un effectif dix fois moins important. Nous retrouvons pour ces données une moyenne d'occurrence normale de 14,65 occurrences par néologisme et une médiane à 2.

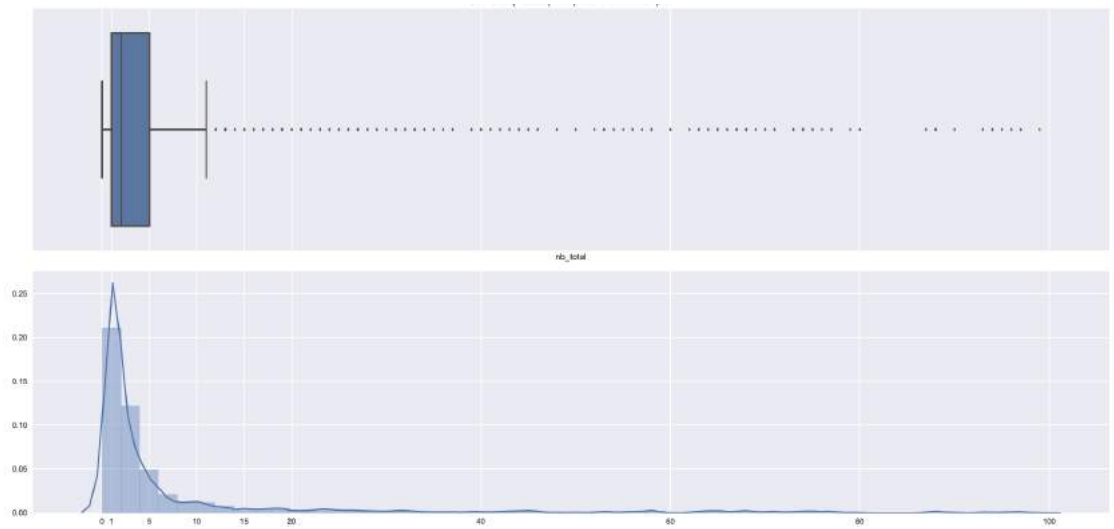


FIGURE 7.3 – Distribution des suffixes dans Néoveille

Nous présentons dans le tableau 7.2 les 20 suffixes les plus productifs sur la période.

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, montebourgisme*
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	*snapchatteur, 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, hollanderies, 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*

TABLE 7.2 – Liste des suffixes repérés dans Néoveille (31/08/2018)

La triple suffixation en *-iser, -isation et -isateur/trice* constitue la famille suffixale la plus productive dans le corpus journalistique, dénotant diverses transformations sociétales (*twitterisation, macronisation...*) et les agents de ces transformations. 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 dans notre corpus, le suffixe *-ifier* n'étant pas représenté.

Nous avons considéré que les morphèmes créateurs de verbes *-er,-ir*, d'adjectifs (*-al, é(e)(s)*, -el) et d'adverbes (*-ment*) étaient exclus des procédés néologiques. Ces suffixes constituent la frontière avec les flexions proprement dites, car ils sont réduits à l'instruction catégorielle sans instruction sémantique (vrais suffixes dérivationnels) et sans instruction de sous-catégorisation syntaxique (vrais suffixes flexionnels).

Une particularité des suffixes est de très souvent s'appliquer « en groupe »à partir d'une même base, notamment lorsque celle-ci provient d'un autre système (emprunt) ou constitue un nom propre. Les domaines sémantiques couverts par les suffixes permettent en effet de désigner les différents composants sémantiques autour d'un noyau. Prenons l'exemple des personnalités politiques (voir (Cartier, 2018c) pour plus de détails).

Les suffixes vont permettre de désigner les participants à l'action politique : les militants (*macroniste*), les sympathisants (*macronien*), avec d'éventuelles connotations (*macronieux, macronolâtre*). Le nom d'agent en *-eur/-euse* est a priori bloqué puisqu'il doit s'appliquer à une activité (verbe ou nom dérivé) (**macroneur, *sarkoz(y)eur*). Cependant, il existe un compte Twitter satirique dont le titre utilise ce nom d'agent [7]. Et sans être attestées dans notre corpus, les formations *Vallseur* et *Trumpeur* sont disponibles, avec la collision avec *valser* et *tromper*. D'autres formations plus spécifiques sont également attestées : *macronista(s), trumpista(s)*, le premier étant influencé par la diffusion du premier, désignant les sympathisantes d'origine mexicaine, puis les sympathisantes en général. Le suffixe *-ette* permet également, par analogie avec *suffragette(s)*, de désigner les militantes (avec ou sans majuscule initiale) : *trumpette(s), macronnette(s), sarkozette(s)* [8].

Les mêmes formants (*-iste, -ien, -âtre*), auxquels il faut ajouter (*-esque*) permettent de créer l'adjectif, soit relationnel (*administration/thèse/éléphant trumpiste*), soit qualificatif / prédicatif (*film trumpiste, trolls trumpistes*). Les bases nominales sur lesquelles portent ces adjectifs dénotent des rôles ou fonctions (ministre, maire, sympathisant, etc.) des organisations (administration, équipe, etc.), des actions et processus (décision, mesures, lois, etc.), les idées et sentiments (ambition, objectifs, etc.), ainsi que les noms d'évaluation (erreur, échec, réussite, etc.), traçant le périmètre élargi du champ sémantique de l'action politique. Notons également la paire de fractolexèmes *-phobe, -phile* est

7. « Nous sommes Tous à être Moi ! Macroner c'est siphonner, s'emparer de. Le Macroneur est quelqu'un qui Macrone ! »(`https://twitter.com/Palafox181`, consulté le 27/07/2018).

8. Ce suffixe diminutif issu du latin a connu deux événements qui affectent aujourd'hui son sémantisme : d'une part, la formation suffragette(s), emprunté en 1906 à l'anglais, qui permet de former, par analogie, des lexies désignant les militantes liées à la base ; d'autre part, dans les années 1990, les mesurettes, puis les balladurettes et les jupettes, désignant des « mesures (de faible portée) en faveur de l'industrie automobile », qui expliquent certains emplois de macronnette. Notons, enfin, deux autres emplois de Trumpette : pour désigner Marine Le Pen, d'une part, comme « militante numéro un », et pour désigner Donald Trump lui-même, en focalisant sur la connotation négative du suffixe (par exemple, « Tais-toi, trumpette! », `http://www.courrier-picard.fr/112589/article/2018-05-28/tais-toi-trumpette`).

également attestée pour toutes les personnalités.

On trouve également d'autres formants productifs pour désigner un territoire socio-géographique dominé par l'homme politique (fracto-composition : *Macronistan, Sarkoland*), l'univers conceptuel construit autour de la politique de l'individu (suffixation : *macronie, macronitude, macronisme, macronité*). Dans notre corpus, ces néologismes ne sont présents que pour les trois personnalités les plus populaires (*Trump, Macron, Sarkozy*). De même pour les fractolexèmes en *-logie, -logue, logiste* désignant la « science » (ou les individus s'adonnant à cette science) liée à la politique menée par la personnalité. La formation en *-ite*, désignant une affection maladive, est attestée pour toutes les personnalités. La formation en *-age* n'est attestée que pour Emmanuel Macron sur Twitter sous forme de hashtag, avec une collision avec *macaronage*[9]. La formation en *-ette*, par analogie avec *mesurette(s)* (puis *balladurettes* et *jupettes*), est attestée pour tous ceux qui ont exercé le pouvoir en France : *macro(n)nette, sarkozette, hollandette, fillonnette.*

On compte également un certain nombre d'autres formations nominales par fracto-composition: *-gate* (pour toutes les personnalités, avec élision de la voyelle ou de la syllabe finale : *sarkogate, poutingate, lepengate*, etc.), *-xit* (attesté pour l'ensemble des personnalités, mais présent dans la presse française seulement pour Macron, Sarkozy et Trump : *macronxit, trumpxit, sarkoxit*), *-economics* (*macronomics*, compocation sous influence de *trumponomics*), et *-mania* (*macro(n)mania, lepenmania, sarkomania*, etc.)

Du point de vue verbal, la conversion par le morphème *-er* est a priori bloquée pour les noms propres, à moins de sélectionner un trait typique (et généralement péjoratif) de la personne. C'est bien ce qui se passe. *Macroner* n'est pas présent dans notre corpus, mais Google indique près de 90 000 occurrences, principalement suite aux définitions proposées par Bernard Pivot sur son compte Twitter en avril 2016 :

> Macroner. Verbe irrégulier. Déf.: marcher, marcher surtout de gauche à droite. Syn.: zigzaguer. Ex.: sur l'eau Jésus macronait-il ? (`https://twitter.com/bernardpivot1/status/723371706379694080?lang=fr`)

Pour *Trump*, la flexion aboutit à un jeu de mots par collision avec *tromper* (14 occurrences dans Néoveille) : « Comment avons-nous pu nous "trumper" à ce point sur l'efficace montée du populisme aux États-Unis ? » (Huffington Post, 19/11/2016). Angela Merkel a eu droit à son verbe en allemand dès 2010 (*merkeln*, néologisme de l'année 2010[10]) dans un sens également péjoratif : « Merkeler, c'est ne pas décider, ne pas agir, tergiverser… » (Le Figaro, 31/08/2015). *Sarkoz(y)er* n'est par contre pas attesté. D'autres personnalités ont pu également être raillées par cette formation (*cahuzaquer, filloner*, etc.)

Deux formations suffixales aboutissant à des verbes sont répandues pour l'ensemble de nos personnalités : les formations en *-iser* et en *-ifier*. Les verbes résultants désignent le processus « naturel » (*-iser*) ou volontariste (*-ifier*) de transformation de la société liée à la politique menée. On notera la cohorte de suffixation en *-iser* permettant de créer le déverbal (*-isation*) et l'agent (*-isateur/isatrice*), permettant de compléter le champ.

9. `https://twitter.com/hashtag/macronage`
10. `http://www.owid.de/artikel/407474`

Pour *-ifier*, bien moins fréquent, le nom d'agent est bloqué de par ses connotations et n'a aucune attestation, même dans Google.

Concernant l'ancrage socio-pragmatique, les suffixations, comme les préfixations, n'ont aucun marquage particulier, ils sont répandus dans tous les types de discours. Leur émergence a les mêmes caractéristiques que les préfixations.

Citons également un cas de dérivation inverse, le verbe *pimper*. Depuis un siècle environ, le français ne connaît plus qu'une réalisation de la racine *pimp*, attesté pourtant dès le 12ème siècle avec un sens verbal et différents dérivés [11]. Il s'agit du participe présent *pimpant*. Jusque vers la fin du 19ème siècle, la forme verbale était attestée, sous la forme d'une expression, *pimper des prunelles*, dont la dernière attestation date de 1879, et de l'emploi intransitif *être pimpé*, marqué comme régional dans le TLFI et attesté dans Giono [12]. La radical a diffusé pendant la guerre de cent ans en Angleterre, où il s'est implanté dans des sens restés très proches de ceux attestés dans le FEW. Aujourd'hui, la forme *pimp*, en anglais [13], est attestée comme nom (signifiant 'proxénète', sans connotation), et comme verbe. avec un emploi intransitif ('agir comme un proxénète', 'faire le maquereau') et un autre transitif, marqué comme populaire, argotique ('to adapt or embellish in an ostentatious manner', 'embellir ou adapter d'une manière ostentatoire'). On trouve également, dans le même sens, *to pimp up* et *to pimp out*. On trouve par exemple des traces de ce dernier usage dans la fameuse émission de téléréalité américaine *Pimp my ride*(signifiant *Tune ta caisse*, et traduit *Pimp ton char* dans la version québecoise de l'émission) au début des années 2000 qui consistait à embellir de vieilles voitures. Mais le sens en anglais n'est pas restreint aux voitures, tout objet pouvant être ainsi transformé. On trouve dans le corpus Néoveille un certain nombre d'attestations du verbe *pimper* qui rappelle cet usage (voir figure 7.3).

Le principal emploi est bien l'emploi transitif, avec une très grande variété de compléments (*soirée, cheveu, mur, plats, missions, chaussure, déco, look, apéros, photos, ongles, etc.*). On notera un emploi réflexif (*se pimper*). On constate également que l'ensemble des contextes est lié à la presse féminine ou à des journaux populaires (20 minutes) ou à obédience anglosaxonne (Slate). Il est évident que *pimper* est un emprunt adapté à l'anglais : le corpus NOW [14] comprend 1126 occurrences avec les mêmes emplois et la même extension de compléments. La pénétration en français est d'autre part limitée à une presse encline à l'utilisation d'emprunts. Cependant, étant donné l'historique du radical, resté en français sous la seule forme participiale, mais forme toujours vivace, rend le radical toujours disponible, et on peut soutenir une analyse du néologisme comme dérivation inverse, ou en tout cas un mélange d'influence étrangère et de réactivation du

11. Voir l'entrée *pimp-* dans le Französisches Etymologisches Wörterbuch (FEW) : `https://apps.atilf.fr/lecteurFEW/index.php/page/lire/e/13662`

12. `http://stella.atilf.fr/Dendien/scripts/tlfiv5/advanced.exe?8;s=3242097540;`

13. Nous utilisons ici la définition du Collins : `https://www.collinsdictionary.com/dictionary/english/pimp`. On pourra consulter également le *American Heritage Dictionary* qui explicite les mêmes définitions, ainsi que la version anglaise du Wiktionary, qui propose encore plus de sens, la plupart liés, selon les auteurs, à des emplois populaires afrcano-américains `https://en.wiktionary.org/wiki/pimp`.

14. Ce corpus comprend des pages web anglo-saxonnes analysées depuis 2010 : `https://corpus.byu.edu/now/`

Date	Journal	Extrait
14/06/16	grazia.fr	*...ez vos meilleures combinaisons avec notre sélection de nuisettes pour pimper vos journées !...*
26/06/16	cosmopolitan.fr	*...fort. Réfléchissez donc à un moyen de sortir de la vôtre. Est-ce de " pimper " vos missions au sein de votre boite ? Ou, au contraire, de changer ...*
06/07/16	grazia.fr	*...Et si l'on choisit une option sobre, on peut toujours pimper sa chaussure avec des lacets colorés ou encore de jolies languettes. ...*
20/07/16	elle.fr	*...Objectif n3 : pimper la déco de sa maison de location (qui s'est réveillée hideuse)....*
19/08/16	cosmopolitan.fr	*...atienter jusqu'au prochain vrai shampoing. Et mettre un headband pour pimper son look en attendant !...*
21/08/16	huffingtonpost.fr	*...Cette recette de feta marinée va pimper tous vos sandwichs et apéros...*
09/09/16	cosmopolitan.fr	*... convoités, vous trouverez Mayfair, Clarendon ou encore Lark. De quoi pimper vos photos et en faire de vraies petites bombes de couleurs et d'effe...*
28/09/16	cosmopolitan.fr	*...plus jeunes, on avait tendance à utiliser tout et n'importe quoi pour pimper nos ongles. A commencer par le contenu de nos trousses : Stabilos, b...*
27/10/16	cosmopolitan.fr	*...C'est un peu ça, l'astuce de Kate : pimper le classique pour le moderniser, mais sans jamais en faire trop. Car ...*
08/11/16	la-croix.com	*...Un Acadien se pimpe, pour s'habiller avec élégance....*
27/11/16	huffingtonpost.fr	*...Comment "pimper" votre chocolat chaud ?...*
02/12/16	cosmopolitan.fr	*... Un tube de fromage fondu pour pimper vos plats...*
07/12/16	grazia.fr	*...Les fêtes de fin d'année sont l'occasion idéale pour " pimper " une coiffure. Pour une fois dans l'année, vous pouvez tout vous per...*
08/01/17	cosmopolitan.fr	*...Pour pimper votre queue de cheval, vous pouvez utiliser des élastiques colorés ou...*
02/03/17	elle.fr	*...Clairement, les stickers sont à bannir ! Pour pimper vos murs, pensez à la frise, au lé de papier-peint, à la décoration m...*
22/03/17	grazia.fr	*...bles de cet été. Pour vous, Grazia a repéré 5 pièces accessibles pour pimper votre garderobe....*

TABLE 7.3 – Contextes d'emploi de *pimper*

radical [15].

15. On consultera également l'article du Figaro retraçant l'historique de cette racine : `http://www.lefigaro.fr/langue-francaise/expressions-francaises/2017/12/03/37003-20171203ARTFIG00031-allez-vous-pimper-votre-sapin-de-noel.php` et la note de l'Académie française : `http://www.academie-francaise.fr/pimper-les-legumes-anciens`.

Chapitre 8

Composition en français contemporain (2015-2018)

Sommaire

8.1 Définitions

(Sablayrolles et Pruvost, 2016) distinguent la composition proprement dite et la composition par amalgame. La composition proprement dite entre lexies comprend quatre sous-classes : la composition « simple » (entre deux lexies : *arbre-feuille, attrape-mouche*), 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 + composition, terme forgé par (Cusin-Berche, 1999), *héliport : hélicoptère>héli et aéroport>port*), la fracto-composition [1] (combinaison de deux lexies dont la première est dans une forme liée, *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 : *optipessimisme*).

1. sur la base de fractolexèmes, également appelés quasi-lexèmes ou quasi-préfixes (voir chapitre 3 et (Renner, 2015) pour une revue de ce concept)

8.2 Composition simple

Dans le corpus Néoveille, nous avons repéré 1 410 néologismes pour 12 497 occurrences. Leur distribution en fréquence est très particulière, puisque 692 ont une seule occurrence (49%), et, en calculant les quartiles (25%, 50%, 75%), on obtient une boîte à moustache et une distribution très écrasée (voir figure 8.1). Cela dénote une particularité des composés simples (et des autres composés), à savoir d'être très souvent des formations ponctuelles, des *nonce-word* créés pour une occasion, et qui ne seront pas repris. La moyenne d'occurrences est de 8,59 et la médiane est située à 2. La dispersion reste importante (890), étant donné, d'une part, que nous avons conservé des néologismes moins récents, afin de suivre leur cycle de vie et, d'autre part, certaines formations ont eu un succès rapide très important (*macron-compatible, positive-attitude, etc.*).

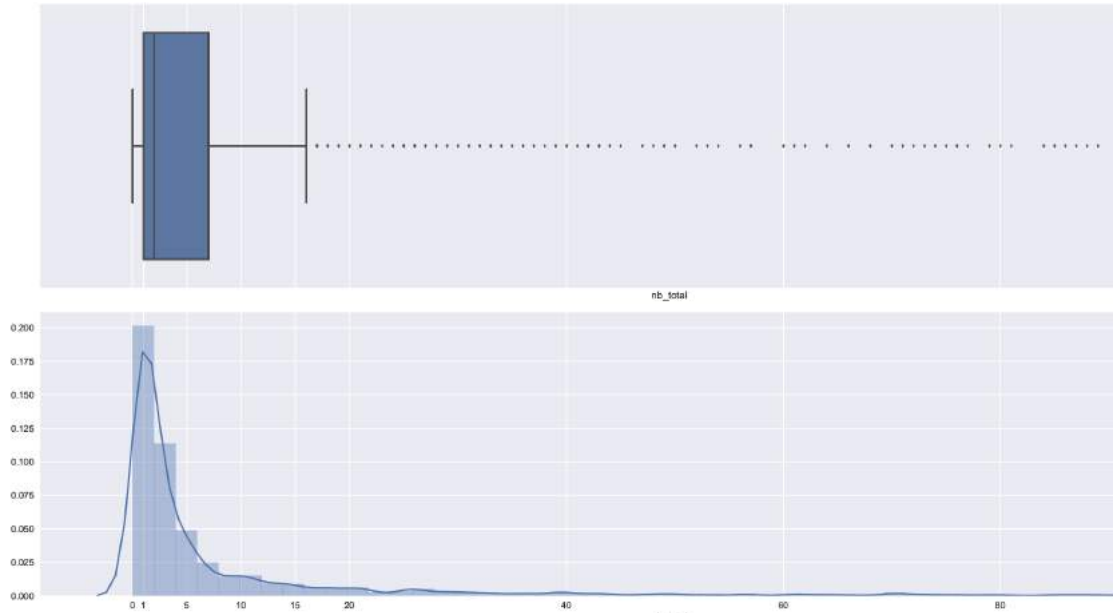


FIGURE 8.1 – Distribution des composés simples dans Néoveille

La distribution en terme de parties du discours du dérivé est restreinte, puisque 87% sont des noms, et 13% des adjectifs. Il n'existe en français pas de possibilité de créer des verbes par ce procédé, sauf à utiliser une suffixation sur le résultat d'une composition nominale ou adjectivale (*tire-bouchonner*).

