Système de prédiction de néologismes formels : le cas des N suffixés par –IER dénotant des artefacts

Aurélie Merlo[1]
(1) STL UMR 8163, rue du Barreau BP 60149 59653 Villeneuve d'Ascq Cedex
aurelie.merlo@etu.univ-lille3.fr

RESUME

Nous présentons ici un système de prédiction de néologismes formels avec pour exemple la génération automatique de néologismes nominaux suffixés par –IER dénotant des artefacts (*saladier, brassière, thonier*). L'objectif de cet article est double. Il s'agira (i) de mettre en évidence les contraintes de la suffixation par –IER afin de les implémenter dans un système de génération morphologique puis (ii) de montrer qu'il est possible de prédire les néologismes formels. Ce système de prédiction permettrait ainsi de compléter automatiquement les lexiques pour le Traitement Automatique des Langues (TAL).

ABSTRACT

Prediction Device of Formal Neologisms : the Case of –IER Suffixed Nouns Denoting Artifacts

We'll introduce here a device that can predict neologisms using an example the automatical generation of nominal neologisms suffixed by –IER denoting artifacts (*saladier, brassière, thonier*). The aim of this article is double. We will first address the –IER suffixation constraints in order to take them into account on the implementation of our morphological generator. Second, we will describe our method to predict formal neologisms. Such a method will permit to automatically enrich NLP lexicons.

MOTS-CLES : morphologie constructionnelle, néologie, génération morphologique, incomplétude lexicale.
KEYWORDS : constructional morphology, neology, morphological generation, lexical incompleteness.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 57–70,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Certaines applications de Traitement Automatique des Langues (TAL) comme les traducteurs automatiques utilisent des lexiques (Sproat, 1992). Cependant, ces lexiques comme les dictionnaires (Sablayrolles, 2000 ; Sablayrolles, 2008) sont lacunaires dans la mesure où ils ne peuvent pas contenir l'ensemble des unités lexicales d'une langue. L'absence d'une unité lexicale dans un lexique de TAL peut alors poser problème. Ce problème, nommé dans la littérature scientifique l'incomplétude lexicale, a déjà été abordé dans de nombreux travaux. Certains de ces travaux proposent une typologie des mots inconnus (Dister & Fairon, 2004 ; Maurel, 2004 ; Cartoni, 2006 ; Blancafort & *al.*, 2010). D'autres travaux présentent des solutions pour palier cette incomplétude lexicale. Les travaux de (Dister & Fairon, 2004) soumettent la solution d'un repérage des mots inconnus dans un corpus québécois grâce au système GlossaNet. (Cartoni, 2006) propose l'implémentation de règles de construction afin d'analyser les mots construits [1]. (Blancafort & *al.,* 2010) proposent quant à eux un enrichissement dynamique d'un lexique TAL. Cela passe d'abord par une annotation en corpus des tokens inconnus pour une classification automatique. Puis, une validation manuelle est alors nécessaire pour décider de l'ajout temporaire ou définitif dans le lexique.

Dans le cadre de cet article, nous proposons une approche pour palier en partie l'incomplétude lexicale des lexiques de TAL. Dans la mesure où les néologismes formels ont une large place dans les mots inconnus (Maurel, 2004 ; Cartoni, 2006), nous proposons de les prédire afin de les intégrer dans les lexiques de TAL. Notre hypothèse est que les nouvelles unités lexicales qui satisfont les contraintes linguistiques liées à un procédé constructionnel sont prédictibles. Mais seront-elles pour autant attestées dans l'usage ? Afin de vérifier notre hypothèse, nous avons élaboré un générateur automatique de néologismes formels à base de contraintes linguistiques. Nous avons choisi de générer automatiquement des néologismes formels nominaux suffixés par –IER[2] dénotant des artefacts (*saladier, brassière, thonier*). Le choix de cette suffixation s'explique par sa diversité référentielle (Corbin & Corbin, 1991).

Les objectifs de cet article sont alors (i) de mettre en évidence les contraintes linguistiques de la suffixation par –IER et (ii) de montrer qu'il est possible de prédire les néologismes formels.

