Extension dynamique de lexiques morphologiques pour le français à partir d'un flux textuel

Benoît Sagot[1] Damien Nouvel[1] Virginie Mouilleron[1] Marion Baranes[2,1]
(1) Alpage, INRIA & Université Paris-Diderot, 75013 Paris
(2) viavoo, 92100 Boulogne Billancourt
{prenom.nom}@inria.fr

RÉSUMÉ

L'incomplétude lexicale est un problème récurrent lorsque l'on cherche à traiter le langage naturel dans sa variabilité. Effectivement, il semble aujourd'hui nécessaire de vérifier et compléter régulièrement les lexiques utilisés par les applications qui analysent d'importants volumes de textes. Ceci est plus particulièrement vrai pour les flux textuels en temps réel. Dans ce contexte, notre article présente des solutions dédiées au traitement des mots inconnus d'un lexique. Nous faisons une étude des néologismes (linguistique et sur corpus) et détaillons la mise en œuvre de modules d'analyse dédiés à leur détection et à l'inférence d'informations (forme de citation, catégorie et classe flexionnelle) à leur sujet. Nous y montrons que nous sommes en mesure, grâce notamment à des modules d'analyse des dérivés et des composés, de proposer en temps réel des entrées pour ajout aux lexiques avec une bonne précision.

ABSTRACT

Dynamic extension of a French morphological lexicon based a text stream

Lexical incompleteness is a recurring problem when dealing with natural language and its variability. It seems indeed necessary today to regularly validate and extend lexica used by tools processing large amounts of textual data. This is even more true when processing real-time text flows. In this context, our paper introduces techniques aimed at addressing words unknown to a lexicon. We first study neology (from a theoretic and corpus-based point of view) and describe the modules we have developed for detecting them and inferring information about them (lemma, category, inflectional class). We show that we are able, using among others modules for analyzing derived and compound neologisms, to generate lexical entries candidates in real-time and with a good precision.

MOTS-CLÉS : Néologismes, analyse morphologique, lexiques dynamiques.

KEYWORDS: Neologisms, Morphological Analysis, Dynamic Lexica.

1 L'incomplétude lexicale et les néologismes

Tout comme les dictionnaires de langues, par définition lacunaires, les lexiques utilisés pour des applications en Traitement Automatique du langage (TAL) doivent être régulièrement complétés afin de refléter au plus près les réalités linguistiques et limiter ainsi l'incomplétude lexicale. Cependant ce processus continu de mise à jour ne peut suffire à lui seul, ne serait-ce que par le coût humain d'une telle tâche. Il est donc utile de disposer de modules d'analyse permettant

© ATALA

d'extraire automatiquement de nouvelles entrées lexicales et les ajouter, après validation manuelle ou automatique, dans des ressources lexicales. La mise au point de tels outils est plus particulièrement intéressante pour le traitement des données textuelles récentes, voire des corpus dynamiques comme un flux de dépêches d'agence, produites en temps quasi-réel.

Étant donné un outil de TAL et un texte à traiter, certains tokens [1] sont *inconnus* : à partir du lexique, l'outil ne parvient pas à les analyser comme mots-formes simples ou combinaisons régulières de tels mots-formes (par exemple, *donne-moi* est inconnu en tant que tel des lexiques de référence mais analysable comme combinaison typographique des mots-formes *donne* et *-moi*). Dans cet article, nous utilisons comme référence le Le*fff* (Sagot, 2010) et l'ensemble des mentions d'entités nommées répertoriées dans la base Aleda (Sagot et Stern, 2012) [2].

Même en se restreignant au niveau morphologique (où une entrée, ou *lemme*, peut être réduite à une forme de citation, une catégorie et une classe flexionnelle), construire automatiquement de nouvelles entrées lexicales candidates n'est pas une tâche simple. Outre la non-correspondance systématique entre tokens et formes, traiter des tokens inconnus est rendu complexe par leur grande variabilité, comme décrit par de nombreux auteurs. Adaptant ainsi la typologie des inconnus proposée par Blancafort San José *et al.* (2010), nous pouvons distinguer :

– les **tokens invalides**, induits notamment par des erreurs de tokenisation ;
– les **inconnus orthographiques**, produits de façon consciente (économie scripturale), par erreur (mauvaise connaissance de l'orthographe), ou en raison d'instabilités orthographiques (notamment pour les emprunts, les constructions préfixales ou les associations : *co-fondateur*, *coproducteur*, *microalgues*, *micro-ondes*, *électro-mécanique*, *électroencéphalogramme*) ;
– les **inconnus typographiques** (absence de tirets ou de blancs typographiques obligatoires) ;
– les **nombres, sigles** et autres unités de ce type (*A380*, *L-334-1*) ;
– les **emprunts non-adaptés**, qui ne sont pas encore rentrés dans le système morphologique de la langue et ne disposent pas encore de paradigmes morphologiques complets
– les **inconnus lexicaux**, formes correctes absentes des ressources de référence (emprunts adaptés, créations lexicales, entités nommées nouvelles ou rares, mentions inconnues d'entités connues, etc.) ; parmi eux, il convient de distinguer les mentions d'entités nommées d'une part et le reste d'autre part, que nous qualifierons de **néologismes** dans la suite de cet article [3].

En fonction de leur nature, les néologismes peuvent être considérés comme analysables (au moins une partie de l'inconnu est reconnaissable à travers sa morphologie) ou non analysables (leur forme n'est pas reconnaissable à travers leur morphologie ou leur orthographe). Dans les faits, presque tous les néologismes devraient pouvoir être analysables hors contexte, en s'appuyant sur des dictionnaires, ou en contexte, en s'appuyant sur les dépendances syntaxiques auxquelles il prend part (Han et Baldwin, 2011).

En TAL, les néologismes analysables hors-contexte à travers leur morphologie peuvent être traités à partir d'algorithmes de racinisation (Lovins, 1968). Ceci permet de rattacher un néologisme à d'autres unités lexicales connues des ressources de référence (par exemple, *zippable* à *zipper*). Il

1. Un token est défini comme une unité typographique constituée d'un caractère de ponctuation ou d'une séquence d'au moins un caractère ne comportant pas d'espace et délimitée par des espaces et/ou des caractères de ponctuation.