Le tableau 8.1 présente la répartition par schéma syntaxique.

Ces données appellent plusieurs commentaires quantitatifs:

— Parmi l'ensemble des schémas syntaxiques des composés, assez variés (14), le schéma N-N (ou plus rarement NN) est majoritaire de façon écrasante (81%) ; Suivent des schémas à tête adjectivale (ADJ-ADJ) et verbale (V-N), puis deux autres schémas à tête nominale (ADj-N et N-ADJ) et prépositionnelle (Prep-N) ;

— Seules deux catégories sont générées par les schémas des composés : des noms, de façon également écrasante (en nombre d'occurrences produites mais également

Schéma syntax.	Cat. résultat	Nbre de néol. diff.	Nbre total d'occ.	Exemples
N(-)N	N	1149	8452	*concept-car': 37, 'tram-trains': 29, 'chien-robot': 22, 'smartphone-phare': 22, 'macronleaks': 20, 'art-thérapeute': 15, 'loi-travail': 5, 'nuit-debout': 3, 'beauf-attitude': 2, 'mélenchonsphère': 2, 'serial-graffeur': 1, 'twittomanie': 1, 'costumegate': 1, 'drone-livreur': 1, 'baby-athlétisme': 1, 'mediabashing': 1, 'actrice-youtubeuse': 1, 'filiation-compatible': 1, 'robot-livreur': 1, 'rybkagate': 1*
ADJ-ADJ	ADJ	105	1451	*'morts-amoureux': 6, 'radical-compatible': 4, 'social-populiste': 3, 'libéral-universaliste': 1, 'dieselo-dépendant': 1,'social-patriotique': , 'pop-rétro': 1, 'médiatico-boursière': 1, 'nationaliste-isolationniste': 1, 'juridico-technique': 1, 'électro-alternatifs': 1*
V-N	N	41	114	*attrape-touriste': 28, 'vide-jardins': 4, 'croque-cake': 2, 'porte-additions': 2, 'passe-miroir': 2, 'porte-bières': 2, 'pare-ballons': 1, 'attrape-œil': 1, 'trouble-sommeil': 1, 'lance-gaz': 1, 'redresse-paupières': 1, 'couvre-bottes': 1, 'pisse-debout': 1*
ADJ-N	N	39	481	*'social-réformiste': 17, 'social-réformisme': 14, 'social-souverainiste': 10, 'sexy-attitude': 7, 'ecolo-business': 4, 'national-catholicisme': 4, 'national-communiste': 2, 'green-attitude': 1, 'social-entrepreneur': 1, 'social-chauviniste': 1*
N-ADJ	N ou ADJ	32	1148	*macron-compatibles': 459, 'macron-compatible': 66, 'apprenti-jihadiste': 8, 'euro-compatible': 7, 'charia-compatible': 3, 'lepéno-compatible': 1*
PREP-N	N	17	105	*sans-portable': 35, 'après-concert': 23, 'après-vol': 3, 'contre-capitalisme': 1, 'arrière-festival': 1, 'contre-indicateur': 1, 'contre-moraliste': 1, 'contre-féministe': 1*
ADV-N	N	10	154	*tout-voiture': 61, 'tout-mobile': 13, 'mal-fonctionnement': 10, 'tout-hôpital': 6, 'tout-marchand': 3*
ADV-ADJ	N ou ADJ	5	327	*tout-inclus': 193, 'mal-inscrits': 19*
N-PREP-N	N	4	61	*bac-à-sable': 38, 'diplomate-en-chef': 19, 'football-pour-tous': 3, 'cul-en-zone': 1*
ADJ-PREP-Vinf	ADJ	3	115	*prêt-à-penser': 60, 'prêts-à-gaver': 49, 'prêt-à-gaver': 6*
V-ADV	N	2	61	*vivre-ensemble': 36, 'savoir-vivre-ensemble': 25*
V-Vinf-ADJ	N	1	1	*savoir-être-seul': 1*
N-PREP-N	N	1	19	*diplomate-en-chef': 19*
ADV-PREP-N	N	1	8	*tout-en-anglais': 8*

TABLE 8.1 – Schémas syntaxiques productifs en composition

en nombre de schémas productifs), et des adjectifs (seulement trois schémas productifs : ADJ-ADJ, N-ADJ et ADV-ADJ) ;

— Les composés générés sont plus ou moins fréquents, selon les schémas: la moyenne d'occurrences par schéma oscille entre 1 (V-Vinf-Adj, mais avec un seul néologismes) et 63 (ADV-ADJ, pour cinq unités). La moyenne du schéma N-ADJ (35) est également tronquée, à cause de deux variantes de *macron-compatible* très fréquentes. Les autres schémas ont une fréquence moyenne et une médiane très proches (entre 5 et 10).

Nous passons maintenant à un commentaire plus qualitatif sur le schéma N(-)N. Ce schéma dénote des combinaisons variées (*médecin-nutritioniste, médecin-régulateur, instagrammeuse-bloggueuse, scientifique-justicier, journaliste-confesseur, poisson-sanglier, veste-électrocardiogramme, bus-cuisine*, etc.).

La relation sémantique entre les deux noms, en reprenant la distinction entre relation subordonnée et coordinative ((Arnaud et Renner, 2014)) est le plus souvent coordinative [2], lorsque la catégorie sémantique des deux noms est identique (*raciste-terroriste, piscines-terrasses, skippeur-journaliste, consommateur-contribuable, artiste-botaniste*). On rencontre plus rarement la relation subordinative, soit dans une rela-

2. Pour s'en assurer, on peut utiliser le test proposé par (Arnaud et Renner, 2014) : *un chasseur-cueilleur est un chasseur,* et *un chasseur-cueilleur est un cueilleur.* En cas d'échec du test, nous sommes dans un cas de subordination : *un pneu neige est un pneu* mais **un pneu neige n'est pas une neige.*

tion attributive (*cravate-ficelle,ciné-spectacle, maison-jardin, etc.* [3]), soit plus rarement dans une relation prédicative (*livre-oreiller, mercredi-loisirs, sœur-terreur, loi-travail, bénéfices-santé, action-réseau, emploi-rebond ect.*).

Plus étonnant, on rencontre dans le corpus nombre de construction où la tête est non pas à droite [4] mais à droite. Cela concerne essentiellement des composés comprenant un emprunt (*fashion-victim, serial-graffeur, concept-car, etc.*) mais également des schémas influencés par la structure anglaise avec la tête à droite (*rebelle-attitude, art-thérapeute, macron-compatibilité*).

Lexies productives. On pourrait s'attendre à ce que, pour les composés, la productivité ne puisse s'appliquer qu'aux schémas syntaxiques des composants. Mais il existe également un certain nombre de lexies qui ont une productivité indéniable. Nous présentons dans le tableau 8.2 les lexies les plus productives à gauche, et dans le tableau 8.3 les lexies les plus productives à droite [5].

Pour ce qui concerne les lexies à gauche, 238 ont une productivité non nulle. 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 schèmes 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-liker, prêt-à-agir*, etc.

8.3 Composition savante et hybride

Ces composés sont beaucoup plus rares : 68 composés savants pour 479 occurrences, avec une moyenne de 7 occurrences et 33 composés hybrides, pour 213 occurrences et une moyenne de 6 occurrences. La distribution ne présente pas d'intérêt étant donné ce faible nombre. On remarque là encore quelques formants productifs (tableau 8.4) :

8.4 Fracto-composition

La fracto-composition est un procédé non-négligeable, puisque nous en avons identifié 791 (soit 3,52% du total), pour 7 039 occurrences (soit 0,97% du total), avec une moyenne de 9 occurrences. La distribution fréquentielle est présentée dans la figure 8.2.

3. On peut utiliser pour s'assurer qu'il s'agit bien d'une attribution du test analogique proposé par (Arnaud et Renner, 2014) : *une cravate-ficelle est une cravate qui ressemble à une ficelle.* Il s'agit ici d'une relation métaphorique où l'on sélectionne un ou plusieurs traits du subordonné. Dans les cas de subordination relationnelle, le test ne fonctionne pas et on doit effectuer une prédiction pour reconstituer le sens du composé :

4. Rappelons que, dans la composition et dans les syntagmes nominaux, l'ordre canonique en français,

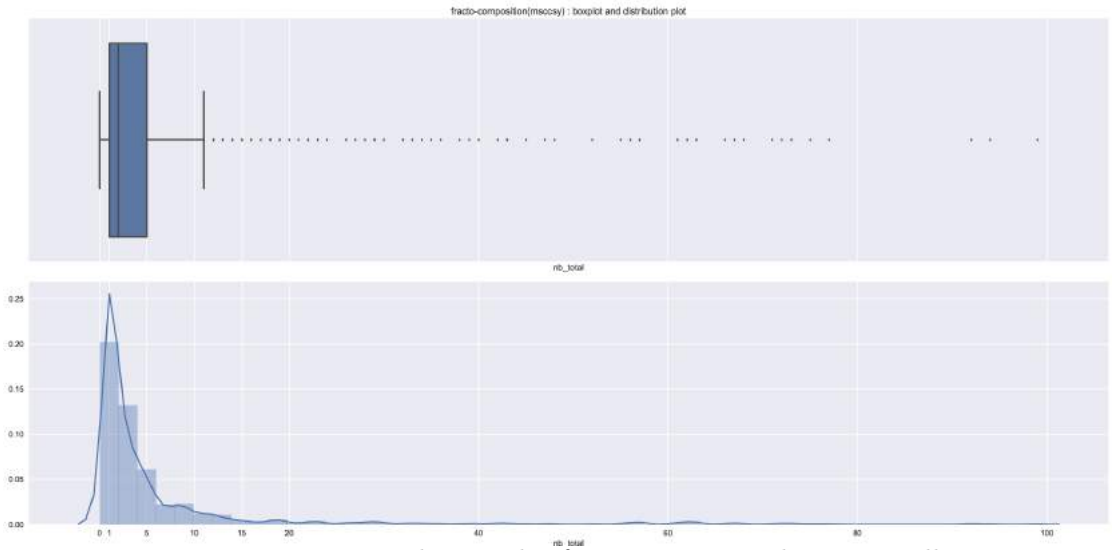


FIGURE 8.2 – Distribution des fracto-composés dans Néoveille

On retrouve ici une distribution proche de celle des préfixes, avec les premiers quartiles assez minces, et un nombre non-négligeable de lexies ayant une forte fréquence. Cette impression initiale est confirmée si l'on regarde les fracto-lexèmes les plus productifs de cette catégorie (tableau 8.5).

On constate que leur productivité est comparable à celle des préfixes, ce qui confirme leur statut d'affixoïdes. Parmi ces affixoïdes, on remarque également, pour l'ensemble d'entre eux, la possibilité de construction avec des noms, des verbes et des adjectifs. Certains formants peuvent se construire avec ou sans trait d'union avec la base lexicale. Ils ont également la particularité formelle d'être limités à deux syllabes. Le sens est compositionnel. On notera enfin que certaines lexies produites ont une forte fréquence, signe de leur implantation dans le lexique.

8.5 Compocation

83 compocations ont été relevées dans Néoveille. Certaines sont moins récentes (*rançongiciel, hackathon, infotainment, combishort, burqini, gréviculteurs, trotscoot, énergiculteur, gangsterrorisation,blogistador*), d'autres liées à une actualité ou une mode récente (*frexit, instameal, twictée, animour, instapreneurs,smombies, trotscoot,bobopulisme, mammouphant, histotainment, énergiculteur, gangsterrorisation, esthéduction, twitcheuse, aoûthlétisme, macronpoly, volontourisme, cataflics, mockumentaire*).

Prenons le cas de *Frexit*, l'une des productions du suffixoïde *-exit*, dont la première occurrence date du *Grexit* en 2012, terme forgé par des économistes pour prévoir les conséquences de la sortie de la Grèce de la zone Euro. Dès 2012, le suffixe suscite nombre

comme dans l'ensemble des langues romanes, est d'avoir la tête à gauche, et le subordonné à droite.

5. Nous avons tenu compte de l'ensemble des composés pour effectuer ces calculs.

de néologismes par compocation : *Spexit, Italexit, Swexit, etc.*. Les premières (très rares) mentions de *Frexit* datent de cette période, mais c'est à partir de 2015-2016 que l'emploi explose, avec le *Brexit*, et la campagne présidentielle en France qui débute dès l'été 2016. Les figures 8.3, 8.4, 8.5, 8.6 et 8.7 présentent les différentes informations disponibles sur la plateforme Néoveille depuis cette date.

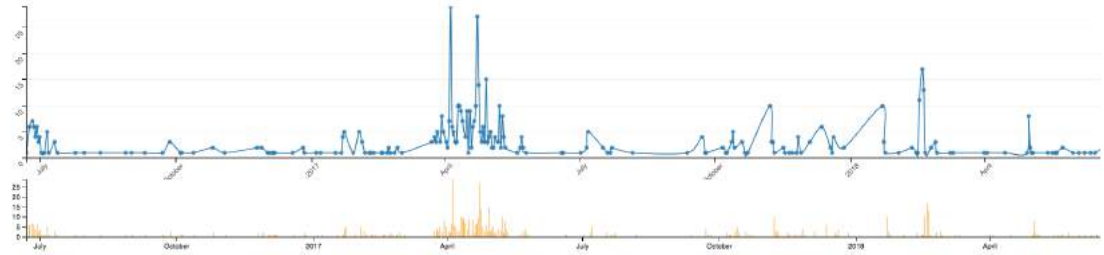


FIGURE 8.3 – Distribution temporelle de *Frexit* de 2016 à 2018

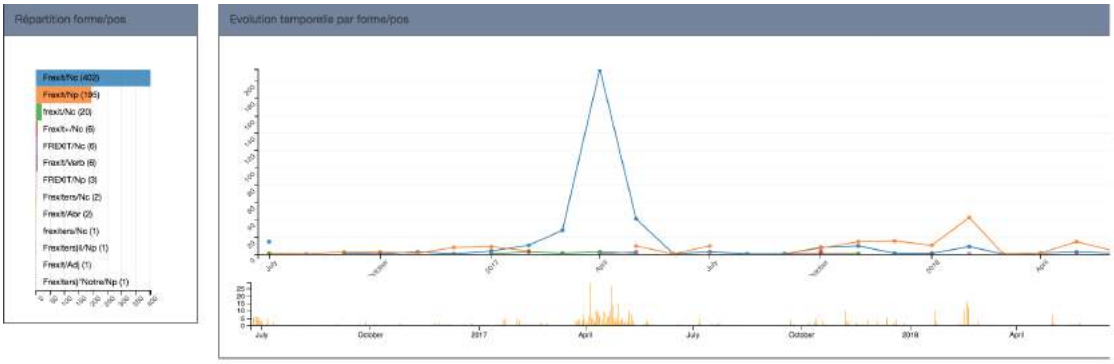


FIGURE 8.4 – Famille morphologique de *Frexit* de 2016 à 2018

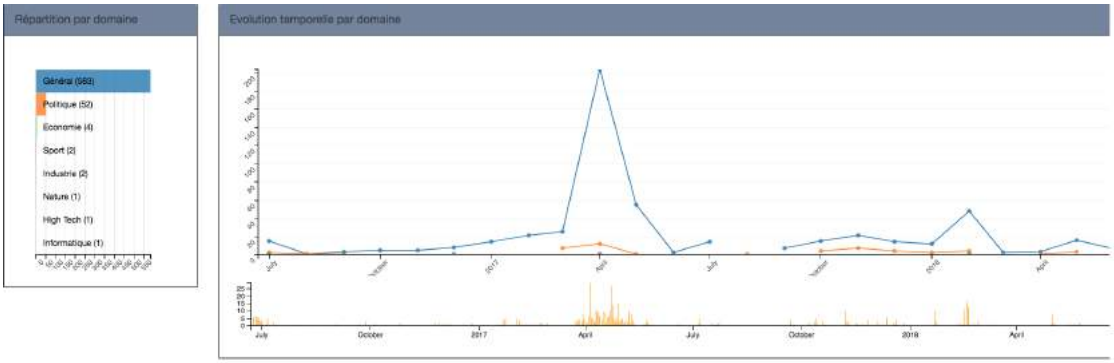
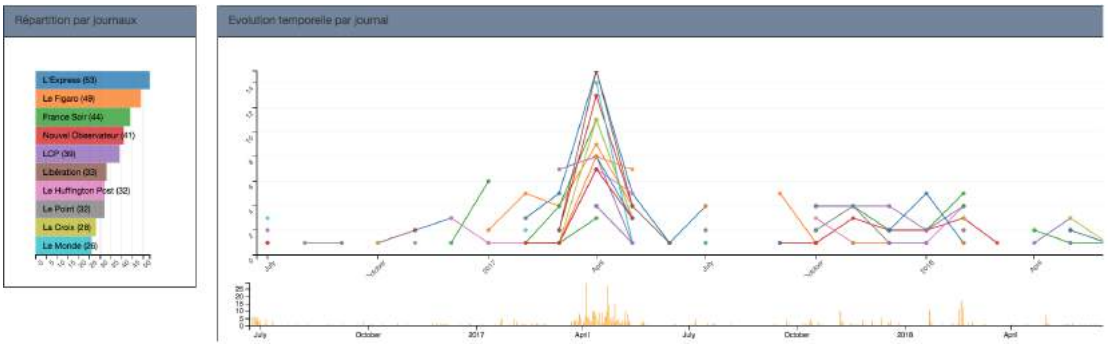


FIGURE 8.5 – Distribution temporelle par domaine de *Frexit* de 2016 à 2018

On constate tout d'abord qu'un pic d'usage a bien eu lieu durant la campagne présidentielle, spécialement durant l'entre-deux tours, la sortie de l'Europe étant l'un des thèmes de campagne de Marine Le Pen. Les emplois se sont ensuite tassés, même si on



Date	Journal	Domaine	Extrait
2018-09-25 00:00:00	Le Huffington Post	Général	...Frexit ...
2018-09-25 00:00:00	Nouvel Observateur	Général	..." Interrogé sur l' idée d' un " Frexit " , ou départ de la France de l' UE ...
2018-09-25 00:00:00	20 Minutes	Général	...Interrogé sur l' idée d' un " Frexit " , ou départ de la France de l' UE ...
2018-09-25 00:00:00	Libération	Général	...En revanche , pas question d' évoquer un Frexit ...
2018-09-24 00:00:00	L'Usine Nouvelle	Industrie	...Lire l' article Au-delà du Brexit , les tentations du Frexit ou du Swexit (la sortie de la Suède) ...
2018-09-24 00:00:00	L'Usine Nouvelle	Industrie	...eau du bain Au-delà du Brexit , les tentations du Frexit ou du Swexit (la sortie de la Suède) ...
2018-09-21 00:00:00	Libération	Général	...plus de fédéralisme comme Macron et ceux qui veulent un Frexit , il y a une place pour une droite patriote ...
2018-09-21 00:00:00	Le Figaro	Général	...ni une " Europe fédérale " , ni un " Frexit " ...
2018-09-18 00:00:00	Courrier International	Général	...ne dédaignait pas alors évoquer l' idée d' un " Frexit " , puisse un jour vanter l' idée du continent ...
2018-09-17 00:00:00	BFM TV	Général	..., alors qu' il vient de sortir un livre intitulé Frexit : ue : en sortir pour s' en sortir(édition l' ...
2018-09-17 00:00:00	Le Huffington Post	Général	...Frexit ...
2018-09-17 00:00:00	Marianne	Général	...Frexit ...
2018-09-17 00:00:00	Le Figaro	Général	...France , imaginez ce que serait la coût d' un frexit ...
2018-09-15 00:00:00	France Soir	Général	...la sortie de la France de l' Union européenne (Frexit) ...