Nous commencerons par un état de l'art des générateurs morphologiques existants afin de mettre en évidence l'originalité de notre système de génération automatique de néologismes formels. Puis, nous procèderons à l'analyse de la suffixation par –IER afin de mettre en évidence ses contraintes linguistiques qui seront par la suite implémentées au sein de notre système de génération morphologique. Nous commenterons les résultats obtenus afin de montrer les avantages et les limites de notre approche. Enfin, nous terminerons cet article sur les perspectives de recherche en morphologie constructionnelle que permet cette première approche de la prédiction des néologismes formels.

[1] Nous pensons ici à DériF (Namer, 2009).
[2] Nous avons adopté cette convention d'écriture afin de rassembler sous –IER les suffixes allomorphiques –*ier(e)* et –*er(e)*.

2 État de l'art des générateurs morphologiques

Il existe actuellement des systèmes de génération morphologique permettant de générer automatiquement des formes fléchies et/ou des formes dérivées.

Le premier de ces systèmes que nous présentons brièvement est le système PILAF (Procédures Interactives Linguistiques Appliquées au Français) (Courtin & *al.*, 1994). Le système PILAF a été élaboré dans le laboratoire CLIPS afin de réaliser des tâches d'analyse et de génération morphologique. Le système est composé d'une grammaire, d'un dictionnaire et de deux moteurs : l'un est consacré à l'analyse, l'autre à la génération. Ce système est capable de reconnaître et de générer 250 000 formes du français à l'aide d'un lexique comportant 35 000 entrées. Ce système, destiné à la morphologie flexionnelle, utilise un formalisme qui « permet le codage de règles de la morphologie dérivationnelle » (Courtin & *al.,* 1994 : 101).

Le « système dérivationnel » de (Tzoukermann & Jacquemin, 1997) est élaboré également pour fonctionner aussi bien en analyse qu'en génération. En génération, le système repose sur le principe de la concaténation et pour éviter la surgénération, (Tzoukermann & Jacquemin, 1997) proposent trois niveaux de filtrage : un filtrage lexical (les dérivés sont générés à partir de bases attestées dans un dictionnaire), un filtrage d'attestation en corpus et un filtrage sémantique collocatif (les dérivés sélectionnés doivent apparaître en contexte en collocation avec un même mot).

Le système Intex (Silverztein, 1993), dont la dernière version se nomme NooJ, est un environnement dans lequel il est possible d'élaborer des descriptions formalisées applicables sur corpus. Cet environnement comprend des dictionnaires électroniques (DELAF) et offre la possibilité de réaliser ses propres grammaires sous forme de graphes. Le système Intex est utilisé en traitement automatique de corpus mais également pour de la génération flexionnelle et/ou dérivationnelle.

Le système GéDériF est le pendant de DériF (Namer, 2009). C'est un système de génération automatique d'unités lexicales construites qui a une triple fonction : (i) « produire un lexique d'unités lexicales construites absentes des dictionnaires », (ii) « enrichir ce lexique d'informations constructionnelles et sémantiques, (iii) « constituer des micro-familles constructionnelles » (Dal & Namer, 2000). Les formes générées sont validées par une recherche en ligne. L'inconvénient est que « le générateur ne peut donc choisir la base des unités qu'il va générer automatiquement que parmi les mots déjà construits » (Dal & Namer, 2000) et analysés par DériF.

La présentation de ces systèmes de génération morphologique permet de mettre en évidence une limite commune. Ces systèmes offrent la possibilité de générer des mots dérivés et implémentent un certain nombre de règles dérivationnelles. Or, ces systèmes ne prennent pas en compte les contraintes sémantiques pouvant peser dans un processus dérivationnel.

À présent, dans le cadre de l'analyse de la suffixation par –IER, nous allons montrer en quoi notre système de génération morphologique est original.

3 Analyse de la suffixation par -IER

3.1 Méthodologie

Nous allons mettre en évidence les contraintes linguistiques pesant sur la suffixation par –IER. Nous avons choisi cette suffixation pour sa diversité référentielle (Corbin & Corbin, 1991). Dans le cadre de cet article, nous étudierons de plus près les dérivés nominaux suffixés par –IER dénotant des artefacts (*saladier, brassière, thonier*). Nous précisons que les résultats présentés sont extraits d'une étude plus large de la suffixation par –IER que nous avons réalisée dans le cadre d'un mémoire de recherche (Merlo, 2011).