2. Dans ce travail, nous laissons de côté les inconnus contextuels, c'est-à-dire les tokens qui ne sont connus de la référence que comme composants de composés mais qui apparaissent dans d'autres contextes que ces composés (par exemple, *instar* si on le trouvait ailleurs que dans le composé *à l'instar de/du*.

3. Nous considérons donc comme étant un néologisme toute unité lexicale valide qui est nouvelle par rapport aux lexiques de référence, et non, comme c'est souvent le cas, par rapport à un usage supposé connu et vérifiable. Puisqu'il ne s'agit pas d'inconnus, nous ne traitons pas non plus des cas où une forme graphique connue est employée avec une catégorie inconnue du lexique (conversion) ou avec un sens nouveau (néologie sémantique).

© ATALA

est également possible de déduire les catégories morphosyntaxiques des néologismes à partir de propriétés de ses affixes morphologiques, si l'on parvient à les identifier. Dans le cas de *zippable*, par exemple, le suffixe *–able* est un bon indicateur de la catégorie adjectif et de la classe flexionnelle marquant le pluriel par un suffixe *–s*. Enfin, le lemme d'un néologisme peut être obtenu par analogie avec les entrées de la référence, par consultation de ressources complémentaires, ou à l'aide de lemmatiseurs (Schmid, 1994; Chrupała *et al.*, 2008).

Ces méthodes d'identification et d'analyse peuvent être combinées à d'autres systèmes de filtrage ou de traitement qui prennent en compte ou non le contexte. En linguistique de corpus, il s'agit généralement de descriptions formalisées sous forme de dictionnaires, et de transducteurs à états finis (Maurel et Piton, 1998; Dister et Fairon, 2004). Ces formalismes sont plus ou moins puissants en fonction de l'organisation des transducteurs, des types de dictionnaires associés et de la variété des traits qu'il est possible d'utiliser à travers eux.

Mais de telles approches supposent que l'on ait su identifier les néologismes parmi l'ensemble des inconnus. Si dans certains cas il s'agit d'une tâche aisée (sigles, nombres), distinguer un néologisme d'un inconnu orthographique ou d'un emprunt non-adapté est moins immédiat. Dans cet article, notre objectif est triple : (1) mettre en évidence les phénomènes constructionnels dont procèdent les néologismes, (2) montrer qu'il est possible d'identifier et d'analyser automatiquement ces néologismes, et (3) étendre ainsi automatiquement le lexique morphologique de référence, ici le Le*fff* .

Nous présentons en partie 2 l'architecture que nous adoptons pour étudier les tokens inconnus. Parmi ces derniers, la partie 3 étudie les mécanismes morphologiques de construction des néologismes que nous relevons. Nous décrivons en partie 4, après un état de l'art, les modules TAL qui nous permettent de traiter ces éléments. Enfin, nous conduisons une évaluation dont les résultats sont présentés en partie 5.

2 Traitement des inconnus dans le corpus

2.1 Architecture d'identification et d'analyse des inconnus

Traiter automatiquement les tokens inconnus afin d'enrichir le lexique nécessite au préalable la mise en place d'une architecture logicielle robuste. La figure 1 présente l'organisation générale des traitements utilisés et, pour certains, développés spécifiquement au sein de la chaîne de traitement SxPipe (Sagot et Boullier, 2008). Nous réalisons en préliminaire une étape de filtrage (`Filtres`), que nous évoquerons à la section suivante lors de la description du corpus.

Nous appliquons ensuite certains modules de prétraitement de SxPipe [4] (Sagot et Boullier, 2008). Nous nous restreignons ici à la tokenisation du texte, à la détection de motifs par automates (nombres, dates, sigles) et à la reconnaissance d'entités nommées à l'aide de la base Aleda (Sagot et Stern, 2012) et de quelques motifs contextuels. Nous obtenons finalement des treillis de formes à partir desquels les modules implémentés pour le traitement des inconnus peuvent opérer. On peut noter que les ambiguïtés d'analyse ainsi créées ne concernent jamais les tokens inconnus qui font l'objet des traitements ultérieurs.

4. Parmi les options disponibles, nous désactivons celles qui cherchent à corriger les fautes d'orthographe ou qui décomposent la reconnaissance des tokens (par dérivation ou par composition).

© ATALA

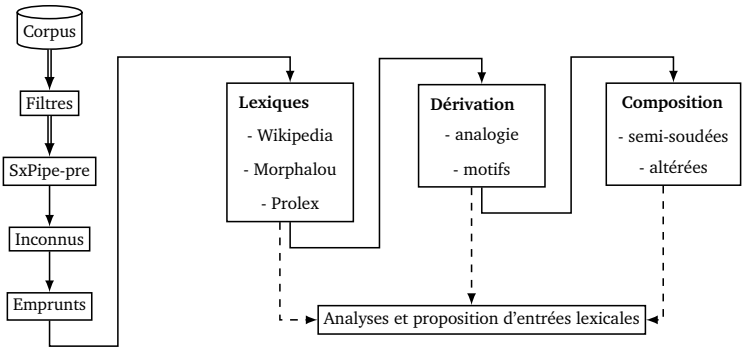

FIGURE 1 – Chaîne de traitement des inconnus

Le module `Inconnus` identifie les tokens inconnus et les étiquette comme tels. Comme nous l'avons évoqué, ces derniers peuvent relever de nombreuses catégories. Le module `Emprunts` (Baranes, 2012) permet d'écarter les tokens empruntés à l'anglais non-adaptés [5]. Parmi les inconnus restants, nous cherchons à repérer les formes qui auraient intérêt à être ajoutées au lexique, soit parce que ce dernier n'est pas suffisamment complet (`Lexiques`), soit parce que ce sont des créations lexicales (en particulier ceux créés par `Dérivation`, ou par `Composition`).