FIGURE 8.6 – Distribution temporelle par journal de *Frexit* de 2016 à 2018 et exemples de contextes

Date	Journal	Domaine	Extrait
2018-09-25 00:00:00	Le Huffington Post	Général	...Frexit ...
2018-09-25 00:00:00	Nouvel Observateur	Général	..." Interrogé sur l' idée d' un " Frexit " , ou départ de la France de l' UE ...
2018-09-25 00:00:00	20 Minutes	Général	...Interrogé sur l' idée d' un " Frexit " , ou départ de la France de l' UE ...
2018-09-25 00:00:00	Libération	Général	...En revanche , pas question d' évoquer un Frexit ...
2018-09-24 00:00:00	L'Usine Nouvelle	Industrie	...Lire l' article Au-delà du Brexit , les tentations du Frexit ou du Swexit (la sortie de la Suède) ...
2018-09-24 00:00:00	L'Usine Nouvelle	Industrie	...eau du bain Au-delà du Brexit , les tentations du Frexit ou du Swexit (la sortie de la Suède) ...
2018-09-21 00:00:00	Libération	Général	...plus de fédéralisme comme Macron et ceux qui veulent un Frexit , il y a une place pour une droite patriote ...
2018-09-21 00:00:00	Le Figaro	Général	...ni une " Europe fédérale " , ni un " Frexit " ...
2018-09-18 00:00:00	Courrier International	Général	...ne dédaignait pas alors évoquer l' idée d' un " Frexit " , puisse un jour vanter l' idée du continent ...
2018-09-17 00:00:00	BFM TV	Général	..., alors qu' il vient de sortir un livre intitulé Frexit : ue : en sortir pour s' en sortir(édition l' ...
2018-09-17 00:00:00	Le Huffington Post	Général	...Frexit ...
2018-09-17 00:00:00	Marianne	Général	...Frexit ...
2018-09-17 00:00:00	Le Figaro	Général	...France , imaginez ce que serait la coût d' un frexit ...
2018-09-15 00:00:00	France Soir	Général	...la sortie de la France de l' Union européenne (Frexit) ...

FIGURE 8.7 – Exemples de contextes pour *Frexit* de 2016 à 2018

constate son retour de manière faible mais régulière. On constate que le terme, jusqu'à 2016 cantonné à la presse économique, passe durant la campagne présidentielle dans le domaine général, avec une grande diversité de journaux. On note également la formation

de dérivés (frexiteur(s)), ainsi que l'apparition d'une forme en minuscule. Les exemples montrent également que le terme est employé parfois par mention, parfois directement sans mention, et sans glose. Il est probable que le terme disparaisse dans les années à venir, étant donné sa forte liaison avec une situation historique, et son statut de nom propre. Mais la règle de formation créée par la compocation à base de la suffixation en *-exit* pourrait par contre avoir un sort tel que celui qui a été fait à *-gate*.

A l'inverse, de nombreuses compocations sont des *nonce-words*. On notera *bobopulisme* terme apparu après la déclaration de candidature d'Emmanuel Macron à l'élection présidentielle début 2017, créé par ses adversaires politiques pour désigner deux traits de son caractère, qui n'a pas eu la popularité escomptée, même s'il reparaît de temps à autres.

8.6 Mot-valisation

On compte 15 mots-valises dans le corpus Néoveille. La totalité sont des hapax ou des quasi-hapax : *mardigital, talibanlieusard, yogapero, pandattitude, etc.*.

Lexie	Total documents	Total occurrences	Eléments uniques	Liste éléments-fréquence
social	93	96	14	*{'social-populiste': 3, 'social-entrepreneur': 1, 'social-écologique': 8, 'social-chauviniste': 1, 'social-collabo': 1, 'social-dirigiste': 1, 'social-médéfisme': 1, 'social-réformisme': 14, 'social-souverainisme': 1, 'social-conservateur': 1, 'social-libérale': 35, 'social-réformiste': 17, 'social-souverainiste': 11, 'social-patriotique': 1}*
livre	170	176	13	*{'livre-ordinateur': 1, 'livre-fondateur': 7, 'livre-révélation': 2, 'livre-oreiller': 2, 'livre-bilan': 11, 'livre-programme': 74, 'livre-déballage': 2, 'livre-jeux': 1, 'livre-univers': 3, 'livre-tuteur': 1, 'livre-spectacle': 1, 'livre-choc': 51, 'livre-photo': 20}*
écrivain	111	124	12	*{'écrivain-photographe': 2, 'écrivain-réalisateur': 2, 'écrivain-voyageuse': 3, 'écrivain-voyageur': 63, 'écrivain-journaliste': 11, 'écrivain-essayiste': 1, 'écrivain-économiste': 1, 'écrivain-scénariste': 1, 'écrivain-éditeur': 5, 'écrivain-pianiste': 1, 'écrivain-aviateur': 30, 'écrivain-chroniqueur': 4}*
tout	387	402	11	*{'tout-droit': 48, 'tout-pétrole': 13, 'tout-voiture': 63, 'tout-images': 38, 'tout-suspendu': 7, 'tout-inclus': 201, 'tout-flash': 2, 'tout-en-anglais': 8, 'tout-mobile': 13, 'tout-hôpital': 6, 'tout-marchand': 3}*
robot	37	38	11	*{'robot-sommelier': 2, 'robot-explorateur': 3, 'robot-aspirateur': 11, 'robot-journaliste': 1, 'robot-compagnon': 10, 'robot-mannequin': 4, 'robot-pasteur': 1, 'robot-cuisinier': 2, 'robot-livreur': 1, 'robot-poisson': 2, 'robot-plongeur': 1}*
médecin	170	173	11	*{'médecin-nutritionniste': 55, 'médecin-chirurgien': 1, 'médecin-acupuncteur': 7, 'médecin-biologiste': 10, 'médecin-régulateur': 22, 'médecin-hygiéniste': 7, 'médecin-anesthésiste': 53, 'médecin-explorateur': 2, 'médecin-directeur': 9, 'médecin-blogueur': 1, 'médecin-animateur': 6}*
national	38	40	10	*{'national-communiste': 2, 'national-conservatrice': 1, 'national-populiste': 16, 'national-économisme': 1, 'national-conservateur': 9, 'national-protectionniste': 2, 'national-catholicisme': 4, 'national-chrétien': 2, 'national-catholique': 1, 'national-islamiste': 2}*
acteur	53	54	9	*{'acteur-marionnettiste': 2, 'acteur-réalisateur': 21, 'acteur-spectateur': 1, 'acteur-metteur': 1, 'acteur-vulgarisateur': 1, 'acteur-rappeur': 1, 'acteur-catcheur': 5, 'acteur-humoriste': 4, 'acteur-phare': 18}*
artiste	33	34	8	*{'artiste-botaniste': 1, 'artiste-tatoueur': 25, 'artiste-chercheur': 1, 'artiste-agitateur': 1, 'artiste-marionnettiste': 1, 'artiste-entrepreneur': 2, 'artiste-pochoiriste': 1, 'artiste-explorateur': 2}*
apprenti	61	63	8	*{'apprenti-voleur': 1, 'apprenti-chanteur': 10, 'apprenti-jihadiste': 8, 'apprenti-djihadiste': 33, 'apprenti-guitariste': 2, 'apprenti-dictateur': 5, 'apprenti-archéologue': 3, 'apprenti-coiffeur': 1}*
contre	72	75	7	*{'contre-favorable': 63, 'contre-capitalisme': 1, 'contre-populisme': 2, 'contre-indicateur': 1, 'contre-mobilisation': 6, 'contre-moraliste': 1, 'contre-féministe': 1}*
journaliste	25	25	7	*{'journaliste-réalisateur': 10, 'journaliste-animateur': 5, 'journaliste-fondateur': 5, 'journaliste-infographiste': 1, 'journaliste-activiste': 1, 'journaliste-publiciste': 1, 'journaliste-confesseur': 2}*
maître	46	48	7	*{'maître-queue': 1, 'maître-démolisseur': 1, 'maître-serviteur': 1, 'maître-assembleur': 5, 'maître-restaurateur': 33, 'maître-descendeur': 1, 'maître-composteur': 6}*
euro	21	22	7	*{'euro-réformiste': 3, 'euro-compatible': 8, 'euro-libéralisme': 1, 'euro-mondialisme': 1, 'euro-réformisme': 6, 'euro-socialiste': 2, 'euro-mondialisation': 1}*
chef	140	161	7	*{'chef-économiste': 52, 'chef-restaurateur': 8, 'chef-prévisionniste': 5, 'chef-négociateur': 79, 'chef-accessoiriste': 8, 'chef-fondateur': 8, 'chef-logisticien': 1}*
bus	72	72	6	*{'bus-relais': 24, 'bus-bureau': 2, 'bus-abri': 1, 'bus-tunnel': 2, 'bus-macrons': 33, 'bus-tram': 10}*
fashion	59	59	6	*{'fashion-victims': 27, 'fashionsphère': 1, 'fashion-champagne-networking': 1, 'fashion-erreur': 2, 'fashion-victim': 27, 'fashion-duel': 1}*
comédien	23	24	6	*{'comédien-scénariste': 2, 'comédien-metteur': 4, 'comédien-réalisateur': 4, 'comédien-auteur': 3, 'comédien-producteur': 3, 'comédien-humoriste': 8}*
après	38	39	6	*{'après-nazisme': 1, 'après-séisme': 6, 'après-franquisme': 3, 'après-vol': 3, 'après-concert': 24, 'après-ubérisation': 2}*
double	54	55	6	*{'double-analyse': 3, 'double-menton': 19, 'double-primes': 4, 'double-ceinture': 1, 'double-arrivée': 3, 'double-finaliste': 25}*
sociale	31	50	6	*{'sociale-écologiste': 27, 'sociale-centriste': 1, 'sociale-souverainiste': 16, 'sociale-réformiste': 4, 'sociale-darwiniste': 1, 'sociale-protectionniste': 1}*
libéral	10	10	6	*{'libéral-laissez-faire': 1, 'libéral-gaulliste': 2, 'libéral-universaliste': 1, 'libéral-réformisme': 1, 'libéral-démocratique': 4, 'libéral-étatisme': 1}*
homme	58	62	6	*{'homme-nation': 4, 'homme-fusée': 43, 'homme-poisson': 11, 'homme-bouc': 2, 'homme-monde': 1, 'homme-gomme': 1}*

TABLE 8.2 – Lexies productives à gauche dans les composés simples

Lexie	Total documents	Total occurrences	Eléments uniques	Exemples
phare	1146	1169	53	{'actrice-phare': 6, 'médicament-phare': 6, 'idée-phare': 91, 'assophare': 1, 'association-phare': 9, 'loi-phare': 13, 'marque-phare': 43, 'arguments-phare': 11, 'division-phare': 10, 'projet-phare': 33, 'course-phare': 17, 'année-phare': 2, 'nation-phare': 13, 'concertphare': 5, 'feuilleton-phare': 4, 'restaurant-phare': 1}
choc	949	1068	45	{'match-choc': 4, 'étude-choc': 3, 'recrutement-choc': 1, 'accessoire-choc': 4, 'affiche-choc': 3, 'déclaration-choc': 65, 'enquête-choc': 16, 'victoire-choc': 23, 'vidéos-choc': 2, 'récitchoc': 3, 'ouvrage-choc': 4, 'démission-choc': 138, 'propositionschoc': 33, 'groupe-choc': 1, 'article-choc': 4, 'vidéo-choc': 48}
phares	246	254	23	{'valeurs-phares': 22, 'programmes-phares': 10, 'sprinteursphares': 1, 'produits-phares': 8, 'entreprises-phares': 9, 'expositions-phares': 3, 'pièces-phares': 16, 'figures-phares': 10, }
sphère	57	57	19	{'foodosphère': 4, 'politosphère': 3, 'infosphère': 1, 'géosphère': 2, 'parentosphère': 1, 'youtubosphère': 1, 'mélenchosphère': 4, 'jihadosphère': 7, 'crêposphère': 1, 'podcastosphère': 2, 'laïcosphère': 1}
clé	116	120	12	{'portefeuille-clé': 8, 'test-clé': 3, 'alliance-clé': 3, 'sujets-clé': 35, 'groupe-clé': 6, 'vote-clé': 1, 'logiciel-clé': 1, 'créneau-clé': 1, 'instants-clé': 5, 'sujet-clé': 54, 'geste-clé': 2, 'substance-clé': 1}
star	141	142	11	{'youtubeur-star': 28, 'quarterback-star': 5, 'égéries-star': 3, 'boxeur-star': 1, 'produit-star': 59, 'blogueuse-star': 3, 'footballeur-star': 15, 'architecte-star': 23, 'médicament-star': 1, 'e-star': 1, 'ado-star': 3}
compatible	87	96	9	{'radical-compatible': 4, 'naturocompatible': 1, 'bobocompatible': 3, 'euro-compatible': 8, 'macron-compatible': 74, 'lepénocompatible': 1, 'bio-compatible': 1, 'filiation-compatible': 1, 'charia-compatible': 3}
gate	113	129	8	{'smartphonegate': 3, 'penelopegate': 67, 'monkeygate': 2, 'filtergate': 9, 'costumegate': 1, 'tobaccogate': 2, 'couscousgate': 44, 'rybkagate': 1}
clés	93	111	8	{'atouts-clés': 1, 'critères-clés': 12, 'dossiers-clés': 55, 'donnéesclés': 13, 'fonctions-clés': 10, 'projets-clés': 15, 'accroche-clés': 1, 'valeurs-clés': 4}
attitude	167	169	6	{'beauf-attitude': 2, 'écoattitude': 1, 'green-attitude': 1, 'rebelleattitude': 3, 'positive-attitude': 155, 'sexy-attitude': 7}
éclair	32	33	6	{'progression-éclair': 7, 'casse-éclair': 2, 'renchérissement-éclair': 1, 'indépendance-éclair': 1, 'célébrité-éclair': 2, 'ascension-éclair': 20}
surprise	100	107	6	{'concert-surprise': 26, 'démission-surprise': 57, 'libérationsurprise': 1, 'issue-surprise': 2, 'titre-surprise': 11, 'annulationsurprise': 10}
robot	28	28	5	{'chien-robot': 22, 'taxi-robot': 3, 'sondes-robot': 1, 'poissonrobot': 1, 'mission-robot': 1}

TABLE 8.3 – Lexies productives à droite dans les composés simples

Formant	**Nbre total d'occurrences**	**Nbre de lexies uniques**	**Exemples**
socio	400	83	{'socio-psychologique': 3, 'socio-traitres': 1, 'socio-psychique': 1, 'socio-histoire': 4, 'socio-esthétique': 4, 'socio-environnementales': 1, 'socio-environnemental': 1, 'socio-religieuses': 21, 'sociostatistiques': 2, 'socio-libertaire': 2, 'socio-spatial': 5}
télé	938	63	{'télé-conseil': 4, 'télé-opérable': 1, 'télémarketers': 17, 'télé-géré': 1, 'téléverse': 1, 'télé-évangéliste': 3, 'télécratie': 4, 'télé-opérés': 2, 'télé-dépendants': 1}
politico	115	19	{'politico-mystique': 2, 'politico-dramatique': 1, 'politico-institutionnelle': 5, 'politico-littéraire': 3, 'politico-politique': 9, 'politico-ethnique': 33}
écolo	84	14	{'écolo-artiste': 1, 'écolo-chic': 5, 'écolo-légitimiste': 1, 'écolo-chics': 5, 'écolo-bobos': 12, 'écolo-terroristes': 2, 'écolo-centriste': 3, 'écolo-communiste': 23, 'écolo-bio': 3, 'écolo-compatibles': 20, 'écolo-responsables': 3, 'écolo-conservatrice': 1, 'écolo-régionalistes': 1, 'écolo-libertaire': 4}
économico	47	13	{'économico-financières': 16, 'économico-juridique': 2, 'économico-sportif': 3, 'économico-politiques': 10, 'économico-financiers': 2, 'économico-sociales': 5, 'économico-culturelle': 2, 'économico-budgétaire': 2, 'économico-religieux': 1, 'économico-corporatif': 1, 'économico-sociétale': 1, 'économico-sportive': 1, 'économico-commerciale': 1}
climato	23	12	{'climato-réalistes': 3, 'climato-sensible': 2, 'climato-négationnisme': 3, 'climato-populistes': 1, 'climatosensible': 1, 'climatoréalistes': 2, 'climatoégoïstes': 5, 'climato-dépendant': 1, 'climatofanatiques': 1, 'climatofatalistes': 2, 'climato-populiste': 1, 'climatoalarmiste': 1}
techno	22	11	{'techno-sceptique': 1, 'techno-capitaliste': 1, 'techno-activistes': 1, 'techno-centriste': 2, 'techno-optimisme': 2, 'techno-critique': 1, 'techno-civique': 1, 'techno-optimiste': 5, 'techno-pessimiste': 5, 'techno-culturels': 1, 'techno-marchande': 2}
psycho	89	10	{'psycho-somatique': 1, 'psycho-criminalistique': 1, 'psychotraumatisme': 61, 'psycho-terroriste': 1, 'psycho-spirituel': 3, 'psycho-humanitaire': 1, 'psycho-criminologique': 2, 'psycho-corporelle': 7, 'psycho-rigidité': 4, 'psycho-traumatique': 8}
socialo	20	10	{'socialo-socialiste': 2, 'socialo-marxiste': 8, 'socialo-humaniste': 1, 'socialo-syndicale': 1, 'socialo-trotskisme': 1, 'socialo-gauchiste': 2, 'socialo-gaulliste': 2, 'socialo-étatiste': 1, 'socialo-sioniste': 1, 'socialo-machin': 1}
neuro	54	10	{'neuro-méningées': 3, 'neuroplasticité': 5, 'neuro-développemental': 25, 'neuro-cardio': 2, 'neuro-méningée': 3, 'neurocérébrale': 1, 'neurotoxicité': 11, 'neuro-endocrinien': 2, 'neurofonctionnel': 1, 'neuroévolutives': 1}
anarcho	19	9	{'anarcho-royaliste': 3, 'anarcho-mitterrandiste': 1, 'anarcho-féministe': 1, 'anarcho-situationniste': 3, 'anarcho-royalisme': 1, 'anarcho-violents': 1, 'anarcho-individualiste': 2, 'anarcho-capitaliste': 2, 'anarcho-communiste': 5}
historico	19	8	{'historico-archéologue': 1, 'historico-fantastique': 3, 'historico-politique': 4, 'historico-sociologique': 2, 'historico-touristique': 2, 'historico-mystique': 2, 'historico-géographique': 4, 'historico-dramatique': 1}
ethno	26	8	{'ethno-fascisme': 1, 'ethnopluralisme': 1, 'ethnopluraliste': 1, 'ethno-différentialiste': 1, 'ethno-nationaliste': 9, 'ethno-politique': 3, 'ethno-nationalistes': 9, 'ethnopolitique': 1}
afro	65	8	{'afro-féminisme': 6, 'afrofeministe': 1, 'afro-marxiste': 1, 'afro-optimistes': 1, 'afro-trap': 13, 'afro-optimisme': 6, 'afro-futuriste': 4, 'afro-féministe': 33}
érotico	23	8	{'érotico-machiste': 3, 'érotico-romantique': 10, 'érotico-humoristique': 1, 'érotico-sadique': 2, 'érotico-rigolo': 4, 'érotico-poétique': 1, 'érotico-fantastiques': 1, 'érotico-sexuel': 1}
technico	8	7	{'technico-judiciaires': 1, 'technico-politiques': 1, 'technico-physique': 1, 'technico-artistique': 1, 'technico-mystique': 1, 'technico-juridique': 2, 'technico-politique': 1}
islamo	77	7	{'islamopsychose': 5, 'islamo-conservatisme': 4, 'islamo-terrorisme': 1, 'islamo-nationaliste': 5, 'islamo-fasciste': 8, 'islamo-conservateur': 38, 'islamo-trotskiste': 16}
euro	22	7	{'euro-réformiste': 3, 'euro-compatible': 8, 'euro-libéralisme': 1, 'euro-mondialisme': 1, 'euro-réformisme': 6, 'euro-socialiste': 2, 'euro-mondialisation': 1}
aqua	15	6	{'aquatonique': 1, 'aqua-trampo': 1, 'aquazumba': 2, 'aquastress': 1, 'aqualagon': 8, 'aqualogie': 2}
militaro	19	6	{'militaro-diplomatique': 5, 'militaro-marxiste': 1, 'militaro-économique': 3, 'militaro-mafieux': 1, 'militaro-conservateur': 1, 'militaro-politique': 8}
géo	1157	6	{'géo-localisation': 13, 'géosphère': 2, 'géoéconomique': 3, 'géolocalisation': 54, 'géocroiseurs': 32, 'géolocaliser': 1053}
diplomatico	13	5	{'diplomatico-politique': 1, 'diplomatico-économique': 2, 'diplomatico-judiciaires': 4, 'diplomatico-médiatique': 1, 'diplomatico-sportives': 5}
juridico	17	5	{'juridico-médiatique': 1, 'juridico-technique': 1, 'juridico-fiscal': 2, 'juridico-financier': 12, 'juridico-corporatiste': 1}
électro	16	5	{'électro-sensible': 1, 'électroencéphalographes': 10, 'électro-alternatifs': 1, 'électro-mécaniques': 2, 'électro-informatique': 2}