Le corpus élaboré dans le cadre de ce mémoire provient de l'extraction du *TLF* d'une liste de candidats à la suffixation par –IER. Cette liste a été nettoyée afin de ne retenir que les lexèmes construits par la suffixation par –IER (élimination des emprunts par exemple du type *manager*). Nous avons par la suite sélectionné 530 lexèmes construits que nous avons analysés afin de faire apparaître les informations nécessaires à la prédiction de néologismes suffixés par –IER. Par conséquent, nous avons procédé à une analyse morpho-sémantique, morphophonologique et graphique de la suffixation par –IER.

Pour cet article ne retenant que les dérivés dénotant des artefacts, l'étude portera sur 119 lexèmes construits et 132 acceptions[3].

3.2 Analyse morpho-sémantique

3. 2. 1. Approche référentielle

L'analyse morpho-sémantique des dérivés suffixés par –IER a consisté en l'étude de la référence des dérivés et de leur base. Nous nous sommes inspiré de l'approche par classe référentielle proposée dans (Roché, 1998)[4]. Par convention, les classes référentielles seront entre crochets ([récipient] par exemple).

À partir des définitions du *TLF*, nous avons déterminé la classe référentielle des acceptions des dérivés de notre corpus d'étude. Notre analyse morpho-sémantique a pu confirmer l'existence de sept classes référentielles pour les dérivés suffixés par –IER dénotant des artefacts (*cf.* FIGURE 1 ci-dessous) que (Roché, 1998) avait mis en évidence.

[3] Nous avons choisi de prendre en compte les acceptions lorsque celles-ci n'étaient pas issues d'une dérivation sémantique. Ex : *médailler* (*s. v. médailler* dans le *TLF*) désigne à la fois un meuble contenant des médailles et un recueil de médailles.

[4] Une classe référentielle est définie comme la projection de propriétés sémantiques (Temple, 1996).

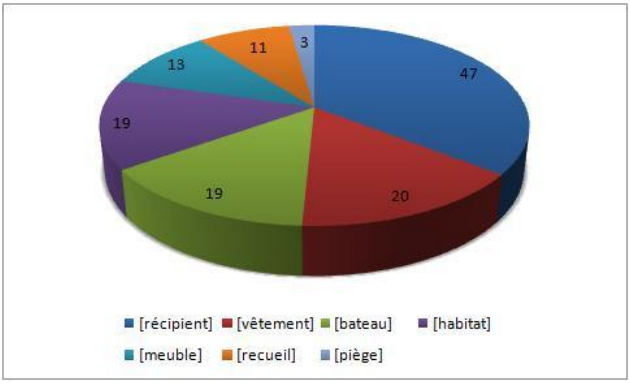

FIGURE 1 – Classe référentielle des dérivés suffixés par –IER dénotant des artefacts (contrainte n°1)

Du côté de la sortie, nous savons à présent que les dérivés suffixés par –IER dénotant des artefacts peuvent avoir pour référent un nom de récipient, un nom de vêtement, un nom de bateau, un nom d'habitat ou un nom de meuble. Néanmoins, il nous reste à déterminer si la suffixation par –IER sélectionne les bases en fonction d'une contrainte référentielle. Toujours à partir des définitions du *TLF*, nous avons déterminé la classe référentielle des bases de notre corpus (*cf.* TABLE 1 ci-dessous).

Classe référentielle du dérivé	Classe référentielle de la base	Nombre d'acceptions	Exemples
[bateau]	[poisson]	7	*baleinier*
	[partie du bateau]	5	*boulinier*
	[minéral]	3	*méthanier*
	[personne]	1	*négrier*
	[boisson]	1	*pinardier*
	[fruit]	1	*bananier*
	[rangement]	1	*vraquier*
[habitat]	[animal]	15	*pigeonnier*
	[personne]	2	*garçonnière*

	[minéral]	1	*terrier*
	[procès]	1	*volière*
[meuble]	[objet]	9	*bonnetière*
	[végétal]	3	*grainier*
	[durée]	1	*semainier*
[piège]	[animal]	3	*ratière*
[récipient]	[aliment]	22	*beurrier*
	[objet]	11	*bouquetier*
	[liquide]	8	*tisannière*
	[animal]	3	*turbotière*
	[minéral]	2	*sablier*
	[mode de cuisson]	1	*daubière*
[recueil]	[écrit]	7	*chansonnier*
	[objet]	2	*médaillier*
	[végétal]	2	*herbier*
[vêtement]	[partie du corps]	17	*jambière*
	[animal]	1	*grenouillère*
	[meuble]	1	*tablier*