2.2 Données : le flux de dépêches AFP

Nous conduisons nos études, expériences et évaluations sur un volumineux corpus de dépêches AFP (francophones), collectées entre 2007 et 2013. Nous en sélectionnons trois sous-parties afin de mener nos expériences : des dépêches entre le 24 juin et le 3 juillet 2009 (`AFP-annot`), l'intégralité des dépêches de l'année 2009 (`AFP-2009`) et 200 dépêches tirées au hasard entre le 1er et le 14 janvier 2013 (`AFP-eval`). L'opération de filtrage consiste à écarter les énoncés ne comportant pas assez de caractères en minuscules ou pas assez de mots. Cela permet d'éliminer les tableaux de résultats sportifs, sommaires, agendas, signatures et autres éléments qui ne sont pas à proprement parler du contenu linguistique.

Le tableau 1 donne les caractéristiques générales de ces corpus. On peut constater que les occurrences d'inconnus sont d'autant plus redondantes que le sous-corpus est grand. Les corpus `AFP-annot` et `AFP-2009` sont utilisés à fins d'études. En particulier, `AFP-annot` est annoté manuellement en inconnus selon la classification de Blancafort San José *et al.* (2010) [6]. Nous écartons certaines classes d'inconnus de cette étude (mots commençant par des chiffres ou des majuscules) afin de se focaliser sur les créations lexicales. Le tableau 2 indique la répartition des inconnus selon ces classes. Nous y vérifions l'importance du phénomène de créations lexicales, que nous assimilons aux néologismes et sur laquelle nous concentrons nos efforts.

La figure 2 nous renseigne sur l'évolution des inconnus distincts (repérés par la chaîne de

5. Des néologismes empruntés à des formes anglophones peuvent alors ne pas être repérés (*cardio-training*, *box-office*, etc.), mais ces erreurs représentent moins de 1% des tokens inconnus.

6. Le travail d'annotation manuelle a été réalisé sous la responsabilité et avec les outils de l'entreprise Syllabs, dans le cadre du projet ANR EDyLex.

© ATALA

Corpus	Dépêches	Tokens	Inconnus	Distincts
AFP-annot	2 535	1 060 378	6 208	2 782
AFP-2009	311 981	94 967 771	907 570	107 496
AFP-eval	200	73 353	729	489

TABLE 1 – Volumes

Type	Inconnus	Distincts
Créations lexicales	3556	1301
Erreurs typographiques	902	622
Formes lexicalisées	531	272
Emprunts adaptés	480	174
Forme avec clitique	302	222
Entités nommées	227	114
Mots étrangers	177	108
Formes non-autonomes	141	43
Variantes graphiques	55	27
Tokens alphanumériques	42	18
Inclassables	10	7

TABLE 2 – Inconnus de AFP-annot

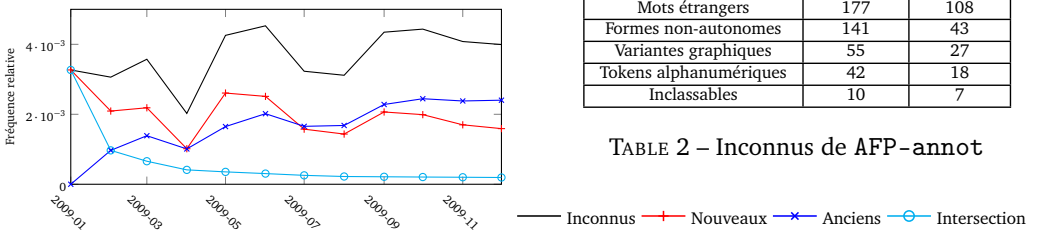

FIGURE 2 – Temporalités des inconnus distincts

traitement) par mois au sein du corpus AFP-2009 en fréquence relative[7]. Au fil de l'année, les accumuler permet d'indiquer pour chaque mois le nombre de nouveaux inconnus à traiter (*Nouveaux*), et leur intersection depuis le début de l'année (*Intersection*). Malgré les volumes de données que nous manipulons, il semble que le nombre de nouveaux inconnus apparaissant tous les mois diminue relativement peu. Cette apparition continuelle de nouvelles entrées, dans un corpus aussi contrôlé que des dépêches AFP, nous confirme la nécessité de mettre en place des mécanismes dynamiques pour traiter ces éléments.

3 Analyse linguistique des néologismes

Il existe de nombreuses formes de néologismes qui ne sont pas toutes aussi faciles à identifier comme telles. Sablayrolles (1997) recense une centaine de typologies plus ou moins profondes des néologismes. Sablayrolles *et al.* (2011) proposent un classement général des néologismes en trois classes qui correspondent aux phénomènes de glissement de sens, aux phénomènes d'affixation et de composition ainsi qu'aux phénomènes d'économie du langage. Il propose également un classement plus fins en 24 classes.

De nombreux travaux ont permis d'élaborer des outils afin d'étudier et de décrire la néologie. Le logiciel *SEXTAN* (Cabré *et al.*, 2003) permet d'identifier automatiquement les inconnus et de les présenter à un lexicographe. Le laboratoire LDI dispose d'une plateforme de traitement des néologismes, *NEOLOGIA* (Cartier, 2011), qui permet de collecter et rechercher des néologismes dans la base de données, d'en observer l'évolution (depuis l'état néologique jusqu'à celui de mot intégré dans les dictionnaires) et, dans certains cas, la disparition.

Les néologismes issus d'emprunts sont étudiés depuis une quarantaine d'années, notamment à travers les travaux de L. Guilbert dans les années 70 puis J. Rey-Debove et M. Yaguello dans les années 80. Aujourd'hui, de nombreuses recherches s'intéressent plus particulièrement à l'adaptation des formes empruntées (Anastassiadis-Syméonidis et Nikolaou, 2011). Walther et

7. La fréquence est rapportée aux nombre de tokens par mois.

© ATALA

Sagot (2011) abordent ainsi la question de l'adaptation de verbes néologiques issus de l'anglais.

Dans le cadre de la définition des néologismes donnée plus haut, nous nous intéressons dans cette partie aux inconnus morphologiquement analysables, et notamment aux créations lexicales formées par dérivation ou par composition (et éventuellement altération) de mots connus.