TABLE 8.4 – Liste des formants savants et modernes les plus productifs

Fracto-lexème	Nbre total d'occurrences	Nbre de lexies uniques	Exemples
néo	1468	232	{'néo-parlementaire': 3, 'néo-spécialiste': 5, 'néo-autoritaire': 2, 'néo-bourgeois': 6, 'néo-ottomane': 6, 'néo-réformistes': 1, 'néo-auberge': 1, 'néo-djihadisme': 1, 'néo-païennes': 2, 'néo-courant': 1, 'néo-post-punk': 1, 'néo-députés': 48, 'néo-centenaire': 1, 'néo-chevènementisme': 1, 'néo-bacheliers': 41, 'néo-zed': 4, 'néo-metal': 1, 'néo-baba': 1, 'néo-professionnel': 12, 'néo-députées': 48, 'néo-comédie': 1, 'néo-classic': 1, 'néo-renaissance': 7, 'néo-maurrassien': 2, 'néo-lepéniste': 1, 'néo-victorienne': 1, 'néo-internationaux': 14, 'néo-muralisme': 1, 'néo-quadra': 1}
bio	325	94	{'bio-affinité': 1, 'bio-accessibilité': 1, 'bio-acousticien': 1, 'bio-éthiques': 7, 'bio-fertilisant': 1, 'bio-senseurs': 1, 'bio-hacker': 2, 'bio-indication': 1, 'bio-économie': 7, 'bioproduits': 2, 'bio-matériaux': 6, 'biocompatibilité': 2, 'bio-seau': 3, 'bio-artificielle': 3, 'bio-sceptiques': 1, 'bio-conservateurs': 4, 'biocapacité': 13, 'bio-imprimante': 1}
cyber	1233	64	{'cyberintelligence': 1, 'cybersexisme': 25, 'cyberdivision': 1, 'cybercentre': 2, 'cyberincidents': 3, 'cyberinfractions': 1, 'cyberoffensive': 4, 'cyberclients': 2, 'cyber-sabotage': 1, 'cyberprostituée': 2, 'cyberassurances': 3, 'cyberharcelée': 2, 'cyberarmement': 4, 'cybersouveraineté': 3, 'cyberescroqueries': 1, 'cyberpatrouilles': 3, 'cybercoopération': 1, 'cyberrisque': 7}
télé	938	63	{'télé-conseil': 4, 'télé-opérable': 1, 'télémarketers': 17, 'télé-géré': 1, 'téléverse': 1, 'télé-évangéliste': 3, 'télécratie': 4, 'télé-évangélistes': 1, 'télé-transporte': 3, 'télédéclarants': 4, 'télé-transportant': 3, 'télé-surveillance': 2, 'téléversés': 4, 'télé-assistance': 4}
éco	352	52	{'écoprêts': 1, 'écotechnologies': 1, 'écoattitude': 1, 'éco-parc': 8, 'éco-contributions': 3, 'écovolontaires': 2, 'écothèque': 2, 'éco-chèque': 2, 'éco-bouteille': 1, 'écobox': 4, 'écovignette': 1, 'écoproduits': 2, 'éconature': 1, 'éco-compatibles': 2, 'éco-hôtels': 3, 'écofiction': 2, 'éco-prêt': 63, 'éco-organisme': 92, 'éco-ludiques': 1, 'éco-vallée': 9, 'éco-éducation': 1}
agro	357	35	{'agroécologique': 32, 'agromafia': 1, 'agropôle': 5, 'agro-poètes': 1, 'agro-environnementales': 28, 'agro-écologique': 32, 'agromanager': 1, 'agroécologiques': 84, 'agrobiopole': 2, 'agrosourcés': 1}
crypto	118	16	{'cryptorévolution': 1, 'cryptobotanique': 1, 'crypto-virus': 2, 'crypto-marxistes': 2, 'cryptocontinents': 1, 'crypto-monnaie': 48, 'crypto-hécatombe': 1, 'cryptodevise': 9, 'cryptomator': 2, 'crypto-érotique': 1, 'crypto-mathématiciens': 2, 'cryptodevises': 2, 'crypto-socialisme': 1, 'cryptomonnaie': 40, 'crypto-deleuzien': 2, 'cryptojacking': 3}

TABLE 8.5 – Liste des fracto-lexèmes les plus productifs

Chapitre 9

Emprunts en français contemporain (2015-2018)

Sommaire

Parmi l'ensemble des procédés néologiques, l'emprunt occupe une place à part puisque le matériau provient d'un autre système linguistique. Il occupe également une place de choix dans les études néologiques, car il s'agit d'un phénomène intrinsèque aux langues, la très grande majorité des langues se développant non pas dans un environnement clos et étanche à toute influence externe, mais au contraire dans un écosystème dont elles ne sont qu'un composant. Cela est encore plus vrai maintenant que les moyens de communication permettent des échanges au niveau mondial, chacun étant exposé, via la communication électronique, de façon quasi instantanée, à des messages linguistiques en provenance de la quasi-totalité des autres régions du monde, dans d'autres langues, et pouvant en subir les influences. Dans ce cadre, évidemment, la *lingua franca* anglo-américaine qui s'est imposée à partir des années 50, est la première langue empruntée. Nous renvoyons, pour une modélisation des différents types d'emprunts, au chapitre 5. Nous ferons ici une analyse des emprunts et les phénomènes néologiques associés repérés par la plateforme Néoveille. Ce chapitre s'inspire fortement d'un article récent (Cartier, 2018a).

9.1 Aperçu général

9.1.1 Distribution des emprunts

Les emprunts représentent 6,36 % du contingent des néologismes repérés dans Néoveille entre 2015 et juin 2018, soit 1 429 formes uniques. Le nombre d'occurrences révèle une spécificité des emprunts, puisque de ce point de vue, ils représentent 18 %, (soit une moyenne de 92 occurrences en moyenne) du total. Cela est dû à un certain nombre d'emprunts dont nous souhaitions analyser l'implantation et non simplement l'émergence (notamment toutes les lexies des réseaux sociaux, et leurs dérivés : *Facebook, Twitter, Instagram*, etc. De plus, de nombreux emprunts, déjà implantés ou en voie de l'être, ne sont pas comptabilisés ici, car ils ressortissent à d'autres matrices néologiques, notamment les affixations (cela est encore vrai pour les *buzzwords* liés aux réseaux sociaux, empruntés il y a quelques années, qui ont très rapidement été intégrés à la morphologie productive du français : *facebooker, facebookage, instagrammeur, twitos, tweetter, twitterisation*, etc.). Il faut également ajouter aux 1 430 emprunts environ un millier de xénismes que nous conservons, puisqu'ils peuvent, à un moment ou à un autre, en cas d'emploi plus fréquent, devenir des emprunts. La distribution des fréquences (figure 9.1) reflète ces spécificités des emprunts, avec une boîte à moustache très large, et la présence de nombreux emprunts à forte fréquence.

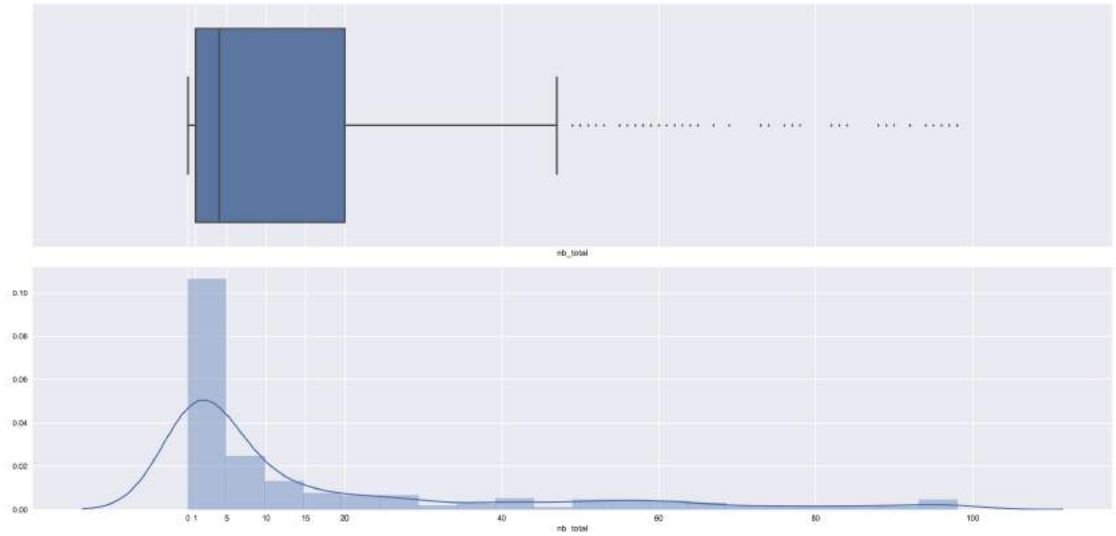


FIGURE 9.1 – Distribution des emprunts dans Néoveille

9.1.2 Langues source

Pour les emprunts, la langue source la plus représentée est la *lingua franca* anglo-américaine à environ 91 %, suivi de l'espagnol (4,5 %), de l'arabe (3 %) et de l'italien (2 %). 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, les premiers renvoient à des aires culturelles variées (Chesley et Baayen 2010). On peut également y voir une différence dans la perception collective de l'anglais international, perçu comme une langue de prestige, 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. La forte imprégnation du vocabulaire anglo-saxon n'est pas une particularité française ni un processus récent (Pulcini et al., 2012 : 2-3). Dans l'aire francophone, comme ailleurs, plusieurs facteurs permettent de l'expliquer : d'une part, depuis la fin de la Seconde Guerre mondiale, la puissance économique et politique américaine, qui s'est manifestée notamment par de nombreuses migrations culturelles et technologiques vers les États-Unis, une politique économique d'exportation massive, la stabilisation de l'anglais comme *lingua franca* dans de nombreux domaines et le nombre croissant d'anglophones non-natifs, a évidemment favorisé les emprunts à l'anglo-américain. Le développement de la radio, de la télévision et du cinéma a été après les années 70 un nouveau facteur d'essor. Enfin, l'avènement de la communication électronique, à partir des années 1995-2000, puis le développement des réseaux sociaux, depuis 2010, n'ont fait qu'accroître ce phénomène, qui est global. Ces raisons expliquent un taux important d'emprunts dans certains domaines (informatique, industrie, monde du travail en général, pratiques sociales). Par ailleurs, il existe à l'évidence un prestige de l'anglais américain dans certaines couches sociales francophones : presse féminine, presse internationale disposant d'une version locale (Slate, Huffington Post). On trouve dans notre corpus des traces à la fois de la diffusion dans toute la communauté linguistique d'emprunts liés aux réseaux sociaux, mais également dans le domaine des pratiques professionnelles et de loisirs (voir tableau 9.1).

9.1.3 Répartition par parties du discours

La distribution par parties du discours est la suivante, pour les emprunts : 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 via le morphème flexionnel *-er*, et de très nombreuses affixations appliquées sur une base empruntée.

9.1.4 Répartition par journaux et domaines

En ramenant les chiffres globaux précédents aux paramètres diastratiques et diatopiques disponibles sur la plateforme, on obtient les résultats suivants (figure 9.2) :

Les journaux les plus productifs sont relativement différenciés : pour l'ensemble des néologismes, 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). Par contre, pour les emprunts, la presse la plus prolifique est représentée par le Huffington Post, Elle, L'Express, Libération puis L'Équipe et France Soir. On peut voir dans cette modification de la distribution une trace des types de presse où les emprunts apparaissent et/ou se rencontrent le plus souvent : la presse internationale disposant d'une version française

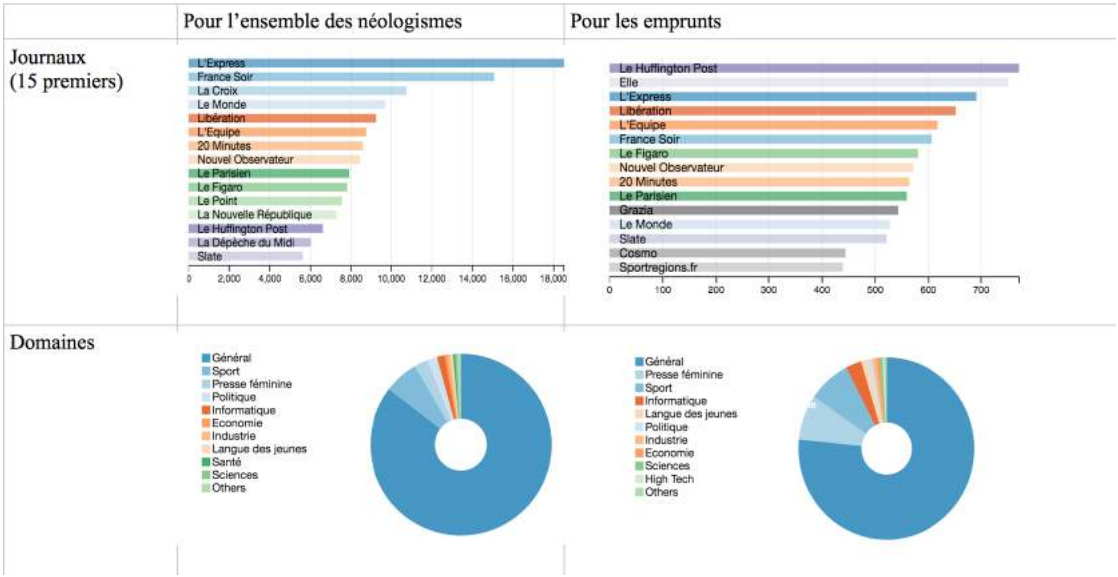


FIGURE 9.2 – Comparatif des distributions par domaines et par journaux (tous néologismes versus emprunts)

(Huffington Post et Slate), la presse féminine (Elle, Grazia parmi les quinze premiers), la presse populaire (France Soir, Le Parisien), la presse généraliste (Libération, L'Express), la presse sportive (L'Équipe) et dans une moindre mesure la presse informatique et les journaux économiques.

Cette interprétation est confirmée par la répartition en domaines : pour l'ensemble des néologismes, trois domaines sont prédominants : sport (10 %), presse féminine (10 %) et informatique (5 %) ; les tendances sont accentuées pour les emprunts : presse féminine (15 %), sport (12 %) et informatique (9 %). Ces domaines ont des occurrences pour près de 50 % des emprunts relevés. Si nous croisons emprunts et presse féminine, les journaux les plus productifs sont : Grazia, Elle, Styles, Madame Figaro et Cosmo. Si nous focalisons sur Grazia, on constate que près de 70 % des néologismes qui s'y trouvent sont des emprunts ! La presse féminine, certains titres en particulier, ont donc cette caractéristique de recourir massivement aux emprunts (et plus particulièrement aux anglicismes). Cette caractéristique est encore confirmée lorsque nous étudions les seules premières occurrences des emprunts : près de 50 % d'entre elles se trouvent dans la presse féminine, ce qui appuie l'idée que non seulement ce type de presse recourt massivement aux anglicismes, mais constitue également un lieu privilégié d'émergence de ce type de néologismes, jouant le rôle d'innovateur. Nous présentons dans le tableau 9.1 des exemples d'emprunts selon les domaines.

9.2 Cycle de vie des emprunts

Pour un descriptif des spécificités du cycle de vie des emprunts, nous renvoyons au chapitre 7, section 7.2.4.2 où nous détaillons l'émergence et la diffusion des emprunts.

Presse féminine	Informatique	Sport	Autres
Styling, contouring, fashion-week, shopper, coming-out, e-shop, lifestyle, street-food, casual, hype, street-wear, lipstick, blur, slow-food...	*Geek, boost, hashtag, playlist, emoji, tag, datacenter, ransomware, snapchat, big-data, instagrammer, scrapbooker, pop-up, blockchain, deep learning, chatbot, hoax, hotspot, live-streaming, animoji, podcaster, web-réputation...*	*Wild-card, snowboard, running, aquabike, snorkeling, futsal, kiteboarding, wakesurfing...*	*Think-tank, fake-news, talk, prime-time, call, push, reviewer, subprime, flat-tax, coworking, guest-star, fact-checking, pop-up store, debrief, car-jacking, bashing, start-up(eur), bingewatcheur, data-journalisme, ghost-writing, sugardaddy, ...*

TABLE 9.1 – Exemples d'emprunts par domaine

9.3 Emprunt de patrons lexico-syntaxiques productifs

Les emprunts à l'anglais ne se limitent pas 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 de formants à fonctionnement affixal : pour ne citer que les plus productifs, citons *e-* (formateur de verbes et de noms, dans le sens « ayant une caractéristique électronique ou numérique »), *-y* (formateur d'adjectifs, dans le sens vague de « ayant la plupart des caractéristiques de la base ») et, moins récent, *-ing* (formateur de noms d'action ou d'événement). Dans le corpus Néoveille, ces formants ont une productivité importante : 86 lexies pour le premier (soit emprunts directs, moins de dix : *e-voting, e-shopping*, etc. soit hybrides : *e-défilé, e-vendeur, e-marché, e-citoyenneté*, etc.), 22 pour le second (uniquement par emprunts directs : *jazzy, buggy, girly, cosy, healthy, creepy, skinny, glowy, flashy, bluesy, catchy, crazy, bitchy, smoky, edgy, flexy, wavy*, etc.), 303 pour le dernier. Le morphème *-ing* est attesté 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. Cependant, à 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). (Mudrochová 2017) étudie une trentaine de formes attestées dans Le Petit Robert, entre 1996 et 2002. 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. C'est un des signes les plus flagrants d'un code-switching lié au prestige dans plusieurs journaux (de façon conséquente dans le Huffington Post et la presse féminine).

Une autre caractéristique des emprunts à l'anglais concerne l'émergence de patrons lexico-syntaxiques productifs. Notamment : les formations en *-gate* (62 occurrences : *dieselgate, couscousgate, penaltygate, penelopegate*, etc.), *-friendly* (41 lexies : *vidéo-friendly, nudistes-friendly, lobby-friendly, éco-friendly*, etc.), *street-* (26 lexies : *street-*

style, street-artiste, etc.), *food-* (29 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.), *serial-* (2 lexies : *serial-buteur, serial-cendrilloneur*). Nous relevons également dans notre corpus 144 occurrences du patron N/ADJ-Ving (*car-jacking, home-staging, speed-dating, speed-watching, binge-viewing, ride-sharing*). Ce dernier cas semble montrer l'émergence d'une nouvelle structure syntaxique en N/ADJ N, clairement empruntée à l'anglo-américain (voir (Cartier et Viaux, 2017)). La diffusion, l'implantation de ces formants et schémas lexico-syntaxiques nous semble un phénomène plus profond que les simples emprunts. En reprenant l'hypothèse de l'impulsion synthétique (*synthetic imperative*) (Picone 1991 ; 1996), le français, langue analytique, tendrait maintenant à plus de formations synthétiques, ce qui se manifeste clairement avec la productivité des schémas empruntés à l'anglais N/ADJ-N, ainsi que l'implantation du suffixe *-ing*.

9.4 Emprunts et politique linguistique

Les lexies empruntées peuvent entrer en concurrence avec des lexies déjà implantées, ou avec les recommandations des organismes associés aux politiques linguistiques de la langue réceptrice. Dans le cadre francophone, la politique linguistique est particulièrement active, à la fois pour la métropole et dans la zone québécoise. De nombreuses études ont déjà été faites sur le sujet, récentes (Saugera 2017 ; Humbley 2010 ; Steuckardt 2008) et moins récentes (Picone 1996). Beaucoup démontrent que les propositions, dans le cadre métropolitain, ne donnent pas les résultats escomptés, mais les outils de suivi sont encore très impressionnistes. Dans le cadre du projet Néonaute (Cartier *et al.*, 2018a; Cartier *et al.*, 2018b), une interface spécifique a été mise en place pour étudier l'implantation comparative d'un groupe de lexies. Ainsi, nous pouvons étudier par exemple l'implantation respective de *hashtag* versus les termes préconisés par la DGLFLF. La figure 9.3 donne les comptages globaux dans le corpus Néoveille, montrant l'écrasante domination de l'emprunt, le terme préconisé *mot dièse* ne représentant qu'à peine 10% des emplois.

La figure 9.4 montre l'évolution de la distribution des emplois, qui ne montre pas d'évolution sensible en faveur de la version préconisée.