TABLE 1 – Classe référentielle des bases (contrainte n°2)

Grâce à cette analyse de la classe référentielle des bases, à présent nous savons que pour générer par exemple un nom de recueil, la suffixation par –IER peut sélectionner des bases dont le référent désigne un écrit, un objet ou un végétal. Cette table permet également de mettre en évidence des préférences sémantiques telle que pour former un nom de meuble, la suffixation par –IER sélectionne de préférence une base dont le référent désigne un aliment (22 acceptions). Dans le cadre de la génération automatique de néologismes formels suffixés par –IER, nous ne tiendrons pas compte de ces préférences.

3.2.1 Variation en genre

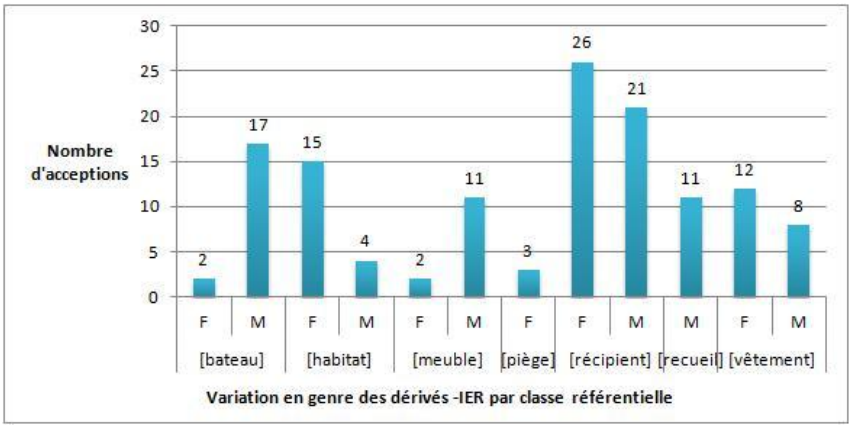

FIGURE 2 – Variation en genre des dérivés suffixés par –IER dénotant des artefacts (contrainte n° 3)

Prédire les néologismes suffixés par –IER dénotant des artefacts nécessite également de déterminer les cas de variation en genre. Selon (Roché, 1998 : 45), l'attribution du genre pour les artefacts s'applique par le biais d'« un déterminé implicite ou d'un terme générique » (*un (bateau) pétrolier*) ou par défaut au masculin lorsqu'il n'y a pas de déterminé implicite. Sur la FIGURE 2 ci-dessous, nous avons répertorié le nombre de formes féminines et masculines par classe référentielle des dérivés.

La FIGURE 2 nous permet de faire des choix quant à la flexion en genre des néologismes suffixés par –IER. En effet, nous voyons nettement se démarquer les formes masculines pour les classes référentielles [bateau], [meuble] et [recueil] et les formes féminines pour les classes référentielles [habitat], [piège], [récipient] et [vêtement]. Cependant, en l'absence de tests formels pour l'étiquetage en classe référentielle, nous admettons que l'attribution du genre ici est sujette à caution.

À présent, il nous reste à déterminer les contraintes formelles de la suffixation par –IER.

3.3 Analyse morphophonologique

Deux ensembles de contraintes pèsent plus particulièrement en morphophonologie : les contraintes de taille (Plénat, 1997) et les contraintes dissimilatives[5].

3.3.1 Contrainte de taille

Il s'agit de découvrir ici si la suffixation par –IER comporte des contraintes de taille sur l'entrée et/ou la sortie de ses règles de construction de lexème. Pour cela, nous avons calculé le nombre de syllabes de chaque base et de chaque dérivé de notre corpus (*cf.*

[5] Les contraintes de dissimilation rejettent la contiguïté de deux phonèmes identiques ou similaires.

TABLE 3 ci-dessous).