3.1 Mécanismes de création lexicale

3.1.1 Formes créées par dérivation

Par dérivation nous faisons référence aux phénomènes d'affixation. L'affixation correspond à l'ajout de morphèmes à gauche (préfixation) ou à droite (suffixation) d'un mot connu. Ces mécanismes permettent la construction de néologismes dont la signification est, généralement, immédiatement compréhensible. En cas de suffixation, la catégorie peut être modifiée (énorme → énormitude) [8]. Nous classons parmi les préfixes, les formes « e », « i », « web » et « cyber ». Notons que ces dernières sont parfois décrites, selon les auteurs comme « *combining forms* » (Ahronian et Béjoint, 2008) ou formes hybrides (Sablayrolles *et al.*, 2011).

		Nb. d'occ. (+tiret)	Nb. de formes distinctes	Exemples	
Préfixes	anti(-)	3371(+5986)	171 (+435)	anticrise	anti-fraude
	co(-)	892 (+2809)	54 (+183)	corapporteur	co-skippeurs
	re(-)	2424 (+36)	128 (+14)	rescolarisés	remariage
Suffixes	isation(s)	1992	110	talibanisation	masterisation
	iste(s)	3265	223	jihadiste	lefebvristes
	eur(s)	10192	353	skippeurs	performeur
	naute(s)	35	7	nutrinaute	mobinautes
	itude(s)	19	4	bravitude	merditude

TABLE 3 – Affixes les plus fréquents dans les dépêches AFP

La table 3 donne, pour quelques exemples, les occurrences que nous relevons dans le corpus `AFP-2009`. L'étude systématique de préfixes et de suffixes nous permet d'en constituer une liste utilisée comme motifs pour l'analyses des néologismes.

3.1.2 Formes créées par composition

Comme nous le verrons en section 4, les mécanismes de composition représentent une part importante de néologismes que nous repérons. Tout d'abord, un composé peut être simplement créé par association de constituants. Dans ce cas, les mots sont concaténés (généralement par trait d'union) afin de n'en former plus qu'un. Nous relevons en particulier : (i) les compositions ADJ+N, N+ADJ et ADJ+ADJ = ADJ ou N, (ii) les compositions N+N = N, (iii) les compositions V+N = N. Certains adjectifs sont très productifs pour composer les formes comme en (i). C'est le cas par exemple des adjectifs *super* (ex. *super-héros*) ou *mini* (ex. *mini-chaîne*). Les compositions à partir de noms communs (ii) permettent de générer d'autres noms composés. Les compositions issues de verbes et de noms communs (iii) permettent de créer des noms communs qui font

8. Certains mots suffixés peuvent également être classés dans les constructions par apocopes qui eux-mêmes peuvent être associés à des mots-valises.

© ATALA

référence, par exemple (Villoing, 2003) à des instruments (*ouvre-lettre*), des lieux (*coupe-gorge*), des agents (*gratte-papier*), des procès (*lèche-vitrine*), etc.

Dans les composés créés à partir de deux adjectifs ou plus (chiraco-villepinistes), les premiers composants manifestent la perte de leur autonomie au profit du dernier adjectif par un mécanisme d'altération en *–o*. Ce mécanisme permet notamment la construction de termes dans des domaines de spécialités (*socialo-communiste*), ou d'adjectifs formés avec des gentilés (*franco-allemand*).

On peut enfin relever le cas particulier des compositions dans lesquelles le premier adjectif est en réalité une base latine ou grecque munie de ce morphe *–o* [9]. Ces composés sont souvent des termes savants du domaine médical (*cardio-vasculaire*).

Créations lexicales		Nb d'occ.	Formes distinctes	Exemples
par association avec altération	composé d'un gentilé	7797	616	américano-taiwanais
	autres	2956	239	politico-judiciaire
	total	10753	855	chiraco-villepinistes
par association sans altération	–	284	109	satiristes-polémistes

TABLE 4 – Formes inconnues créées par association dans les dépêches AFP

Le tableau 4, qui résume les études sur corpus sur la composition, montre la forte productivité des composés, notamment des formes en *–o* , qui demandent un traitement spécifique.

4 Analyse morphologique automatique des néologismes

4.1 État de l'art

Plusieurs approches peuvent permettre d'analyser les néologismes par dérivation ou par composition. Nous nous appuyons sur les travaux en morphologie qui s'intéressent au regroupement de mots en familles morphologiques [10]. C'est le cas de Bernhard (2010) qui a mis en place deux systèmes d'apprentissage non supervisés (*MorphoClust* et *MorphoNet*) ou de Hathout (2010) dont le système (*Morphonette*) mêle analogie et informations sémantiques. Dans certains cas, les travaux portent sur la prédiction de formes potentielles de la langue (Neuvel et Fulop, 2002) ou afin de compléter un quadruplet d'analogie (Lepage, 1998).

La plupart des systèmes décrits dans la littérature mettent l'accent sur l'analyse de la construction d'un mot. Le système à base de règles (créées manuellement) *Dérif* (Hathout et Namer, 2011) détermine les éléments à partir desquels sont construits des unités lexicales, par dérivation ou par composition. D'autres systèmes réalisent cette tâche de manière non supervisée, par exemple par analogie formelle (Lavallée et Langlais, 2011) ou par segmentation (Goldsmith, 2001; Creutz et Lagus, 2005). En ce qui concerne les mécanismes compositionnels, si Mathieu-Colas (2010) se penche sur la création lexicale par trait d'union, nous n'avons pas connaissance de systèmes implémentés spécifiquement à leur sujet.

Certains travaux montrent qu'il est possible d'ajouter à l'analyse de la construction d'un mot la prédiction de son lemme et de ses traits morphologiques en s'appuyant sur un système

9. Ces « compositions néoclassiques » existent en français mais ne correspondent plus aux formes d'origines.
10. Une famille morphologique est composée de mots partageant une base lexicale commune (*écrit, écrire, écrivain,...*).

© ATALA

d'apprentissage supervisé couplé à de l'analogie (Stroppa et Yvon, 2006). Disposer de ces informations supplémentaires ont, par exemple, permis à Dal et Namer (2000) (avec *GéDéRif*), ainsi qu'à (Tanguy et Hathout, 2002) (avec *Webaffix*), de proposer un système qui, pour chaque forme nouvelle, calcule ses dérivés et vérifie leur validité sur internet.