Les figures 9.5 et 9.6 montrent respectivement pour *mot-dièse* et *hashtag* la distribution par journaux, montrant que les journaux populaires, féminins et d'obédience anglo-saxonne utilisent plutôt le terme anglais, tandis que les journaux nationaux sont plus enclins à utiliser la variante préconisée (on notera la répartition partagée pour *France soir*).

9.5 Conclusion et perspectives

Nous avons présenté dans ce chapitre les résultats d'analyse concernant les emprunts en français contemporain (2015-2018). Les travaux ont été menés dans le cadre de la

Termes préconisés

Terme	Statut	Fréquence BNF	Fréquence Néoveille	Date dernière mise à jour
mot-dièse	néologisme	0	396	2018-06-22 03:20:01
mot dièse	variante-N	0	62	2018-06-22 03:15:31
mots dièse	variante-N	0	5	2018-06-22 03:26:02
mots dièses	variante-N	0	2	2018-06-22 03:16:56
mots-dièse	variante-N	0	50	2018-06-22 03:23:03
mots-dièses	variante-N	0	50	2018-06-22 03:27:25
Total		**0**	**565**	

Termes concurrents

Terme	Statut	Fréquence BNF	Fréquence Néoveille	Date dernière mise à jour
hash tag	concurrent	0	1	2018-06-22 03:20:16
hashtag	concurrent	0	1955	2018-06-22 03:26:47
mot clé	concurrent	0	619	2018-06-22 03:21:08
mot clic	concurrent	0	3	2018-06-22 03:24:01
mot-clé	concurrent	0	1001	2018-06-22 03:25:50
mot-clic	concurrent	0	124	2018-06-22 03:21:37
hash tags	variante-C	0	1	2018-06-22 03:19:27
hashtags	variante-C	0	1955	2018-06-22 03:18:06
mots clics	variante-C	0	0	null
mots-clé	variante-C	0	344	2018-06-22 03:16:26
mots-clés	variante-C	0	344	2018-06-22 03:24:21
mots-clic	variante-C	0	13	2018-06-22 03:28:13
mots-clics	variante-C	0	13	2018-06-22 03:16:32
Total		**0**	**6373**	

FIGURE 9.3 – Comptages globaux de l'emprunt *hashtag* versus termes préconisés DGLFLF

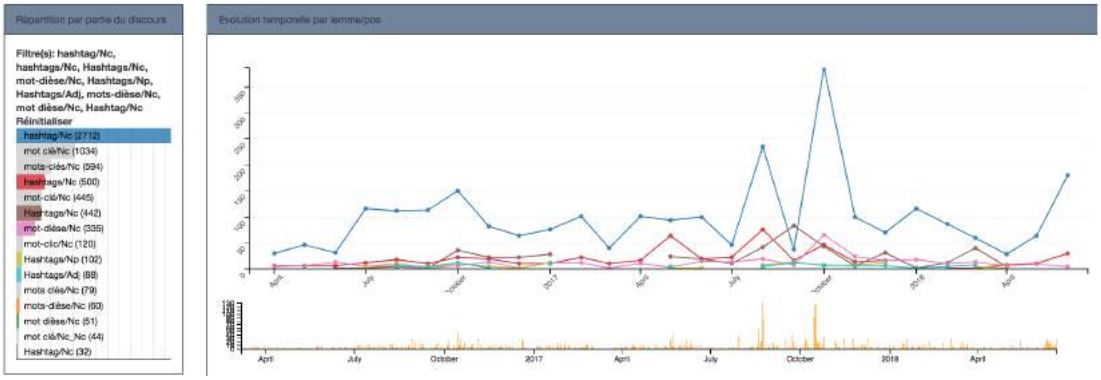


FIGURE 9.4 – Distribution de l'emprunt *hashtag* versus termes préconisés DGLFLF

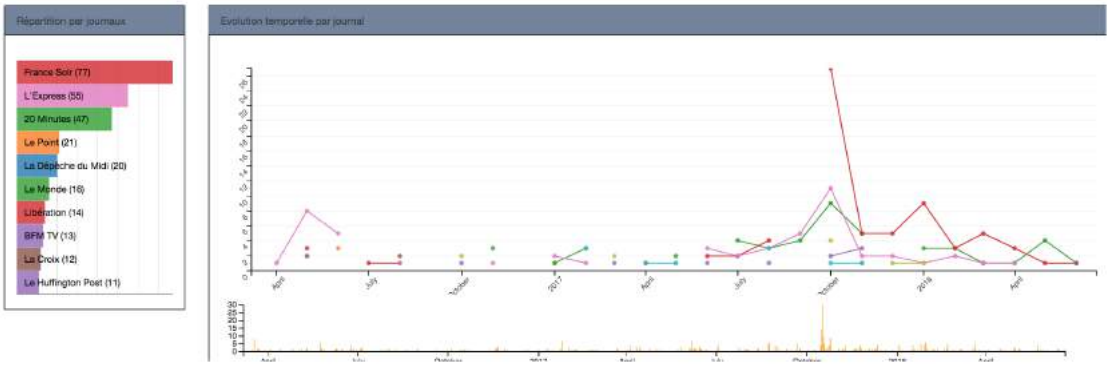


FIGURE 9.5 – Distribution par journaux de *mot-dièse*

plateforme Néoveille, qui permet de détecter semi-automatiquement, de décrire linguistiquement et de suivre l'évolution temporelle des néologismes sous trois points de vue complémentaires : évolution temporelle, évolution linguistique et socio-pragmatique. De

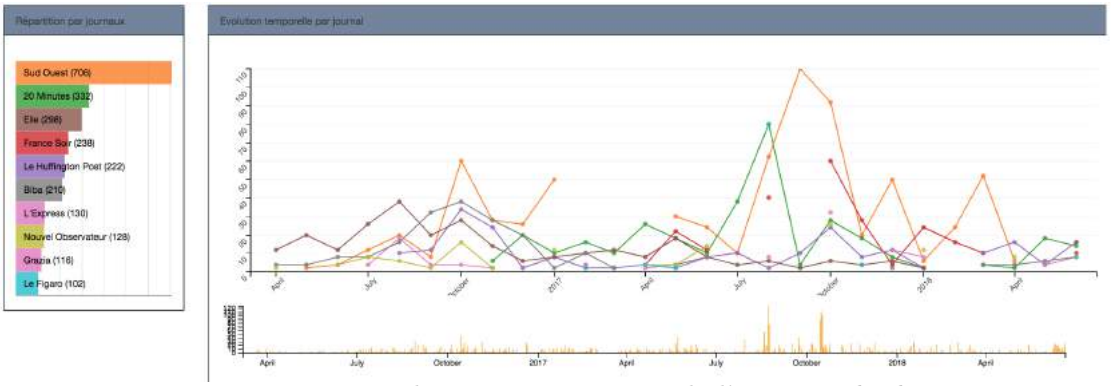


FIGURE 9.6 – Distribution par journaux de l'emprunt *hashtag*

notre étude des emprunts découlent plusieurs enseignements : tout d'abord, les emprunts en français contemporain représentent environ 6 % (soit 1 429 lexies) des formations nouvelles en français. À ce chiffre, il faut ajouter environ un millier de xénismes, ainsi qu'un nombre conséquent d'autres formations sur une base empruntée. Plus de 90 % de ces emprunts proviennent de la lingua franca anglo-américaine, tandis que les xénismes sont bien plus diversifiés. Il existe par ailleurs des domaines privilégiés de recours aux emprunts : la mode, le sport, les domaines technologiques et économiques. Les journaux les plus prolifiques appartiennent à la presse féminine, à la presse magazine parisienne, ainsi qu'à la presse populaire. Les emprunts, enfin, comme l'ensemble des néologismes, sont à plus de 75 % des hapax ou des quasi-hapax, ou plus exactement des néologismes à faible ou très faible diffusion. Il est difficile de dire si les emprunts à l'anglo-américain sont liés à des nécessités dénominatives, ou bien motivés par l'expression d'une identité liée au prestige de l'anglais. En effet, depuis plusieurs dizaines d'années, et de manière plus prégnante encore depuis l'avènement des communications numériques, l'anglo-américain est quasiment une seconde langue intégrée à laquelle nous sommes confrontés régulièrement, à la fois au niveau personnel et professionnel. Cette situation est vraie pour la majorité des langues actuelles. Cela aboutit à faire de cette lingua franca globale une ressource immédiatement disponible pour l'innovation lexicale, et qui fournit également des moules morphosyntaxiques productifs spécifiques, dotés d'un trait synthétique de plus en plus implanté en français. L'étude des néologismes à plus forte fréquence montre que, si les domaines des réseaux sociaux et des pratiques sociales spécifiques connaissent une grande popularité, les emprunts sont présents dans de nombreux domaines.

Bilan et perspectives

Le travail présenté dans ce document a porté sur le *dynamisme lexical des langues* avec trois objectifs principaux : un **objectif théorique**, en tentant de formuler un modèle qui puisse rendre compte de sa place dans l'économie générale des langues ; un **objectif computationnel**, en tentant de proposer des méthodes pour construire une plateforme de suivi des évolutions lexicales sur corpus dynamique, et en formulant des algorithmes pour détecter les néologismes formels ; un **objectif descriptif**, en détaillant, sur la base du système Néoveille, les tendances néologiques du français contemporain de 2015 à aujourd'hui.

Nous ferons dans cette conclusion un bilan des travaux effectués pour chacun de ses aspects, ainsi qu'une synthèse méthodologique. Nous évoquerons enfin quelques pistes de travail pour le futur.

9.6 Modélisation des langues

9.6.1 Dynamisme des langues : langue/discours, synchronie/diachronie

Nous avons tout d'abord élaboré les grandes lignes d'une conception dynamique de la langue qui permette de reconnaître l'existence et la place du changement linguistique : les langues sont des systèmes qui à la fois sont extrêmement stables dans le temps, mais qui sont aussi en continuelle évolution, de par les changements qui se produisent dans le monde extérieur, de par les mouvements de populations qui induisent des contacts entre langues et entre communautés humaines aux traditions variées, et de par l'histoire multiforme de chacun des individus constituant une communauté linguistique. Le changement linguistique touche toutes les composantes de la langue : la phonologie, la morphologie, la syntaxe, le lexique et jusqu'aux routines pragmatiques et discursives. L'innovation lexicale est dans ce cadre sans doute le phénomène le plus visible et le plus intense : matérialisée par de nouvelles formes lexicales et par de nouveaux usages, elle ne bouleverse qu'une partie limitée du système et nous permet de désigner les nouveautés technologiques, sociales, économiques, culturelles qui se produisent, d'exprimer nos états intérieurs et de jouer avec la langue. Le changement se manifeste tout d'abord par les innovations continuelles des individus, mais également, pour certaines créations, par leur diffusion puis leur adoption par l'ensemble d'une communauté linguistique. Le changement se manifeste aussi par la baisse et la disparition de certaines unités et de certains usages. Ces évolutions continues s'articulent avec une grande stabilité du système lin-

guistique qui reste le socle permettant la communication et la transmission des langues : les discours sont d'abord là pour que nous puissions communiquer, et transmettre la langue partagée par la communauté.

Pour situer le phénomène du changement linguistique dans l'économie générale des langues, nous avons, en nous appuyant notamment sur les travaux de Coseriu, revu les dichotomies saussuriennes langue/discours et synchronie/diachronie. Si le discours est le lieu de transmission des langues et le lieu de leur modification continue, il faut également considérer que le changement est lui-même contenu dans les langues, car elles sont non pas une nomenclature d'unités à sens fixe, mais une suite de lexies ayant un sens potentiel qui sera réalisé chaque fois de manière différente dans les discours. Les langues prévoient de plusieurs façons la création lexicale : par des modes de formation spécifiques (réduction/transformation des unités, dérivation, composition, création polylexicale), par les contacts avec d'autres langues qui permettent des emprunts lexicaux, des calques et même des procédés de formation. Le discours lui-même ne doit pas pour autant être considéré comme une suite accidentelle de faits inaccessibles à l'étude linguistique : au contraire, les situations de communication imposent des contraintes fortes sur nos paroles, parce qu'elles sont le fruit de traditions discursives pour leur forme, leur contenu et leur déroulement interne. Même dans les situations apparemment les moins normées, des contraintes pragmatiques pèsent sur l'émergence ou non des innovations linguistiques.

Nous avons également insisté sur la nécessité de prendre une triple perspective pour étudier l'innovation lexicale et le changement linguistique en général : une *perspective linguistique interne*, car il est évidemment nécessaire de décrire les mécanismes linguistiques permettant l'innovation lexicale ; *une perspective cognitive*, car ce sont bien les individus qui font vivre et évoluer les langues, et il faut pouvoir comprendre les processus cognitifs qui président à la préservation comme au changement linguistique. De ce point de vue, le processus de mémorisation et son résultat, l'*entrenchment*, est central. La fréquence d'exposition et d'usage est le principal déterminant de la mémorisation et fonde une étude statistique et probabiliste des langues. Enfin, *une perspective sociolinguistique*, car l'innovation lexicale est avant tout une variation linguistique, et les langues sont traversées de variations et comprennent des variétés qu'il faut mettre au jour pour comprendre l'inscription sociale, économique, culturelle de certaines innovations ; de même, les innovations lexicales, qui sont nombreuses de manière hapaxique, diffusent parfois au sein de la communauté linguistique, et il faut se faire une idée précise des flux et des réseaux de communication par lesquels passent les innovations.

9.6.2 Unités lexicales et innovations lexicales

Après cette modélisation générale, nous avons étudié plus en détail les notions d'unité lexicale et d'innovation lexicale. La conception dynamique des langues s'accompagne d'un principe de continuité entre les unités linguistiques, qui ne peuvent être conçues que comme des prototypes que la tradition a diversement nommé. En partant d'une conception traditionnelle des grammaires du français, distinguant mot, phrase et texte, nous avons tenté de spécifier les caractéristiques spécifiques des unités lexicales tout d'abord en traçant sa frontière supérieure, la proposition énoncée. Nous inspirant de la

conception de (Adam, 1990; Adam, 2005), nous avons indiqué que cette unité combine une prédication, un point de vue énonciatif et une inscription contextuelle et co-textuelle. En deçà de la proposition énoncée, nous inspirant de la notion de *construction*, nous avons montré que l'unité lexicale, ou paire forme-sens, s'étend des morphèmes liés (flexion et affixe) aux morphèmes libres (les lexies au sens traditionnel) et au-delà aux unités polylexicales plus ou moins figées et aux constructions lexico-syntaxiques et syntaxiques. Les formes en sont donc diverses, mais elles partagent toutes cette propriété d'associer une forme à une représentation mentale déterminée qui n'est pas de l'ordre de la proposition énoncée. Nous avons également établi quelques unes des propriétés prototypiques de ces différentes unités, qui ressortissent à trois dimensions : la forme elle-même, les propriétés morphosyntaxiques (partie du discours, formes flexionnelles éventuelles et règles de combinatoire) et la représentation mentale. Nous avons émis l'idée que les parties du discours ne sont que des abstractions de fonctionnements combinatoires prototypiques des lexies qui permettent leur inscription dans les prédications. Les langues isolantes disposent de lexies pures dont la combinatoire est matérialisée par des formes spécifiques isolées et l'ordre des mots permet les relations de dépendances entre les lexies, tandis que les langues agglutinantes agglomèrent ces informations sur les lexies proprement dites. Les langues casuelles et les langues flexionnelles proposent une situation mixte, notamment pour les dernières en agglomérant des propriétés sémantiques spécifiques et générales pour trois parties du discours (nom, adjectif et verbe). Il n'en reste pas moins que les unités lexicales doivent être décrites en distinguant les trois dimensions : formelle, morphosyntaxique (à défaut d'une meilleure dénomination, car il s'agit là également d'une information sémantique) et sémantique ou mentale. Nous avons ensuite décrit les particularités spécifiques des flexions, des affixes et des lexies proprement dites.

Après cette modélisation générale, nous avons établi les propriétés essentielles des innovations lexicales : il s'agit d'unités lexicales qui divergent par rapport à l'usage d'une communauté linguistique donnée, à la fois au niveau de la forme (qu'il s'agisse de la forme au sens classique, orthographique et morphosyntaxique, auquel cas nous avons des innovations formelles ; ou qu'il s'agisse d'une forme au sens d'une combinatoire inhabituelle, auquel cas nous obtenons une innovation sémantique, même si le terme ne rend pas justice à la modification formelle qui se produit) et au niveau du sens, puisqu'une nouvelle conception mentale apparaît. Nous avons également indiqué que nous considérons l'innovation lexicale dès le moment de son émergence et quel que soit son sort futur dans la langue, en limitant le concept au moment où l'unité lexicale est adoptée par l'ensemble d'une communauté linguistique et où donc la divergence n'est plus ressentie. Nous avons indiqué l'importance d'effectuer une typologie des procédés d'innovation lexicale, typologie qui est présentée plus exhaustivement dans le chapitre 5. Enfin, nous avons présenté les principales phases du cycle de vie des innovations, l'émergence, la diffusion et l'adoption, en nous appuyant sur les points de vue lexicologique, sociolinguistiques et psychologiques.

9.6.3 Langue, variations et variétés

Nous avons ensuite révisé la notion de langue au sens structuraliste de langue unique et homogène. Les innovations lexicales sont en effet d'abord et avant tout des variations et ces variations sont liées à des groupes sociolinguistiques et aux individus. Il s'agit donc de pouvoir caractériser sociolinguistiquement les innovations, ce qui passe d'abord par une caractérisation des notions de variations et de variétés.

La conception homogène des langues a été d'abord remise en cause, dans l'ère moderne, par (Weinreich *et al.*, 1968), qui conçoivent les langues comme des *systèmes dynamiques* caractérisés par une « hétérogénéité ordonnée »(*orderly heterogeneity*) : il peut se présenter - et il se présente de façon continue - des divergences d'usage au sein de la communauté linguistique. (Coseriu, 1980, p.5) affirme ainsi que : « Le locuteur (...) se trouve confronté, dans son expérience réelle, à l'état d'une langue historique, dont la synchronie est différenciée des points de vue diatopique, diastratique et diaphasique. Tout locuteur, s'il ne connaît pas la langue historique dans son ensemble, connaît, au moins jusqu'à un certain degré, plus d'un dialecte et plus d'un niveau de langue ; et tout locuteur maîtrise plusieurs styles de langue. » Il existe donc non pas une langue unique (le français, l'italien, l'espagnol, etc.) mais une série de variétés linguistiques qui coexistent et s'interpénètrent.

Ce principe de variabilité intrinsèque des langues est facile à constater : d'abord, chacun d'entre nous, même dans un environnement monolingue, est compétent dans plus d'un code linguistique : nous n'utilisons pas le même vocabulaire ni les mêmes formulations selon que nous sommes dans une situation intime, familiale, amicale ou professionnelle. À l'écrit, nous n'employons ni le même vocabulaire ni la même syntaxe s'il s'agit de rédiger une lettre pour l'administration, un blog ou un travail académique. Les différences parfois considérables entre l'oral et l'écrit sont une autre preuve de l'existence de différentes variétés d'une même langue. Ensuite, il existe à l'évidence des cercles sociaux disposant d'une variété de langue spécifique : la langue des jeunes, la langue des bobos ou des hipsters, la langue des geeks, la langue des bouchers, etc. La variation et le changement linguistique peuvent être mis en évidence avec la notion de *variable linguistique*, lorsque « deux formes différentes permettent de dire "la même chose", c'est-à-dire lorsque deux signifiants ont le même signifié et que les différences qu'ils entretiennent ont une fonction autre, stylistique ou sociale. »(Calvet, 1998: p.76). La variation et le changement linguistique sont donc deux facettes du même phénomène : en synchronie, il y a des variations, qui sont ensuite éventuellement résolues en diachronie par l'adoption d'une des variantes, ou une redistribution du champ sémantique.