Cependant, mis à part Mikheev (1997), peu de travaux étudient spécifiquement l'élaboration et l'évaluation de systèmes de complétion d'un lexique. Dans notre travail, nous mettons en place des outils destinés à traiter des documents susceptibles de contenir de nombreux néologismes, tels que définis en introduction. Il nous faut donc distinguer ces derniers parmi les tokens inconnus, déterminer et évaluer les mécanismes qui permettent de les analyser automatiquement.

4.2 Recherche dans des lexiques externes

Les notions d'inconnu et de néologisme étant définies ici par rapport à un lexique de référence, le Le*fff* , la façon la plus simple de les traiter consiste à les rechercher dans d'autres ressources lexicales librement disponibles. Nous avons fait appel au Wiktionnaire (`fr.wiktionary.com/`), dictionnaire collaboratif, à Morphalou (Romary *et al.*, 2004), lexique morphologique extrait du TLFi, et à ProLexBase (Maurel, 2008), base de noms propres incluant de nombreux gentilés.

Toutefois, l'enrichissement du Le*fff* avec des entrées lexicales manquantes extraites de ces ressources nécessite de rendre ces dernières compatibles avec le Le*fff* , en les transformant en un inventaire d'entrées lexicales associant une forme de citation à une des classes flexionnelles du Le*fff* . Pour chacune de ces trois ressources, nous avons donc construit des outils de conversion automatique vers le formalisme Alexina, puis avons projeté les classes flexionnelles obtenues vers celles utilisées par le Le*fff* . Ce processus, bien que nécessairement imparfait, a permis de détecter des erreurs dans les trois ressources d'origine [11]. Le Wiktionnaire, Morphalou et ProLexBase ont été ainsi transformés en des lexiques Alexina de même grammaire morphologique que le Le*fff* et comprenant respectivement environ 1 million, 400 000 et 125 000 entrées produisant au total 1 100 000 formes fléchies distinctes, parmi lesquelles 700 000 ne sont pas couvertes par le Le*fff* .

Ainsi, un module dédié recherche les néologismes dans ces ressources, et propose autant d'analyses qu'il y trouve d'entrées. Au sein du corpus `AFP-2009`, 18,6% des inconnus analysés distincts le sont grâce à ce module, parmi lesquels 71,5% sont trouvés dans le Wiktionnaire, 32,0% dans Morphalou et 14,9% dans ProLexBase (une entrée pouvant se trouver dans plusieurs lexiques en même temps). Notons que, même s'ils ne sont pas utilisés par les modules présentés ci-après, ces lexiques sont disponibles pour l'ensemble de la chaîne de traitement SxPipe.

4.3 Néologismes construits par dérivation

4.3.1 Analyse par analogie

Comme indiqué en partie 3.1, nous analysons les néologismes construits par dérivation comme l'application de règles d'affixation sur une entrée existante du Le*fff* (ex : *divulgable–divulgation*) [12].

11. Par exemple parce qu'une entrée lexicale se retrouve à associer une forme de citation avec une classe flexionnelle que la grammaire morphologique du Le*fff* considère comme incompatible

12. Le Le*fff* ne comportant pas de noms propres, notre chaîne ne permet pas l'analyse de dérivés dont la base est un nom propre, tels que *zlataner* ou *sarkozysme*.

© ATALA

Le module décrit ici s'inspire de travaux sur l'analogie appliquée à la morphologie. Cette notion, décrite dans les travaux cités en section 4.1, permet d'établir un rapport entre deux paires d'éléments : x est à y ce que z est à t, noté $x : y :: z : t$. Pour la néologie, nous recherchons des règles d'affixation communes à des paires d'entrées du Le*fff*, qui nous permettent de déduire des informations pour des néologismes donnés. Dans le cas de *divulgable*, nous pouvons déduire qu'il s'agit d'un dérivé si nous trouvons conjointement (i) *divulgation* dans le Le*fff* et (ii) une règle de substitution du suffixe *–able* en *–ation* (extraite d'entrées du Le*fff* comme *acceptable-acceptation*). Ce type d'analyse nécessite donc un apprentissage des règles morphologiques. Le nôtre, faiblement supervisé, se fait en trois étapes :

1. Nous apprenons les règles de construction à partir des formes fléchies du Le*fff* en étudiant tous les couples (forme de citation, forme fléchie — reliée ou non à la forme de citation —) qui ont une partie commune d'au moins 5 caractères et qui ne diffèrent que par un suffixe ou par un préfixe (en sélectionnant les règles de fréquence ≥ 40).
2. Les règles sont utilisées pour grouper les entrées du Le*fff* (supposées partager une même base lexicale) afin de constituer des paires de formes accompagnées de règles de transformation pour passer de l'une à l'autre.
3. Cette seconde étape permet d'établir des paires de formes x, y (forme de citation, forme fléchie reliée par flexion ou dérivation), qui vont nous servir à construire des relations analogiques impliquant un inconnu t, relations de la forme $x : y :: t : z$. Pour ce faire, nous remplaçons chaque paire de formes par une règle de réécriture reliant un couple préfixe/suffixe d'input à un couple préfixe/suffixe d'output, ce qui permet de traiter les préfixations, les suffixations, et les dérivés parasynthétiques (cf. table 5). Chacune de ces règles indique la catégorie et les traits morphologiques (genre, nombre) obtenus. De surcroît, en étudiant les couples de mots leur donnant naissance, nous catégorisons chaque règle comme flexionnelle ou dérivationnelle. Nous ne conservons que les règles morphologiques qui ont plus de 80 occurrences (32 508 règles distinctes) dont quelques exemples sont montrés en table 5.

Cat.	Préfixe	Suffixe	Occ	Type	Exemple
adj_Kfp → v_W	_	ées → er	6483	Dérivation	données → donner
v_I12s → nc_fs	_ → dé	is → tion	116	Dérivation	valorisais → dévalorisation
v_P2p → v_W	_	z → r	7074	Flexion	dansez → danser

TABLE 5 – Exemples de règles affixales apprises

Nous sommes ainsi en mesure de déterminer si un inconnu, analysable par ces règles, est une création lexicale. En effet, si nous parvenons à le relier ainsi à une entrée du Le*fff* grâce à une règle dérivationnelle [13]. Le lemme (forme de citation + classe flexionnelle) est obtenu en appliquant l'outil intégré au Le*fff* permettant de calculer pour un triplet (forme, catégorie, étiquette morphologique) l'ensemble des lemmes morphologiquement compatibles avec la grammaire morphologique du Le*fff* d'une part et avec le triplet d'autre part. L'application, en parallèle, de l'étiqueteur morphosyntaxique MElt (Denis et Sagot, 2012) permet alors de ne conserver que les lemmes dont la catégorie est la même que celle proposée par MElt pour l'inconnu (s'il y en a plusieurs, on choisit le mieux pondéré, s'il n'y en a aucun, on déclare l'inconnu inanalysable par ce module). La sortie de notre module, présentée en table 6 nous permet de proposer de

13. Si la règle est flexionnelle, nous considérons qu'il s'agit d'une faute d'orthographe et non d'un néologisme : le Le*fff* étant supposé comporter toutes les flexions possibles, le mot est fléchi selon une classe flexionnelle erronée (*travails* comme pluriel de *travail*).

© ATALA

Composé	Règles appliquée	Famille morphologique	Cat., flexion	F. de citation
blablatons	-ons→-age	blablatage	V_P1p, v-er ; V_Y1p, v-er	blablater
décrocheurs	-eurs→-er, -eurs→-age	décrocher, décrochage...	NC_mp, nc-2m	décrocheur

TABLE 6 – Analyse des dérivés

nouvelles entrées lexicales et permet d'analyser 11,9% des inconnus distincts présents dans le corpus AFP-2009.

4.3.2 Mise au point de motifs dédiés

L'étude menée en section 3.1 nous a permis d'isoler et de décrire certains phénomènes de création lexicale pour lesquels nous mettons au point des mécanismes d'analyse dédiés. En particulier, parmi les préfixes considérés, une proportion importante correspond à un mécanisme de dérivation qui ne modifie pas la catégorie morphosyntaxique et la classe flexionnelle du lemme de base. Ainsi, nous mettons en place un module qui, pour un inconnu donné, recherche un des préfixes standard [14] et vérifie si le composant à droite (concaténé, éventuellement par trait d'union) de ce préfixe est un mot connu du L*efff* . Si tel est le cas, une analyse est proposée qui construit le lemme complet à partir de l'entrée trouvée. Dans le corpus AFP-2009, 16,6% des inconnus analysés distincts le sont grâce à ce module. Un mécanisme similaire est implémenté pour les suffixes *iste*, *isme* et *isation* mais ne traite qu'un très faible nombre d'inconnus (0,2%).

4.4 Analyse des néologismes construits par composition

Afin de traiter les expressions construites par composition, nous nous focalisons en première approche sur la composition marquée par un ou plusieurs tiret(s) '-' dont nous cherchons à décrire la morphologie. Nous sommes alors en mesure d'en identifier simplement les composants et d'interroger le L*efff* pour y rechercher, par ordre de préférence :

(i) l'expression dans laquelle les tirets sont remplacés par un blanc (expression multi-mots),

(ii) chaque composant séparément (composition),

(iii) s'il n'y a que deux composants, le dernier composant et les formes de citation ayant un préfixe commun [15] avec le premier composant (composition avec altération).

Comme l'étude linguistique en partie 3.1 l'a suggéré, dans une grande majorité de cas, le dernier composant impose sa catégorie et sa classe flexionnelle au néologisme construit. S'il y a ambiguïté, nous utilisons la catégorie proposée par l'étiqueteur morphosyntaxique MElt appliqué en parallèle, pour ne conserver que les analyses qui sont compatibles avec cette catégorie (contrairement à la section précédente, nous conservons ici toutes les analyses produites si MElt n'en a proposé aucune compatible avec les entrées suggérées). Nous conservons ainsi ces informations pour proposer la catégorie et la classe flexionnelle de la nouvelle entrée lexicale. Enfin, la forme de citation est construite à partir de tous les composants d'origine, sauf le dernier qui est remplacé par la forme de citation de l'entrée lexicale du L*efff* trouvé.

14. *agri*, *anti*, *après*, *archi*, *contre*, *cyber*, *dé*, *demi*, *dés*, *e*, *ex*, *extra*, *grand*, *hyper*, *im*, *in*, *inter*, *intra*, *mal*, *maxi*, *méga*, *méta*, *mi*, *mini*, *multi*, *non*, *outre*, *para*, *péri*, *pluri*, *poly*, *post*, *pré*, *quart*, *quasi*, *re*, *ré*, *sans*, *semi*, *sous*, *sub*, *super*, *supra*, *sur*, *télé*, *tiers*, *trans*, *ultra*, *uni*, *vice*, *co*.

15. Ce préfixe doit contenir au moins la moitié du composant.

© ATALA

Composé	Analyse(s) en composants	Cat., flexion	Forme de citation
centre-ville	(a) centre ville (NC_mp, nc-2m)	NC_mp, nc-2m	centre ville
député-maire	(b) député (NC_ms, nc-4) + maire (NC_ms,nc-4sse)	NC_ms,nc-4sse	député-maire
lumino-technique	(c) lumino(lumineux) (ADJ_s,adj-ique2) + technique (NC_fs,nc-2f)	NC_fs,nc-2f	lumino-technique

TABLE 7 – Analyse de composés

Nous remarquons qu'en (iii), l'analyse peut fournir de nombreuses hypothèses distinctes (forme de citation, catégorie, classe flexionnelle), notamment lors de la recherche par préfixe commun. Pour pallier cela, nous ajoutons des contraintes dans ce cas : nous nous restreignons aux expressions formée selon le motif *–o* qui décrit correctement, notamment, la composition adjectivale. La table 7 donne quelques exemples de décompositions réalisées selon ces principes.

Dans le cas général, nous remarquons la difficulté de traiter de telles expressions lorsqu'ils sont amalgamés sans marqueurs de séparation (ni espace), puisque l'on tombe dans le cas difficile des mots composés standards (Sag *et al.*, 2002), mais notre étude sur corpus a montré que ce phénomène était marginal.