Une première approche pour caractériser les variations et les variétés a été proposée par Coseriu, qui invoque trois dimensions :

— une dimension diatopique, qui est sans doute le fondement de la création des langues, sur des bases géographiques : c'est d'abord la proximité qui permet la communication linguistique et permet la formation des langues ;
— une dimension diastratique, qui permet d'isoler des variations voire des variétés en liaison avec des sous-groupes de la communauté linguistique : de ce point de vue, la situation se révèle complexe car, historiquement, en tout cas dans les so-

ciétés occidentales, on peut identifier d'abord des variétés conditionnées par la stratification par classes sociales, puis, à l'époque moderne et contemporaine des variétés conditionnées par des communautés de pratiques sans doute plus temporaires ; à ces deux situations correspondent deux méthodologies : la première essaie de corréler les variétés à une macro-structure sociale, en se basant sur des propriétés biologiques, ethnico-culturelles et économiques ; la seconde s'intéresse aux réseaux sociaux locaux, en complétant la première approche d'une étude des réseaux de communication entre les individus, aboutissant à identifier des réseaux denses et moins denses, et identifier des rôles spécifiques permettant de décrire l'émergence et la diffusion des innovations linguistiques ;

— une dimension diaphasique, qui permet de rendre compte de "styles", ou variations individuelles, liées aux interactions des individus dans des situations de communication diverses. Là encore, la situation est complexe, et différentes approches ont été proposées : une approche pour laquelle le style est déterminé par une adaptation à la classe sociale de l'interlocuteur, qui peut être appliqué dans le cadre de relations sociales conflictuelles où certaines variétés sont stigmatisées et d'autres valorisées, et où chacun des individus est le représentant d'une classe sociale et d'un vernaculaire spécifique ; une second approche considère que le style est le résultat d'une adaptation complexe à l'auditoire - cet auditoire étant multiple (les auditeurs, l'auditoire secondaire et l'auditoire imaginé) - et est dépendant des objectifs de communication ; enfin, une dernière approche considère le style comme une expression volontaire déterminée par des objectifs communicatifs.

Dans l'approche sociolinguistique, l'innovation linguistique se définit par l'émergence d'une variante linguistique. Les sociolinguistiques ont proposé différents modèles permettant de définir le mécanisme de sélection et d'adoption des variantes. Labov distingue le changement par en-dessous et le changement par au-dessus : celui par en-dessus concerne l'adoption d'une innovation d'un groupe social considéré comme plus prestigieux et/ou du rejet d'innovations provenant d'un groupe stigmatisé ; il est conscient et conditionné par la volonté d'adopter les pratiques des groupes sociaux les plus valorisés ; le changement par en-dessous est inconscient et consiste à introduire des marqueurs linguistiques propres à un groupe social par les membres eux-mêmes : il s'agit d'un processus inconscient qui est conditionné par l'identification au groupe social. (Giles et Powesland, 1975) a introduit la notion d'adaptation-convergence à l'interlocuteur/auditoire (*Communication Adaptation Theory*) pour expliquer l'adoption de certaines innovations, tandis que (Trudgill *et al.*, 2000) a insisté sur l'importance de la fréquence d'exposition (qui sera plus approfondie par les psycholinguistes).

Au niveau des acteurs de l'innovation, la sociolinguistique a identifié deux structures de réseaux sociaux favorisant l'adoption des innovations : Labov considère que la situation la plus favorable est liée à l'existence d'individus ayant un réseau dense à liens forts et un réseau dense à liens faibles : dans cette configuration, ils sont les diffuseurs par excellence, de par leur prestige au sein de leur propre communauté, et leurs liens nombreux avec d'autres communautés. Un autre modèle, dit du lien faible, considère

que ce sont les individus à liens faibles qui sont les véritables diffuseurs des innovations, car ils sont des ponts entre les communautés, et sont également des innovateurs, de par leur ouverture à des communautés disparates et à leurs idées. Mais ce modèle peut être complété par l'ajout des "hubs", ces individus ayant un réseau dense à liens forts, qui permet la diffusion au sein de chaque communauté.

Nous avons également présenté l'approche de Koch et Oesterreicher qui explicitent une quatrième dimension, universelle et transversale aux trois autres, celle de la proximité-distance communicative, qui permet de caractériser toutes les situations de communication. Construite à partir de la distinction des codes écrit et oral, qui lui sert de fondement, elle permet d'associer à chaque situation de communication une valeur de proximité. Nous avons indiqué deux pistes pour préciser ce modèle : tout d'abord, rétablir le caractère absolument immédiat de l'oral, alors que l'écrit se place dans le continuum proximité-distance ; ensuite, rationaliser les critères permettant d'identifier la proximité des situations de communication. Il nous semble à cet égard que le schéma de communication de (Hymes, 1982) est le plus complet pour établir une liste complète de paramètres.

Nous avons également évoqué, dans ce chapitre, différentes pistes de travail à explorer, notamment l'intérêt d'une étude des flux de communications entre individus à un niveau plus global que les études menées jusqu'à présent, en ajoutant les organes de diffusion d'information qui jouent, dans la période contemporaine, un rôle de diffuseurs et de régulateurs non négligeables et font à l'évidence partie des situations de communication auxquelles nous sommes exposés. Cette approche inductive des variations nous semble plus objective que les études sociologiques par l'établissement a priori de caractéristiques individuelles pour déterminer les corrélations variations/changements linguistiques et les groupes d'humains.

Nous avons également proposé de compléter les dimensions de la variation en ajoutant une dimension diasituationnelle, qui intègre l'analyse de la distance communicative et focalise sur les situations de communication qui sont, avec le paramètre géographique, les groupes sociaux et les individus le quatrième déterminant des variations : en effet, les situations de communication sont dans leur très grande majorité le résultat de traditions discursives et même les situations apparemment les moins normées sont soumises à des contraintes pragmatiques.

9.7 Modèles et méthodes pour la détection et le suivi automatiques des innovations lexicales

Nous avons ensuite abordé l'automatisation de la détection et du suivi des innovations lexicales sur corpus dynamique. Le propos se divise en trois chapitres : le premier évoque la construction d'une plateforme comprenant l'ensemble des composants nécessaires et des traitements pour détecter et suivre les néologismes ; le second chapitre traite du repérage automatique de la néologie formelle ; le troisième, sur la néologie sémantique, fera l'objet d'un travail ultérieur..

Dans le **chapitre 4**, nous avons détaillé les contours d'une plateforme idéale pour la

détection et le suivi (semi-)automatique des innovations lexicales sur corpus dynamique. Après avoir passé en revue différents outils disponibles pour effectuer ces opérations, moteurs de recherche généralistes et spécialisés, outils développés dans le cadre de la linguistique de corpus, et outils spécialisés dans la détection et la gestion des néologismes, nous avons établi l'architecture d'une plateforme idéale : ses composants (un gestionnaire des sources d'information textuelle, un lieu de stockage des corpus récupérés, un gestionnaire de néologismes, un gestionnaire de ressources lexicographiques) et les processus à mettre en œuvre (récupération continue du sources d'information textuelle et leur stockage dans un moteur de recherche, détection automatique des néologismes - formels et sémantiques - et leur stockage dans une base de données pour validation), outils de fouille et de visualisation interactive des contextes et des paramètres sociopragmatiques. Nous avons également insisté sur la nécessité de combiner les traitements informatiques et l'expertise linguistique humaine : aucun système informatique n'est apte à générer des résultats totalement fiables, et aucun expert humain n'est apte à traiter de grandes masses de données. C'est dans la collaboration entre l'humain et la machine que se trouve la solution la plus adéquate, pour que l'humain corrige ou valide les décisions automatiques, et « apprenne »à la machine les décisions prises ; en retour, les traitements automatiques sont progressivement plus fiables. À partir de cette architecture, nous avons présenté les réalisations effectuées dans le système Néoveille et présenté en détail ses différents composants. La conclusion évoque quelques pistes pour l'amélioration de l'existant.

Dans le **chapitre 5**, nous nous intéressons à la détection automatique de la néologie formelle. Nous modélisons tout d'abord plus précisément ce phénomène : délimitation du périmètre respectif de la néologie formelle et de la néologie sémantique, distinction avec le phénomène de flexion et les constructions polylexicales, identification des phénomènes de dérivation, de composition, de réduction/transformation et d'emprunt lexical ; établissement des propriétés distinctives entre affixe, fractolexème (ou affixoïde) et lexie et du continuum entre ces unités ; présentation de la notion de productivité qui est l'une des propriétés importantes permettant de caractériser la néologie formelle. Nous passons ensuite aux méthodes de repérage automatique, en présentant tout d'abord un état de l'art des méthodes précédemment utilisées. Ensuite, nous présentons la méthode utilisée dans Néoveille, consistant à utiliser un dictionnaire d'exclusion et différents filtres pour établir automatiquement une liste de candidats néologismes. Nous ajoutons à cette méthode classique une phase d'apprentissage itératif par intervention de l'expert humain pour corriger les décisions automatiques et réinjecter les décisions humaines dans le système, permettant d'obtenir assez rapidement des améliorations conséquentes de la détection. Une évaluation est présentée, et, en perspective, nous évoquons un prototype mis en place et mobilisant des algorithmes d'apprentissage automatique qui permettraient encore d'améliorer le système actuel.

9.8 Application : Tendances néologiques du français contemporain (2015-2018)

Dans la troisième partie, applicative, nous présentons le travail effectué à partir de la plateforme *Néoveille* sur l'innovation lexicale en français contemporain à partir de corpus dynamique. Il s'agit d'une présentation qui étend et détaille le travail effectué par un groupe de travail composé de linguistes qui ont travaillé sur le sujet depuis septembre 2015.

Dans le **chapitre 6**, après avoir présenté quelques éléments méthodologiques pour effectuer le travail de validation et de description, nous avons détaillé les principales caractéristiques des quelques 22 000 néologismes détectés et validés : du point de vue des procédés néologiques, la préfixation domine largement, avec près de 75% du contingent, suivie par la composition simple, les emprunts lexicaux, la suffixation et la fracto-composition. La très grande majorité des innovations est hapaxique ou de très faible diffusion. La domination de la préfixation peut facilement s'expliquer, puisqu'il s'agit là, avec la suffixation, des procédés morphologiques productifs de la langue, directement accessibles et généralement sans besoin d'une ré-analyse. La répartition par domaines montre que, même si ce sont dans les actualités générales que se produisent la majorité des néologismes, les domaines du sport, de la presse féminine et de l'informatique sont des domaines particulièrement productifs. La répartition par journaux confirme ce constat, avec un plus fort taux de néologismes dans la presse populaire que dans les grands journaux nationaux. La répartition par parties du discours confirme également l'intuition : les noms prédominent très largement, suivis par les adjectifs et les verbes. Concernant le cycle de vie des néologismes, l'étude permet d'affiner la notion d'émergence : il ne s'agit pas seulement de l'événement hapaxique, mais, dans le contexte moderne d'une diffusion accélérée des informations, d'une période temporelle courte au cours de laquelle les lexies apparaissent dans des contextes socio-pragmatiques déterminés, avec une (très) faible fréquence. Généralement, cette émergence se caractérise également par des emplois métalinguistiques et des gloses, sauf pour les procédés d'affixation, pour lesquels le sens est généralement compositionnel et immédiatement accessible. Concernant la diffusion, l'étude permet de la caractériser par plusieurs facteurs : évolution fréquentielle, diffusion hors des contextes socio-pragmatiques initiaux, adaptation et stabilisation phonologique, orthographique et morphosyntaxique (pour les emprunts lexicaux), intégration à la morphologie productive et enfin stabilisation du profil combinatoire.

Dans le **chapitre 7**, nous portons notre attention sur les procédés dérivationnels. Nous présentons tout d'abord quelques définitions des différents phénomènes impliqués : préfixation, suffixation et parasynthèse. Nous détaillons ensuite les formations lexicales par préfixation, le plus gros contingent de néologismes, en montrant que la distribution offre une courbe de Zipf typique, avec une très grande majorité d'hapax et, plus spécifique, un nombre non négligeable de formations ayant une fréquence très faible ou faible. Parmi les 69 préfixes attestés dans notre corpus, cinq formants (*non-, ex-, anti-, quasi-, ultra-*) se détachent, à la fois en nombre de formes uniques produites, mais également en nombre d'occurrences. Ces formants sont sans doute les préfixes les plus typiques. Les

autres préfixes se trouvent dans un continuum qui paraît difficile à catégoriser, même si une quinzaine d'entre eux sont en situation de quasi-non-productivité. Une étude diachronique ultérieure permettra sans doute de suivre les éléments dont l'usage baisse ou augmente. Si nous classons les formants selon la productivité en expansion, le classement est modifié, puisque les plus productifs sont *mini-, post-, co-, multi-, etc.*. Parmi les formants, la plupart s'appliquent à des noms, des adjectifs ou des adverbes, et très peu s'appliquent à des verbes. Nous faisons ensuite une analyse de la série des préfixes évaluatifs exprimant le haut degré, avec la domination de trois formants : *ultra-, super-, hyper-*. Nous terminons cette analyse par l'une des caractéristiques des préfixes (partagée avec les suffixes) : l'absence quasi-généralisée de glose métalinguistique. Nous nous intéressons ensuite aux suffixes : en nombre moindre, ils présentent une distribution classique, avec une très grande majorité de formations hapaxiques. Parmi les suffixes les plus productifs, la triple suffixation en *iser-, isation-, -isatrice/teur* domine, en lien direct avec les transformations sociétales. Les formations génératrices d'agent (*-ien(ne), -iste, -eur/euse*) sont également très fréquentes, avec une application conséquente aux hommes et femmes politiques.

Dans le **chapitre 8**, nous passons aux précédés par composition, dominés par le composition simple, la composition savante, hybride et la fracto-composition. les phénomènes de compocation et de mot-valisation sont beaucoup plus rares. La distribution des procédés par composition est maquée par un plus grand nombre d'hapax que dans la dérivation, il s'agit généralement d'emplois contextuels qui ne sont par la suite plus repris. La composition simple est largement dominée par le schéma syntaxique N-N, suivie par ADJ-ADJ, V-N, ADJ-N, N-ADJ, les neuf autres schémas étant beaucoup plus rares. On remarque également l'existence de schémas lexico-syntaxiques productifs (généralement N-N), avec 292 lexies à gauche productives (*social-,livre-, écrivain-, tour-, robot-, médecin-, etc.*) et une trentaine à droite (*-phare, -choc, sphère, -clé, -star, -compatible, -gate, -attitude*), parmi lesquelles on note trois formations empruntées à l'anglais. Ce même phénomène de lexies productives se produit dans la composition savante, hybride et dans la fracto-composition. Dans ce dernier cas, on voit bien qu'un certain nombre sont devenus des quasi-affixes : réduction à deux syllabes, règle de formation non ou faiblement contrainte, très forte productivité (*néo-, bio-, cyber-, télé-, éco-, agro-, etc.*). Parmi les compocations, sur les 83 formations validées, certaines sont apparues il y déjà une dizaine d'années, tandis que la plupart restent hapaxiques. Nous prenons l'exemple de la diffusion de *Frexit* pour illustrer l'intérêt d'un suivi multidimensionnel : évolution de fréquence, évolution des distributions par domaines et par journaux. On constate une hausse temporaire de son usage durant l'entre-deux-tours des élections présidentielles de 2017 et sa diffusion dans la presse généraliste. Depuis, le terme est d'une faible fréquence.

Enfin, Dans le **chapitre 9**, nous nous intéressons aux emprunts, qui constituent 6% des néologismes, mais près de 18% des occurrences, montrant une diffusion rapide. À ces chiffres, il faudrait ajouter près d'un millier de xénismes et un certain nombre de dérivés construits sur une base empruntée. La très grande majorité proviennent de la *lingua franca* anglo-américaine, avec des contextes privilégiés d'émergence : presse féminine, presse magazine parisienne, presse populaire et informatique. Il est difficile

de faire le départ entre les nécessités dénominatives ou l'expression d'une identité liée au prestige de l'anglais dans la motivation de ces emprunts. Il s'agit de toute façon d'un phénomène global qui touche la très grande majorité des langues, et de manière encore plus rapide depuis l'émergence d'internet puis des réseaux sociaux. Les emprunts lexicaux ne sont pas les seuls représentants, puisque notre corpus comprend toute une série de constructions empruntées et très productives : *-gate, -attitude, -compatible* déjà évoqués plus haut, mais également un schéma plus générique en N-Ving, particulièrement productif depuis l'émergence des formations en N/ADJ-shaming et -N-/ADJ-bashing. L'étude par domaine montrent qu'il n'y a pas de limitation aux réseaux sociaux ou à des pratiques sociales spécifiques. Nous présentons enfin dans ce chapitre les dernières expérimentations menées pour suivre l'évolution différentielle d'une grappe de mots. Dans ce cadre, nous présentons quelques exemples d'emprunts pour lesquels ont été proposés des termes francisés (*hashtag - mot-dièse, digital native - enfant du numérique*). Dans ces deux cas, l'usage du terme anglais l'emporte largement, même si les grands journaux nationaux tentent de diffuser les termes préconisés.

9.9 Perspectives

Nous passons maintenant en revue quelques perspectives de travail que nous envisageons, du point de vue de la description des néologismes, du point de vue de l'automatisation des processus, et du point de vue théorique. Nous terminerons pas quelques considérations méthodologiques.

9.9.1 Description des néologismes

Du point de vue descriptif, le travail présenté pour le français doit encore être approfondi, sur plusieurs points:

— **finalisation des descriptions linguistiques** : nous avons mis en œuvre dans le système Néoveille, notamment grâce aux travaux de Jean-François Sablayrolles, une méthodologie pour décrire successivement le ou les procédés de formation, les caractéristiques linguistiques (forme, combinatoire, sémantique), les caractéristiques socio-pragmatiques des occurrences, et l'évolution diachronique de ces propriétés, pour chacun des néologismes. Sur les 22 000 néologismes validés, environ 40% sont exhaustivement décrits, il reste donc à finaliser le travail descriptif afin de pouvoir générer des tendances générales et détaillées et permettre une exploitation des résultats dans d'autres cadres ;

— **description des procédés de troncation/transformation** : ces procédés, même si un certain nombre de formations ont été repérées et validées (voir site internet), elles n'ont pas encore été analysées et décrites globalement, c'est l'objet d'une thèse en cours de Valeriya Vinogradova que je co-encadre ;

— **néologie sémantique** : cet aspect peut-être plus massif encore de l'innovation lexicale n'est actuellement pas traité et, si des pistes de détection automatique sont actuellement en phase exploratoire et quelques résultats disponibles (Car-

tier, 2016c; Cartier, 2017b; Cartier, 2017c; Cartier, 2018d), le travail n'est pas suffisamment avancé pour mettre en place les programmes dans la plateforme ; il s'agit là du gros chantier à venir, en liaison étroite avec le TAL ;

— **description des tendances néologiques dans d'autres langues** : actuellement le système permet de détecter des néologismes formels pour 11 langues, avec, hors français, trois langues bien avancées : l'italien, le portugais et le russe, avec environ 10 000 néologismes détectés et validés dans chaque langue. Il s'agira de finaliser les descriptions puis d'en faire des analyses. Dans le cadre du projet Néoveille, nous avons pu mobiliser une équipe de linguistes pour le français, mais il sera nécessaire, pour que le système perdure, et s'étende à d'autres langues, de mettre en place un projet plus vaste, par exemple dans le cadre d'un réseau européen.

9.9.2 Détection et suivi automatiques des innovations lexicales

Du point de vue du modèle opérationnel et des méthodes automatiques, plusieurs chantiers sont également ouverts:

— **sources d'information** : tout d'abord, au niveau des sources d'information, le système actuel ne caractérise que très approximativement les textes (date, journal source, type de texte - exclusivement des articles de presse - domaine - assigné par l'éditeur lui-même -) ; il s'agirait de développer des algorithmes pour rendre compte de façon plus fine des caractéristiques des documents sources : auteur(s), thématiques traitées dans les textes eux-mêmes, etc. De ce point de vue, il manque un modèle global des situations de communication qu'il faudrait préalablement expliciter, pour envisager des modèles automatiques. Du point de vue des thématiques, nous avons, dans un autre projet sur un très grand corpus d'archives du web (projet Néonaute avec la BnF : (Cartier *et al.*, 2018b; Cartier *et al.*, 2018a)) mis en œuvre les algorithmes de *topic modeling* qui devraient prochainement être introduits dans le système, permettant ainsi d'avoir une idée plus fine des thématiques traitées par chaque texte ;

— **extension des types de documents sources pris en charge** : actuellement le système fonctionne sur des articles de presse généraliste ou de vulgarisation ; il serait évidemment intéressant de pouvoir ajouter des documents de différentes natures : blogs, réseaux sociaux, etc. Il serait également intéressant d'inclure des textes oraux et multimédia ; pour ce faire, des collaborations doivent être initiées pour construire une plateforme plus vaste permettant l'accès à un plus grand nombre de sources variées ;

— **mise en œuvre de l'approche apprentissage automatique / apprentissage profond pour la détection des néologismes formels** : concernant spécifiquement la néologie formelle, nous avons indiqué en conclusion du chapitre qu'un prototype plus efficace avait pu être défini ; il s'agit maintenant de le mettre en place dans le système. L'intérêt de cette approche est évident : il permet, à partir d'un jeu de référence de lexies néologiques, de lexies non-néologiques et de formes fautives, de construire un modèle pour la détection des nouveaux néolo-

gismes, sans recours à des ressources linguistiques (dictionnaires de référence et d'exclusion) lourdes à mettre en place et à maintenir ;
— **détection de la néologie sémantique** : là encore, il s'agira de mettre en œuvre de manière plus systématique les pistes mises au jour dans différentes communications (voir plus haut). Cela demande également des corpus quantitativement plus importants, et de pouvoir accéder à des corpus diachroniques antérieurs pour constituer un corpus et un modèle de référence ;
— **valorisation de la plateforme** : un dernier point, qui engage les autres tâches, concerne la valorisation de la plateforme. Celle-ci a été développée via un appel à projet IDEX qui a permis d'obtenir des résultats très encourageants, et il va falloir trouver de nouveaux financements et construire un réseau plus conséquent pour faire vivre ce projet. Il s'agira de stabiliser la plateforme afin qu'elle puisse être installée de manière libre par différents centres de recherche, et dotée d'un réseau de chercheurs et d'ingénieurs pour continuer à la développer et permettre l'analyse des innovations lexicales en corpus continu. Une première étape a été franchie avec la mise à disposition de l'ensemble des source sous licence Apache 2.0 : `https://github.com/ecartierlipn/neoveille2016.git`.