Parmi les inconnus distincts traités par nos modules, celui-ci en analyse 57,7% dans le corpus `AFP-2009`. Ce chiffre important est lié à la productivité des mécanismes de création lexicale. Ceci nous confirme que l'incomplétude lexicale relève de mécanismes au-delà de la morphologie dérivationnelle et que des moyens peuvent être mis en œuvre pour permettre leur analyse.

5 Construction et évaluation d'entrées lexicales néologiques

Tous les modules que nous avons décrits, lorsqu'ils parviennent à analyser un inconnu, fournissent pour chaque élément son lemme, c'est-à-dire sa forme de citation, sa catégorie et sa classe flexionnelle. En conséquence, nous sommes en mesure de récupérer les analyses produites afin de proposer de nouvelles entrées à ajouter au lexique. L'ordre de ces modules importe : le premier qui parvient à analyser un inconnu interrompra le processus d'analyse. Comme nous le verrons ci-dessous, la précision des modules d'analyse nous permet d'examiner les analyses dès la première occurrence.

Comme indiqué plus haut, l'objectif de ce travail est double : construire des entrées lexicales flexionnelles à ajouter au L*efff* afin d'en augmenter la couverture sur les dépêches AFP de façon dynamique, mais également extraire des informations constructionnelles concernant ces nouvelles entrées, afin de permettre des traitements ultérieurs (sémantique lexicale y compris pour des applications en TAL, étude quantitative des mécanismes de création lexicale, etc.). Nous avons donc procédé à une évaluation en trois étapes, qui vise à répondre aux questions suivantes pour chaque occurrence d'inconnu : a-t-elle été correctement identifiée comme étant ou n'étant pas un néologisme ? si oui, l'entrée lexicale proposée est-elle correcte, y compris sa classe flexionnelle afin de pouvoir produire correctement ses formes fléchies ? si oui, les informations constructionnelles associées sont-elles correctes ?

Pour cela, nous traitons les inconnus contenus dans le corpus `AFP-eval` tel que décrit en partie 2. Parmi les 489 inconnus distincts, 449 (soit 92%) ont été correctement classés, dont 357 qui ne sont pas des néologismes et 92 néologismes. Les 40 inconnus restant (8% du total) ont

© ATALA

été mal classés, dont 34 néologismes : seulement 6 inconnus ont été analysés à tort comme des néologismes (et ont donc donné lieu à des entrées lexicales candidates erronées). Nous obtenons donc pour la tâche de détection des néologismes une précision de 94% (92/(92+6)) et un rappel de 73% (92/(92+34)), et pour la tâche complémentaire de détection des inconnus non-néologiques une précision de 91% (357/(357+34)) et un rappel de 98% (357/(357+6)).

Les 98 inconnus détectés comme néologismes, y compris les 6 classés par erreur, ont conduit à la création de 93 entrées lexicales candidates, c'est-à-dire d'entrées (forme de citation, catégorie, classe flexionnelle) qui permettent de construire automatiquement toutes les formes fléchies correspondantes. Nous avons évalué cette liste manuellement avec les résultats suivants : 73 sur 93, soit environ 80%, sont totalement correctes, 5 ont la bonne catégorie mais pas la bonne classe flexionnelle (ainsi *point-presse*, considéré comme féminin et prenant un *s* au pluriel), 13 n'ont pas la bonne catégorie (mais souvent les bonnes formes fléchies, car il s'agit fréquemment de confusions nom/adjectif, par exemple *multi-facette*), 1 est douteuse et 1 est totalement erronée (le verbe *multi-voir* pour l'adjectif *multi-vues*).

Parmi les 73 entrées lexicales correctes, 52 ont été construites par l'un de nos modules d'analyse, et non au moyen de lexiques externes. Pour ces 52 entrées nous disposons donc d'informations constructionnelles, de nature à permettre le calcul d'informations supplémentaires telles que la valence ou la sémantique lexicale [16]. Ainsi, ayant correctement analysé *co-attribuer* comme issu d'une dérivation préfixale à partir du verbe *attribuer*, nous pouvons d'une part associer à *co-attribuer* les mêmes informations de valence que celles dont on peut disposer dans un lexique comme le L*efff* , et d'autre part savoir que le sens de *co-attribuer* peut se construire compositionnellement à partir de celui du préfixe *co-* et de celui d'*attribuer* [17]. Nous avons étudié manuellement les informations constructionnelles obtenues au cours du processus d'analyse. Pour cette évaluation, celles-ci se sont toujours avérées correctes. Par exemple, le module d'analyse des dérivés par analogie a correctement relié le verbe néologique *galvaniser* au nom *galvanisation*. Le module d'analyse des composés a su analyser *politico-judiciaire* comme formé par la composition des adjectifs *politique* (avec altération) et *judiciaire*.

6 Conclusion

Le traitement de flux continu de dépêches d'actualité nécessite de maintenir un lexique à jour aussi dynamiquement que possible. Face à cette problématique, nous avons mis au point une chaîne de traitement qui isole les éléments relevant de l'incomplétude lexicale, nous étudions les mécanismes néologiques liés à leur création et implémentons des modules dédiés pour leur analyse morphologique. Ce processus nous permet de récolter des informations (morpho-syntaxiques et flexionnelles), selon la morphologie des néologismes analysés. Nous sommes ainsi en mesure de proposer des entrées lexicales à ajouter au lexique, dès leur première occurrence et avec une très bonne précision.

En perspectives, ce travail pourra donner lieu à des études, à plus grande échelle et en temporalité, des néologismes, ainsi qu'à l'inférence d'autres informations (cadres de sous-catégorisation, classes sémantiques) concernant des entrées inconnues des lexiques. De tels travaux devront

16. Ces informations pourraient également être construites pour les néologismes trouvés dans les lexiques externes.

17. Même si ce dernier point est plus délicat, une des caractéristiques de la morphologie constructionnelle étant précisément le caractère non complètement prédictible de la sémantique résultante.

© ATALA

par ailleurs être évalués dans une configuration orientée-tâche, pour montrer par exemple leur utilité en analyse morphosyntaxique, syntaxique ou sémantique, ainsi par exemple que pour l'indexation de documents.