9.9.3 Modélisation des langues

Du point de vue de la modélisation des langues, là encore, des nombreux points restent problématiques et méritent approfondissement. Il nous semble acquis que : les langues sont des systèmes complexes dynamiques, intrinsèquement stables pour les besoins de la communication entre les membres d'une communauté, et intrinsèquement évolutifs de par les circonstances chaque fois uniques de la communication ; pour les décrire, il faut prendre au moins trois perspectives, l'une linguistique, l'autre cognitive et la troisième sociolinguistique ; du fait de la dynamicité de ces systèmes, de la variété interne à toute communauté linguistique et de la variété des situations de communications, il est illusoire de prétendre décrire une langue dans un état donné de manière systématique. Il est par contre raisonnable de chercher à décrire des régularités et des prototypes d'unités et de comportements linguistiques. Pour ce faire, il nous semble nécessaire de porter nos efforts sur les éléments suivants :

— modélisation plus fine des situations de communications : propriétés socio-pragmatiques des textes sources / situations de communication ;
— approfondissement de la perspective linguistique : il s'agirait de décrire de manière plus formelle les unités linguistiques, d'établir pour chacun des pôles les caractéristiques prototypiques que nous avons établies à gros traits, et également d'étudier l'organisation sémantique des unités : avons-nous à faire à une structure globale, ou à une collection hétérogène de sous-structures plus ou moins organisées ?
— approfondissement des perspectives cognitive et sociolinguistique : de même, afin d'approfondir les perspectives cognitives et sociolinguistiques, il nous faudra établir des collaborations afin d'enrichir la vision globale du phénomène d'innovation lexical ;

— notion de système dynamique complexe : la notion de système énoncée par Saussure et revue par Weinreich, Coseriu et la sociolinguistique aboutit à un questionnement : si nous posons l'existence d'unités linguistiques - placées dans un espace continu - avec trois pôles principaux, les unités lexicales, qui sont des unités représentationelles-dénotationnelles ou paires forme-sens, dotées également d'instructions pour leur combinatoire, les unités propositionnelles-communicationnelles, la proposition énoncée, l'unité véritable de la communication linguistique, et d'éventuelles séquences normées de propositions énoncées formant "texte", comment s'organisent ces différentes unités : s'agit-il d'un système global plaçant, dans un état de langue donné, les différentes unités en relations stables les unes par rapport aux autres, ou bien s'agit-il d'un système partiellement stable, dans lequel certaines zones ne sont pas organisées ?

— modélisation de la néologie sémantique : enfin - nous y revenons de façon récurrente dans cette conclusion -, il nous semble que les travaux sur la néologie sémantique, une fois de premiers résultats obtenus à grande échelle, permettront d'obtenir une vision globale plus fine encore de l'innovation lexicale : l'innovation sémantique, comme l'innovation par procédés morphologiques, est généralement régulière : on génère de nouveaux sens par métaphore, par métonymie, par extension et restriction de sens. Les travaux sur la polysémie régulière combinés avec les approches de la sémantique distributionnelle devraient nous permettre de mieux comprendre les mécanismes cognitifs d'innovation dans leur ensemble, qui touche l'ensemble des lexies, au moins potentiellement.

9.10 Éléments méthodologiques

Le travail présenté ici a suivi une méthodologie inductive combinant le souci de l'automatisation à l'intuition et l'analyse linguistique. Notre objectif principal étant la connaissance des langues, ce travail a emprunté deux voies principales pour parvenir à la modélisation des langues et la description adéquate des phénomènes spécifiques au français : d'une part l'automatisation, dont nous avons déjà dit qu'elle nous semble une *propédeutique* incontournable, à l'ère du numérique, pour valider ou corriger les modèles théoriques ou les intuitions linguistiques. Cette méthode de travail est matérialisée dans la plateforme Néoveille, qui permet une collaboration entre les traitements automatiques et les intuitions et analyses linguistiques. La correction des résultats automatiques est nécessaire, car la machine ne peut (et peut-être ne pourra jamais) gérer l'ensemble des paramètres permettant d'expliquer l'émergence d'une innovation lexicale dans une situation de communication donnée, mais la systématicité de ces traitements et son application à un nombre toujours plus élevé de documents textuels est sans doute le seul moyen qui nous permette d'accéder à cette langue qui est le fruit mouvant de la somme de toutes les expériences linguistiques de l'ensemble des membres d'une communauté linguistique. Si la répétition d'exposition à des situations linguistiques est le principe de notre apprentissage des langues, la machine sera seule capable d'approcher le résultat de cet apprentissage continuel collectif. Mais il faut intégrer aux processus

automatiques l'intuition et l'analyse linguistique humaine, car la machine peut converger vers des régularités, mais seul l'expert linguistique (et tous les membres de le communauté linguistique de manière spontanée) peut décrypter l'interprétation à donner d'un discours particulier, parce qu'il peut mobiliser toutes ses expériences passées. Cette correction humaine des traitements automatiques, réinjectée dans la machine, est donc aussi le moyen d'améliorer au fur et à mesure ces traitements.

Nous voudrions enfin signaler deux autres pistes méthodologiques, non explorées jusqu'à présent et qui nous semblent importantes à considérer pour les travaux futurs. Il s'agit des méthodes propres à la psycholinguistique et à la linguistique cognitive, d'une part, et des méthodes propres à la sociolinguistique, d'autre part. Les expérimentations, d'une part, et les enquêtes, d'autre part, nous paraissent en effet un complément utile pour assoir les descriptions et les analyses de l'innovation lexicale. Il conviendra de susciter des collaborations avec ces deux disciplines afin de renforcer les analyses linguistiques proprement dites. Une autre méthode consisterait également à utiliser l'apprentissage par les foules : étant donné la variabilité des connaissances linguistiques de chacun, un apprentissage par les foules permettrait de mieux approcher l'usage commun à l'ensemble des membres d'une communauté linguistique. Avec les outils disponibles actuellement pour de tels sondages à grande échelle, cette approche semble également prometteuse.

Bibliographie

Adam, J.-M. (1990). *Eléments de linguistique textuelle: théorie et pratique de l'analyse textuelle*. Editions Mardaga.

Adam, J.-M. (1993). Le texte et ses composantes. théorie d'ensemble des plans d'organisation. *Semen. Revue de sémio-linguistique des textes et discours*, (8).

Adam, J.-M. (2005). *La linguistique textuelle: introduction à l'analyse textuelle des discours*. Armand Colin.

Aiden, E. et Michel, J.-B. (2014). *Uncharted: Big data as a lens on human culture*. Penguin.

Alex, B. (2008). Comparing corpus-based to web-based lookup techniques for automatic english inclusion detection. *In Proceedings of LREC 2008*.

Amiot, D. (2004a). Haut degré et préfixation. *Travaux linguistiques du Cerlico*, 17:91–104.

Amiot, D. (2004b). Préfixes ou prépositions ? Le cas de sur (-), sans (-), contre (-) et les autres. *Lexique*, 16:67–83.

Andersen, H. (1988). Center and periphery: adoption, diffusion, and spread. *Historical dialectology: Regional and social*, pages 39–83.

Anthony, L. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research*, 30(2):141–161.

Antoine, F. (2000). *Dictionnaire français-anglais des mots tronqués*, volume 105. Peeters Publishers.

Arnaud, P. J. et Renner, V. (2014). English and French [NN] _N lexical units: A categorial, morphological and semantic comparison. *Word Structure*, 7(1):1–28.

Aronoff, M. (1976). *Word formation in generative grammar*. Cambridge University Press.

Austin, J. L. (1970). *Quand dire, c'est faire*. Paris, Seuil.

Baayen, H. (1992). Quantitative aspects of morphological productivity. *In Yearbook of morphology 1991*, pages 109–149. Springer.

Baayen, H. (1993). On frequency, transparency and productivity. *In Yearbook of Morphology 1992*, pages 181–208. Springer.

Baayen, R. H. (2009). *Corpus linguistics in morphology: morphological productivity*, chapitre 43, pages 900–919.

Backus, A., Dorleijn, M., Bullock, B. et Toribio, A. (2009). Loan translations versus code-switching. *The Cambridge handbook of linguistic code-switching*, pages 75–94.

Bakhtine, M. (1978). Esthétique et théorie du roman. *Paris, Gallimard*.

Baroni, M. et Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Bauer, L. (1983). *English word-formation*. Cambridge university press.

Bauer, L. (2001). *Morphological productivity*, volume 95. Cambridge University Press.

Bell, A. (1984). Language style as audience design. *Language in society*, 13(2):145–204.

Bell, A. (2001). *Back in style: reworking audience design*, pages 139–169. New York, NY: Cambridge University Press.

Benveniste, E. (1966 et 1974). *Problèmes de linguistique générale, tome I et II*. Gallimard, collection TEL.

Benveniste, E. (1970). L'appareil formel de l'énonciation. *langages*, (17):12–18.

Birkenes, M. B., Johnsen, L. G., Lindstad, A. M. et Ostad, J. (2015). From digital library to n-grams: Nb n-gram. *In Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 293–295.

Blanco, X. (2015). Les pragmatèmes: définition, typologie et traitement lexicographique. *Verbum*, 4(4):17–25.

Bloomfield, L. (1926). A set of postulates for the science of language. *Language*, 2(3):153–164.

Blumenthal-Dramé, A. (2012). *Entrenchment in usage-based theories: what corpus data do and do not reveal about the mind*. Walter de Gruyter.

Blythe, R. A. et Croft, W. (2012). S-curves and the mechanisms of propagation in language change. *Language*, pages 269–304.

Booij, G. (2005a). Compounding and derivation. *Morphology and its demarcations*, pages 109–132.

Booij, G. (2005b). Compounding and derivation: evidence for Construction Morphology. *Amsterdam Studies in the Theory and*, pages 1–23.

Booij, G. (2006). inflection and derivation. *In Encyclopedia of Language & Linguistics*, volume 5, pages 654–661.

Booij, G. (2009). Lexical Integrity as a formal universal : a constructionnist view. *In* Scalise, S., Magni, E. et Bisetto, A., éditeurs : *Universals of Language Today*, pages 83–100. Berlin:Springer.

Booij, G. (2010). Morphological analysis. *In* Heine, B. et Narrog, H., éditeurs : *The Oxford Handbook of Grammatical Analysis*, pages 1–23. Oxford University Press.

Breen, J. (2010). Identification of neologisms in japanese by corpus analysis. *eLexicography in the 21st Century: New Challenges, New Applications. Louvain: Presses universitaires de Louvain*, pages 13–22.

Britain, D. (2010). Language and space: The variationist approach. *Language and space: an international handbook of linguistic variation. Berlin: Mouton de Gruyter.*

Cabré, M. T., Domènech, M., Estopà, R., Freixa, J. et Solé, E. (2003). L'observatoire de néologie: conception, méthodologie, résultats et nouveaux travaux. *L'innovation lexicale*, pages 125–147.

Cabré, M. T., Estopà, R. *et al.* (2004). Metodología del trabajo en neología: criterios, materiales y procesos.

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

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

Carnoy, A. J. (1927). *La science du mot. Traité de sémantique.* Louvain, Éditions Universitas, VIII, 426 S.

Cartier, E. (2011a). Néologie et description linguistique pour le tal. *Langages*, (183).

Cartier, E. (2011b). Utilisation des contextes dans le cadre dictionnairique : état des lieux, typologie des contextes, exemple des contextes définitoires. *In VAN CAMPENHOUDT, Marc, LINO, Teresa, COSTA, Rute (dir.) (2011) : "Passeurs de mots, passeurs d'espoir. Lexicologie, terminologie et traduction face au défi de la diversité", Actes des Huitièmes Journées scientifiques du Réseau de chercheurs Lexicologie, terminologie, traduction, Lisbonne, 15-17 octobre 2009. Editions des Archives Contemporaines et Agence universitaire de la Francophonie*, pages 619–632.

Cartier, E. (2016a). *Distributionnalisme et sémantique: état des lieux en traitement automatique des langues*, pages 288–313. Paris: CRL.

Cartier, E. (2016b). Neoveille, système de repérage et de suivi des néologismes en sept langues. *Neologica : revue internationale de la néologie*, (10).

Cartier, E. (2016c). Semantic change tracking through the prism of distributionnalism and construction grammars : an experiment in contemporary french. *In International Conference on Construction Grammar*, Lancaster, United Kingdom. Juiz Da Fora, Brésil.

Cartier, E. (2017a). Neoveille, a Web Platform for Neologism Tracking. *In Proceedings of European Chapter of the Association for Computational Linguistics 2017, Valencia, 3-7 avril 2017.*

Cartier, E. (2017b). Néoveille, a web platform for neologism tracking : the semantic neologisms module. *In Proceedings of the 5th Electronic Lexicography in the 21st Century Conference, Leiden, 19-21 sept. 2017).*

Cartier, E. (2017c). Sémantique lexicale et distributionnalisme : éléments pour le repérage automatique du sens en corpus. *In Proceedings of the 7th Représentations du sens linguistique, Sherbrooke, 25-27 oct. 2017)*.

Cartier, E. (2018a). *Emprunts : modélisation, méthodes de repérage, résultats et analyse dans Néoveille*. Lambert-Lucas.

Cartier, E. (2018b). Neural network for formal neology detection : experiments. *In Paper submitted to XXX*, page to appear.

Cartier, E. (2018c). Néologie et noms propres : modélisation, méthodes de repérage, résultats et analyse dans néoveille. *Cahiers de lexicologie*, (XX).

Cartier, E. (2018d). Néologie sémantique : modélisation, expérimentations automatiques multilingues dans le cadre de néoveille. *In 4ème Congrès international de néologie des langues romanes (CINEO), Lyon, France, 4-6 juil. 2018*.

Cartier, E., Galand, L., Aubry, S. et Stirling, P. (2018a). Néonaute, enrichissement sémantique pour la recherche d'information. *In Atelier Recherche d'Information Sémantique, CORIA-TALN-RJC 2018, Rennes, 14-18 mai 2018*.

Cartier, E., Galand, L., Aubry, S. et Stirling, P. (2018b). Néonaute: mining web archives for linguistic analysis. *In International Internet Preservation Consortium (IIPC) Web Archiving Conference, Wellington (New-Zealand), 12-15 nov. 2018*.

Cartier, E., Sablayrolles, J.-F., Boutmgharine, N., Humbley, J., Bertocci, M., Jacquet-Pfau, C., Tallarico, G. *et al.* (2018c). Détection automatique, description linguistique et suivi des néologismes en corpus: point d'étape sur les tendances du français contemporain. *In Congrès Mondial de Linguistique Française*, volume 46, page 20 p. EDP Sciences, SHS Web of Conferences.

Cartier, E. et Sablayrolles, J.-F. c. (2009). Néologismes, dictionnaires et informatique. *Cahiers de Lexicologie*, 2008-2(93):175–192.

Cartier, E. et Viaux, J. (2017). Etude 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 contemporain. *Folia Litteraria Romanica*.

Chadelat, J.-M. et Pergnier, M. (2000). *Valeur et fonctions des mots français en anglais à l'époque contemporaine*. Editions L'Harmattan.

Chambers, J. K. et Schilling-Estes, N. (2013). *Handbook of Language Variation and Change*, volume 129. John Wiley & Sons, Inc., 2nd édition.

Charaudeau, P. (1995). Une analyse sémiolinguistique du discours. *Langages*, pages 96–111.

Charaudeau, P. (2015). De la linguistique de la langue à la linguistique du discours, et retour. *In* Engwall, G. et Fant, L., éditeurs : *Festival Romanistica. Contribuciones lingüísticas – Contributions linguistiques – Contributi linguistici – Contribuições linguísticas.*, pages 3–12. Stockholm: Stockholm University Press.

Charaudeau, P. (2017). Contrat de communication , contrat de parole. *Publictionnaire. Dictionnaire encyclopédique et critique des publics.*, pages 1–7.

Charaudeau, P. et Maingueneau, D. (2002). *Dictionnaire d'analyse du discours*. Seuil.

Clem, E. (2016). Social network structure, accommodation, and language change. *UC Berkeley Phonetics and Phonology Lab Annual Report*.

Corbin, D. (1987). *Morphologie dérivationnelle et structuration du lexique*, volume 193. Walter de Gruyter.

Corbin, D. (1992). Hypothèse sur les frontières de la composition nominale. *Cahiers de grammaire*, (17):25–55.

Corbin, D. (1999). Pour une théorie sémantique de la catégorisation affixale. *Faits de langues*, 7(14):65–77.

Coseriu, E. (1952). *Sistema, norma y habla: con un resumen en alemán*. Universidad de la República. Facultad de Humanidades y Ciencias. Instituto e Filología. Departamento de Lingüística.

Coseriu, E. (1962). *Teoría del lenguaje y lingüística general: cinco estudios*. Madrid:Gredos.

Coseriu, E. (1964). *Pour une sémantique diachronique structurale*. Centre de philologie et de littératures romanes de l'Université de Strasbourg.

Coseriu, E. (1973 [1958]). *Sincronía, diacronía e historia: el problema del cambio lingüístico*. Gredos Madrid.

Coseriu, E. (1981). Los conceptos de 'dialecto', 'nivel' y 'estilo de lengua' y el sentido propio de la dialectología". *Lingüística española actual*, , III/1(1):1–33.

Coseriu, E. (1982). *Sentido y tareas de la dialectología*. Instituto de Investigaciones Filologicas, Mexico.

Coseriu, E. (1998). Sens et tâches de la dialectologie. *In* Coseriu, E. et Wunderli, P., éditeurs : *Les cahiers dia - études sur la diachronie et la variation linguistique*, pages 17–56. Communication & Cognition, Gent, Belgium.

Coseriu, E. (2007 [1980]). Du primat de l'histoire. *Texto*, XII(2):15p.

Coseriu, E. et Polo, J. (1986). *Introducción a la lingüística*, volume 65. Gredos.

Costin-Gabriel, C. et Rebedea, T. E. (2014). Archaisms and neologisms identification in texts. *In RoEduNet Conference 13th Edition: Networking in Education and Research Joint Event RENAM 8th Conference, 2014*, pages 1–6. IEEE.

Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press on Demand.

Croft, W. (2007). Construction grammar. *In The Oxford handbook of cognitive linguistics*.

Dal, G. (2003). Productivité morphologique: définitions et notions connexes. *Langue française*, pages 3–23.

Darmesteter, A. (1874). *Traité de la formation des mots composés dans la langue française comparée aux autres langues romanes et au latin*, volume 19. A. Franck.

Deroy, L. (2013 [1956]). *L'emprunt linguistique*. Presses universitaires de Liège.

Dijk, T. A. V. (1985). *Handbook of Discourse Analysis*. Academic Press.

Divjak, D., Levshina, N. et Klavan, J. (2016). Cognitive Linguistics: Looking back, looking forward. *Cognitive Linguistics*, 27(4):447–463.

Dubois, J. (1962). *Étude sur la dérivation suffixale en français moderne et contemporain: essai d'interprétation des mouvements observés dans le domaine de la morphologie des mots construits*. Larousse.

Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual review of Anthropology*, 41:87–100.

Edwards, W. F. (1992). Sociolinguistic behavior in a detroit inner-city black neighborhood. *Language in society*, 21(1):93–115.

Evert, S. et Hardie, A. (2011). Twenty-first century corpus workbench: Updating a query architecture for the new millennium. *In Proceedings of the Corpus Linguistics 2011 conference :. University of Birmingham, Birmingham.*

Evert, S. et Hardie, A. (2015). Ziggurat: A new data model and indexing format for large annotated text corpora. *Bański, P. ; Biber, H. ; Breiteneder, E. ; Kupietz, M*, pages 21–27.

Fagyal, Z., Swarup, S., Escobar, A. M., Gasser, L. et Lakkaraju, K. (2010). Centers and peripheries: Network roles in language change. *Lingua*, 120(8):2061–2079.

Fillmore, C. J. (1977). Scenes-and-frames semantics. *Linguistic structures processing*, 59:55–88.

Fillmore, C. J. (1985). Frames and the semantics of understanding. *Quaderni di semantica*, 6(2):222–254.

Fillmore, C. J., Kay, P. et O'connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, pages 501–538.

Finkel, J. R., Grenager, T. et Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. *In Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics.

Firth, J. R. (1957). *Papers in Linguistics 1934–1951*. Oxford University Press.

Flydal, L. (1951). *Remarques sur certains rapports entre le style et l'état de langue*. Norsk tidsskrift for sprogvidenskap.

Fortis, J.-M. (2011). Comment la linguistique est (re)devenue cognitive. *Revue d'Histoire des Sciences Humaines*, 25(2):103.

Fortis, J.-M. (2012). De la grammaire générative à la linguistique cognitive : retour sur un basculement théorique. *Histoire Épistémologie Langage*, 34(1):115–154.

Fradin, B. (2003). *Nouvelles approches en morphologie*. Presses universitaires de France.

Francois, J. et Cordier, F. (2006). Psycholinguistique vs psychologie cognitive du langage : une simple variante terminologique ? *Syntaxe & sémantique*, (7):57–77.

Furiassi, C., Pulcini, V. et González, F. R. (2012). *The anglicization of European lexis*. John Benjamins Publishing.

Gadet, F. (1998). Ğcette dimension de variation que l'on ne sait nommerğ. *Sociolinguistica: Internationales Jahrbuch für Europäische Soziolinguistik= International Yearbook of European Sociolinguistics= Annuaire International de la Sociolinguistique Européenne*, (12):53–71.

Gadet, F. (2007). *La variation sociale en français*. Editions Ophrys.

Gadet, F. et Guérin, E. (2008). Le couple oral / écrit dans une sociolinguistique à visée didactique. *Le français aujourd'hui*, 162(3):21.

Gallois, C. et Giles, H. (2015). Communication accommodation theory. *The international encyclopedia of language and social interaction*, pages 1–18.

Garcia, O., Flores, N. et Spotti, M. (2017). *The Oxford handbook of language and society*. Oxford University Press.

Gardin, B., Lefèvre, G., Marcellesi, C. et Mortureux, M. F. (1974). A propos du «sentiment néologique». *Langages*, (36):45–52.

Gauchat, L. (1905). *L'unité phonétique dans le patois d'une commune*. Niemeyer.

Geeraerts, D. (2010). *Theories of Lexical Semantics*. Oxford University Press.

Geeraerts, D. et Cuyckens, H. (2007). *The Oxford handbook of cognitive linguistics*. Oxford University Press.

Giles, H. et Powesland, P. F. (1975). *Speech style and social evaluation*. Academic Press.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Goldberg, A. E. (2013). Constructionist approaches. *The Oxford handbook of construction grammar*, pages 15–31.

Gottlieb, H. (2005). *Anglicisms and translation*, pages 161–184. Clevendon : Multilingual Matters.

Granovetter, M. (1973). The strength of weak ties: A network theory revisited. *American Journal of Sociology*.

Greenberg, J. H. (1963). *Universals of language*. Cambridge University Press.

Grezka, A., Cartier, E. et Mathieu-Colas, M. (2015). Dictionnaires morphologiques du français contemporain : présentation de morfetik, éléments d'un modèle pour le tal. *In Traitement Automatique des Langues Naturelles (TALN)*, Caen, France. Université de Caen.

Guerin, E. (2008). Le'français standard': une variété située ? *In Congrès Mondial de Linguistique Française*, page 200. EDP Sciences.

Guilbert, L. (1971). De la formation des unités lexicales. *Grand Larousse de la langue française*, 1.

Gérard, C., Falk, I. et Bernhard, D. (2014). Traitement automatisé de la néologie : pourquoi et comment intégrer l'analyse thématique ? *Actes du 4e Congrès mondial de linguistique française (CMLF 2014), Berlin, p. 2627-2646.*

Halliday, M. A. et Hasan, R. (1976). *Cohesion in English.* Longman, London.

Halliday, M. A. K. (1978). *Language as social interpretation of language and meaning.* University Park Press.

Halliday, M. A. K. (2006). *Linguistic studies of text and discourse*, volume 2. A&C Black.

Halskov, J. et Jarvad, P. (2010). Automated extraction of neologisms for lexicography. *In E-lexicography in the 21st century: New challenges, new applications: proceedings of eLex 2009, Louvain-la Neuve, 22-24 october 2009*, pages 405–410.

Hardie, A. (2012). Cqpweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3):380–409.

Harris, Z. (1988). *Language and information.* Columbia University Press.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

Haspelmath, M. et Sims, A. (2013 [2002]). *Understanding morphology.* Routledge.

Hatzfeld, A. et Darmesteter, A. (1890). 1900. *Dictionnaire général de la langue française*, 2.

Haugen, E. (1950). The analysis of linguistic borrowing. *Language*, 26(2):210–231.

Hermann, P. (1886). *Principien der Sprachgeschichte.* 5th ed. 1920, Halle: Max Niemeyer Verlag.

Hippisley, A. et Stump, G. (2017). *The Cambridge Handbook of Morphology.*

Hohenhaus, P. (2005). Lexicalization and institutionalization. *In Handbook of word-formation*, pages 353–373. Springer.

Hymes, D. H. (1982). Toward linguistic competence. *Philadelphia: University of Pennsylvania, Graduate School of Education.*

Hymes, D. H. (1984). *Vers la compétence de communication.* Hatier-Crédif.

Irvine, J. T. et Gal, S. (2009). Language ideology and linguistic differentiation. *Linguistic anthropology: A reader*, pages 402–34.

Jacquet-Pfau, C. (2003). *Du statut de l'emprunt en traitement automatique des langues*, pages 79–97. Paris: Honoré Champion.

Jakobson, R. (1963). *Essais de linguistique générale.* Les Éditions de Minuit Paris.

Jakubicek, M., Kilgarriff, A., McCarthy, D. et Rychlỳ, P. (2010). Fast syntactic searching in very large corpora for many languages. *In Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation.*

Janssen, M. (2008). Neotrack, un analyseur de néologismes en ligne. *Actes du Ier Congrès International de Néologie des langues romanes (Cinéo 2008).*

Janssen, M. (2012a). Neotag : a pos tagger for grammatical neologism detection. *In LREC*, pages 2118–2124.

Janssen, M. (2012b). Neotag: a pos tagger for grammatical neologism detection. *In LREC*, pages 2118–2124.

Kang, B.-J. et Choi, K.-S. (2002). Effective foreign word extraction for korean information retrieval. *Information processing & management*, 38(1):91–109.

Kee, K. F. (2017). Adoption and diffusion. *International encyclopedia of organizational communication*, 1.

Kerremans, D. et Prokić, J. (2018). Mining the web for new words: Semi-automatic neologism identification with the neocrawler. *Anglia*, 136(2):239–268.

Kerremans, D., Stegmayr, S. et Schmid, H.-J. (2012). The neocrawler : identifying and retrieving neologisms from the internet and monitoring on-going change. *Current methods in historical semantics, Berlin etc.: de Gruyter Mouton*, pages 59–96.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlỳ, P. et Suchomel, V. (2014). The sketch engine: ten years on. *Lexicography*, 1(1):7–36.

Kilgarriff, A., Kovář, V., Krek, S., Srdanović, I. et Tiberius, C. (2010). A quantitative evaluation of word sketches. *In Proceedings of the XIV Euralex international Congress*, pages 372–379.

Kirby, J. et Sonderegger, M. (2013). A model of population dynamics applied to phonetic change. *In Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35.

Kleiber, G. (1990). *La sémantique du prototype*, volume 4. Presses universitaires de France.

Knack, R. (1991). *Ethnic boundaries in linguistic variation*, pages 251–272. Academic Press New York.

Koch, P. et Oesterreicher, W. (1985). Sprache der nähe-sprache der distanz. mündlichkeit und schriftlichkeit im spannungsfeld von sprachtheorie und sprachgebrauch. *Romanistisches Jahrbuch*, 36:15–43.

Koch, P. et Oesterreicher, W. (2001). Langage parlé et langage écrit. *In* Holtus, G., Metzeltin, M. et Schmitt, C., éditeurs : *Lexikon der Romanistischen Linguistik*, volume I/2, pages 584–627.

Koch, P. et Oesterreicher, W. (2011 [1990]). *Gesprochene Sprache in der Romania: französisch, italienisch, spanisch*, volume 31. Walter de Gruyter.

Krefeld, T. (2015). L'immédiat , la proximité et la distance communicative. *In* Polzin-Haumann, C. et Schweickard, W., éditeurs : *Manuel de linguistique française*, chapitre 11, pages 262–274. Walter de Gruyter GmbH, Berlin/Boston.

Labelle, M. (2001). Trente ans de psycholinguistique. *Revue québécoise de linguistique*, 30(September):155–176.

Labov, W. (1966). The Social Stratification of English in New York City. *Center for Applied Linguistics, Washington DC*.

Labov, W. (1972). *Sociolinguistic patterns*. Numéro 4. University of Pennsylvania Press.

Labov, W. (1994). *Principles of language change: Internal factors*. Oxford: Blackwell.

Labov, W. (2001). *Principles of language change: Social factors*. Malden, MA: Blackwell.

Labov, W., Ash, S. et Boberg, C. (2008). *The atlas of North American English: Phonetics, phonology and sound change*. Walter de Gruyter.

Lai, S. L. et Ng, V. T. (2014). Collaborative discovery of chinese neologisms in social media. *In Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on*, pages 4107–4112. IEEE.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago.

Langacker, R. W. (1987). *Foundations of cognitive grammar. Volume 1. Theoretical prerequisites*. Stanford University Press Stanford.

Lave, J., Wenger, E. et Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*, volume 521423740. Cambridge university press Cambridge.

Legallois, D. et François, J. (2011). La Linguistique fondée sur l'usage : parcours critique. *Travaux de linguistique*, 62(1):7.

Lejeune, G. et Cartier, E. (2017). Character based pattern mining for neology detection. *In Subword and Character Level Models in NLP Workshop (ScLEM), EMNLP 2017*, page to appear.

Li, W. (1994). *Three generations, two languages, one family: Language choice and language shift in a Chinese community in Britain*. Clevedon, Avon : Multilingual Matters.

Lieber, R. et Štekauer, P. (2014). *The Oxford handbook of derivational morphology*. Oxford Handbooks in Linguistic.

Lipka, L., Handl, S. et Falkner, W. (2004). Lexicalization & institutionalization: the state of the art in 2004. *SKASE Journal of Theoretical Linguistics*, 1(1):1–18.

Lippi-Green, R. L. (1989). Social network integration and language change in progress in a rural alpine village. *Language in society*, 18(2):213–234.

Loubier, C. (2011). *De l'usage de l'emprunt linguistique*. Office québécois de la langue française.

Maingueneau, D. (2005). L'analyse du discours et ses frontières. *Marges linguistiques*, (9):1–12.

Maingueneau, D. (2016). *Les termes clés de l'analyse du discours*. Le seuil.

Mallinson, C. (2007). Social Class, Social Status and Stratification: Revisiting Familiar Concepts in Sociolinguistics. *Working Papers in Linguistics*, 13(2):149–163.

McMahon, A. M. (1994). *Understanding language change*. Cambridge University Press.

Meillet, A. (1904). Comment les mots changent de sens. *L'Année sociologique (1896/1897-1924/1925)*, 9:1–38.

Mel'čuk, I. (2011). Tout ce que nous voulions savoir sur les phrasèmes, mais.... *Cahiers de lexicologie*, 102:129–150.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J. *et al.* (2010). Quantitative analysis of culture using millions of digitized books. *science*, page 1199644.

Mikolov, T., Chen, K., Corrado, G. et Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Milardo, R. M. (1988). *Families and social networks*. Sage Publications, Inc.

Milroy, J. et Milroy, L. (1978). *Belfast: Change and variation in an urban vernacular. Sociolinguistic patterns in British English*. Baltimore: University Park Press.

Milroy, J. et Milroy, L. (1985). Linguistic change, social network and speaker innovation. *Journal of Linguistics*, 21(2):339–384.

Milroy, L. (1980). Language and social networks.

Milroy, L. et Llamas, C. (2013). *Social networks*, chapitre 19, pages 407–427. Wiley Online Library.

Milroy, L. et Milroy, J. (1992). Social network and social class: Toward an integrated sociolinguistic model. *Language in society*, 21(1):1–26.

Nevalainen, T. (2015). Descriptive adequacy of the s-curve model in diachronic studies of language change. *Varieng*, 16.

Ollinger, S. et Valette, M. (2008). La créativité lexicale: des pratiques sociales aux textes. *In CINEO'08*, pages 25–40.

Olsson, F. (2009). A literature survey of active machine learning in the context of natural language processing. Rapport technique.

Puech, C. (2005). L'émergence de la notion de «discours» en france et les destins du saussurisme. *Langages*, (3):93–110.

Renner, V. (2015). Panorama rétro-prospectif des études amalgamatives. *Neologica*, (9):97–112.

Renouf, A., Kehoe, A. et Banerjee, J. (2007). Webcorp: an integrated system for web text search. *Language and Computers*, 59:47.

Reutner, U. (2017). *Manuel des francophonies*, volume 22. Walter de Gruyter GmbH & Co KG.

Rickford, J. R. (1986). The need for new approaches to social class analysis in sociolinguistics. *Language and communication*, 6(3):215–221.

Riegel, M., Pellat, J.-C. et Rioul, R. (2018). *Grammaire méthodique du français*. Presses Universitaires de France.

Riegel, M., Pellat, J.-C. et Rioul, R. (2018 [1994]). Grammaire méthodique du français. *Linguistique nouvelle*.

Ritzer, G. (2004). *Encyclopedia of social theory*. Sage publications.

Rogers, E. M. (2010). *Diffusion of innovations, fourth edition [1953]*. Simon and Schuster.

Rosch, E. (1978). *Principles of categorization. Cognition and categorization*, pages 27–48. Hillsdale, NJ: Lawrence Erbaum Associates.

Rychlỳ, P. (2007). Manatee/bonito-a modular corpus manager. *In 1st Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 65–70.

Sablayrolles, J.-F. (2000a). Lexique et processus. *Cahiers de lexicologie: Revue internationale de lexicologie et lexicographie*, (77):5–26.

Sablayrolles, J.-F. (2000b). *La néologie en français contemporain: examen du concept et analyse de productions néologiques récentes*, volume 4. H. Champion.

Sablayrolles, J.-F. (2002). Fondements théoriques des difficultés pratiques du traitement des néologismes. *Revue française de linguistique appliquée*, 7(1):97–111.

Sablayrolles, J.-F. (2010). Neologia une base de données pour la gestion des néologismes. *Actes du congrés international de néologie des langues romanes. Barcelone: UPF/IULA*, (22):757–766.

Sablayrolles, J.-F. (2017). *Les néologismes. . Créer des mots français aujourd'hui*. Petits guides de la langue française, 29, Paris, Garnier et Le Monde.

Sablayrolles, J.-F. et Pruvost, J. (2016). *Les néologismes*. Presses Universitaires de France-PUF, collection Que sais-je ?

Sagot, B. (2010). The lefff, a freely available and large-coverage morphological and syntactic lexicon for french. *In 7th international conference on Language Resources and Evaluation (LREC 2010)*.

Sagot, B., Nouvel, D., Mouilleron, V. et Baranes, M. (2013). Extension dynamique de lexiques morphologiques pour le français à partir d'un flux textuel. *In TALN-Traitement Automatique du Langage Naturel*, pages 407–420.

Sajous, F. et Hathout, N. (2015). Glawi, a free xml-encoded machine-readable dictionary built from the french wiktionary. *In Proceedings of eLex conference, Herstmonceux, England*, pages 405–426.

Saugera, V. (2017). *Remade in France: Anglicisms in the Lexicon and Morphology of French*. Oxford University Press.

Saussure, d. F. (1916). Cours de linguistique générale, ed. *C. Bally and A. Sechehaye, with the collaboration of A. Riedlinger, Lausanne and Paris: Payot*.

Saussure, d. F. (2002). *Écrits de linguistique générale*. Gallimard.

Schmid, H. (1995). Treetagger, a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43:28.

Schmid, H.-J. (2007). Entrenchment, salience, and basic levels. *The Oxford handbook of cognitive linguistics*, pages 117–138.

Schmid, H.-J. (2008). New Words in the Mind: Concept-formation and Entrenchment of Neologisms. *Anglia - Zeitschrift für englische Philologie*, 126(1):1–36.

Schmid, H.-J. (2015a). A blueprint of the entrenchment-and-conventionalization model. *Yearbook of the German Cognitive Linguistics Association*, 3(1):1–27.

Schmid, H.-J. (2015b). The scope of word-formation research. *Word-Formation. An International Handbook of the Languages of Europe. Volume 1*, pages 1–21.

Schmid, H.-J. (2016). A framework for understanding linguistic entrenchment and its psychological foundations. *Entrenchment and the Psychology of Language Learning: How We Reorganize and Adapt Linguistic Knowledge*, pages 9–36.

Schmid, H.-J. (2017). *A framework for understanding linguistic entrenchment and its psychological foundations in memory and automatization*, pages 2–24. Mouton de Gruyter.

Schultink, H. (1961). Produktiviteit als morfologisch fenomeen. *In Forum der letteren*, volume 2, pages 110–125.

Searle, J. (1972). *Les actes de langages. Essai de philosophie du langage.* Paris. Hermann.

Settles, B. (2014). Active learning literature survey. 2010. *Computer Sciences Technical Report*, 1648.

Shannon, C. E. (2001 [1948]). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55.

Sperber, D. et Wilson, D. (1989). La pertinence. *Communication et cognition*, pages 377–381.

Stefanowitsch, A. et Gries, S. T. (2003). Collostructions: Investigating the interaction of words and constructions. *International journal of corpus linguistics*, 8(2):209–243.

Štekauer, P., Valera, S. et Kőrtvélyessy, L. (2012). *Word-formation in the world's languages: a typological survey.* Cambridge University Press.

Stern, G. (1931). *Meaning and change of meaning ; with special reference to the English language.* Wettergren & Kerbers.

Testenoire, P.-Y. (2015). Ce que les théories du discours doivent à saussure. *Semen. Revue de sémio-linguistique des textes et discours*, (39).

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.

Trudgill, P., Gordon, E., Lewis, G. et Maclagan, M. (2000). Determinism in new-dialect formation and the genesis of new zealand english. *Journal of Linguistics*, 36(2):299–318.

Turpin, B. (1995). Discours, langue et parole dans les cours et les notes de linguistique générale de f. de saussure. *Cahiers Ferdinand de Saussure*, (49):251–266.

Van Dijk, T. A. (2008). *Discourse and context: A sociocognitive approach*. Cambridge University Press Cambridge.

Van Dijk, T. A. (2013). *News as discourse*. Routledge.

Warner, W. L. (1960). Social class in america. a manual of procedure for the measurement of social status.

Weinreich, U., Labov, W. et Herzog, M. (1968). Empirical foundations for a theory of language change.

Wenger, E. (1998). *Communities of practice: Learning, meaning and identity*. New Yor : Cambridge University Press.

Winter-Froemel, E. (2009). Les emprunts linguistiques: enjeux théoriques et perspectives nouvelles. *Neologica*, 3:79–122.

Winter-Froemel, E. (2013). What Does It Mean to Explain Language Change ? Usage-Based Perspectives on Causal and Intentional Approaches to Linguistic Diachrony, or: On S-Curves, Invisible Hands, and Speaker Creativity. *Energeia V*, page 123–142.