Remerciements Ce travail a été financé par le projet ANR EDyLex (ANR-09-CORD-008) et par l'entreprise viavoo.

Références

AHRONIAN, C. et BÉJOINT, H. (2008). Les noms composés anglais et français du domaine d'internet : une radiographie bilingue. *Meta : journal des traducteurs*, 53(3):648–666.

ANASTASSIADIS-SYMÉONIDIS, A. et NIKOLAOU, G. (2011). L'adaptation morphologique des emprunts néologiques : en quoi est-elle précieuse ? *Langages 3*.

BARANES, M. (2012). Vers la correction automatique de textes bruités : Architecture générale et détermination de la langue d'un mot inconnu. *In Actes de Recital 2012*, Grenoble, France.

BERNHARD, D. (2010). Apprentissage non supervisé de familles morphologisques : Comparaison de méthodes et aspects multilingues. *TAL*, 51(2):11–39.

BLANCAFORT SAN JOSÉ, H., RECOURCÉ, G., COUTO, J., SAGOT, B., STERN, R. et TEYSSOU, D. (2010). Traitement des inconnus : une approche systématique de l'incomplétude lexicale. *In Actes de TALN 2010*, Montréal, Canada.

CABRÉ, M., 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.

CARTIER, E. (2011). Néologie et description linguistique pour le tal. *Langages*, 183:105–117.

CHRUPAŁA, G., DINU, G. et van GENABITH, J. (2008). Learning morphology with morfette. *In Proceedings of LREC 2008*, Marrakech, Morocco. ELDA/ELRA.

CREUTZ, M. et LAGUS, K. (2005). Inducing the morphological lexicon of a natural language from unannotated text. *In Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*, pages 106–113.

DAL, G. et NAMER, F. (2000). Génération et analyse automatiques de ressources lexicales construites utilisables en recherche d'informations. *TAL*, 41(2):423–446.

DENIS, P. et SAGOT, B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art pos tagging. *Language Resources and Evaluation*, 46(4):721–736.

DISTER, A. et FAIRON, C. (2004). Extension des ressources lexicales grâce à un corpus dynamique. *Lexicometrica*, Thema 7.

GOLDSMITH, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198.

HAN, B. et BALDWIN, T. (2011). Lexical normalisation of short text messages : Makn sens a# twitter. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, volume 1, pages 368–378.

HATHOUT, N. (2010). Morphonette : a morphological network of french. *CoRR*, abs/1005.3902.

HATHOUT, N. et NAMER, F. (2011). Règles et paradigmes en morphologie informatique lexématique. *In Actes de TALN 2011*, Montpellier, France.

© ATALA

LAVALLÉE, J.-F. et LANGLAIS, P. (2011). Moranapho : un système multilingue d'analyse morphologique fondé sur l'analogie formelle. *TAL*, 52(2):17–44.

LEPAGE, Y. (1998). Solving analogies on words : An algorithm. *In Proceedings of COLING-ACL 1998*, pages 728–735.

LOVINS, J. (1968). *Development of a stemming algorithm*. MIT Information Processing Group, Electronic Systems Laboratory.

MATHIEU-COLAS, M. (2010). *Flexion des noms et des adjectifs composés : principes de codage*. Lexiques, Dictionnaires, Informatique (LDI).

MAUREL, D. (2008). Prolexbase : a multilingual relational lexical database of proper names. *In Proceedings of LREC'08*, pages 334–338, Marrakech, Morocco.

MAUREL, D. et PITON, O. (1998). Un dictionnaire de noms propres pour intex : les noms propres géographiques. *Lingvisticae Investigationes*, 22:279–289.

MIKHEEV, A. (1997). Automatic rule induction for unknown-word guessing. *Computational Linguistics*, 23:405–423.

NEUVEL, S. et FULOP, S. A. (2002). Unsupervised learning of morphology without morphemes. *CoRR*, cs.CL/0205072.

ROMARY, L., SALMON-ALT, S. et FRANCOPOULO, G. (2004). Standards going concrete : from lmf to morphalou. *In Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries*, pages 22–28.

SABLAYROLLES, J. (1997). Néologismes : Une typologie des typologies. *Cahier du CIEL*, 1996-1997:11–48.

SABLAYROLLES, J., JACQUET-PFAU, C. et HUMBLEY, J. (2011). Emprunts, créations 'sous influence' et équivalents. *In Actes des Huitièmes Journées scientifiques du Réseau de chercheurs Lexicologie, terminologie, traduction*.

SAG, I., BALDWIN, T., BOND, F., COPESTAKE, A. et FLICKINGER, D. (2002). Multi-word expressions : A pain in the neck for NLP. *In proceedings of Conferences on Computational Linguistics and Natural Language Processing*.

SAGOT, B. (2010). The L*efff* , a freely available and large-coverage morphological and syntactic lexicon for French. *In Proceedings of LREC 2010*, La Valette, Malte.

SAGOT, B. et BOULLIER, P. (2008). SxPipe 2 : architecture pour le traitement pré-syntaxique de corpus bruts. *TAL*, 49(2):155–188.

SAGOT, B. et STERN, R. (2012). Aleda, a free large-scale entity database for French. *In Proceedings of LREC 2012*, pages 1273–1276, Istanbul, Turquie.

SCHMID, H. (1994). Treetagger. *TC project at the Institute for Computational Linguistics of the University of Stuttgart*.

STROPPA, N. et YVON, F. (2006). Du quatrième de proportion comme principe inductif : une proposition et son application à l'apprentissage de la morphologie. *TAL*, 47(1):33–59.

TANGUY, L. et HATHOUT, N. (2002). Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du Web. *In Actes de TALN 2002*, pages 245–254, Nancy, France.

VILLOING, F. (2003). Les mots composés VN du français : arguments en faveur d'une construction morphologique. *Cahiers de Grammaire*, 28:183–196.

WALTHER, G. et SAGOT, B. (2011). Problèmes d'intégration morphologique d'emprunts d'origine anglaise en français. *In 30th International Conference on Lexis and Grammar*.

 © ATALA