Taille de la base	Taille du dérivé	Exemple	Nombre de lexèmes
1	2	*beurrier*	41
1	3	*oeufrier*	16
2	2	*jarretière*	4
2	3	*grenouillère*	61
2	4	*vinaigrier*	1
3	4	*porcelainier*	9

TABLE 3 – Contrainte de taille pour les dérivés suffixés par –IER dénotant des artefacts (contrainte n° 4)

Ces résultats permettent de mettre en évidence que la suffixation par –IER ne sélectionne pas de base comportant plus de trois syllabes pour former des dérivés quadrisyllabiques. Par ailleurs, ces résultats montrent que la suffixation par –IER sélectionne de préférence des bases dissyllabiques pour former des dérivés trissyllabiques (61 lexèmes concernés). Enfin, à travers ces résultats, nous voyons que la suffixation par –IER utilise de préférence la concaténation avec un nombre important de lexèmes construits comportant une syllabe de plus par rapport à leur base.

Par conséquent, notre système de prédiction de néologismes suffixés par –IER dénotant des artefacts devra effectuer un tri formel (en plus d'un tri sur les classes référentielles) parmi les bases afin de ne sélectionner que les bases entre une et trois syllabes.

3.3.2 Contrainte sur l'attaque

Ce que nous appelons une attaque est le phonème placé au début de la syllabe contenant le suffixe –IER. Ce phonème provient en général de la finale consonantique de la base mais certaines attaques sont parfois des consonnes épenthétiques (comme dans *cafetière*), intercalées entre la base et le suffixe pour éviter un hiatus car « d'une façon générale, la dérivation par *–ier(e)* est gênée par une finale vocalique (sans consonne latente) » (Roché, 1998).

Nous avons relevé les attaques des 119 lexèmes construits de notre corpus (*cf.* FIGURE 3 ci-dessous).

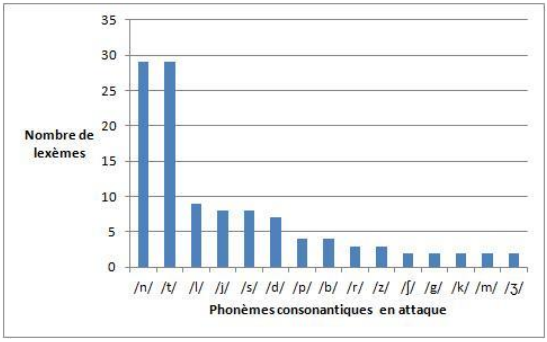

FIGURE 3 – Distribution des phonèmes consonantiques en attaque

Ce graphique montre que les dérivés suffixés par –IER comportant une attaque /n/ ou /t/ sont les plus nombreux. D'une manière générale, la suffixation par –IER tend à sélectionner des bases dont la finale consonantique est une alvéolaire (/t/ et /n/) et au contraire à éviter des bases dont la finale consonantique est un phonème labiodental (/v/ et /f/), bilabial (/p/, /b/ et /m/) ou vélaire (/k/ et /g/). Néanmoins, on ne peut pas parler de contraintes sur l'attaque mais plutôt de préférences ce qui n'est pas le cas lorsque la base se termine par une finale vocalique avec l'apparition systématique d'une consonne épenthétique.

3.4 Analyse graphique

3.4.1 Cas de changements graphiques

L'adjonction du suffixe –IER provoque des modifications graphiques. La première de ces modifications est l'apparition d'une consonne épenthétique entre la finale vocalique de la base et le suffixe –IER (contrainte n°5). Dans le cadre de notre mémoire, nous avons recensé 19 dérivés suffixés par -IER comportant une consonne épenthétique. Parmi eux des noms d'artefacts comme *cafetière* [récipient], *fourmilière* [habitat], *morutier* [bateau] ou encore *tabatière* [récipient]. Nous avons également découvert que le phonème /t/ est particulièrement utilisé comme consonne épenthétique.

Le second cas de modification graphique par adjonction du suffixe –IER est le changement d'accent qui traduit un changement d'aperture pour cause de contrainte dissimilative (contrainte n°6). Ainsi, parmi les noms d'artefacts, les dérivés *chéquier* [recueil] et *négrier* [bateau] ont subi un changement d'accent par rapport à leur base respective *chèque* et *nègre*.

Enfin, la gémination est le dernier cas de modification graphique relevé. La consonne finale de la base à tendance à se doubler avec la concaténation du suffixe –IER. Nous allons détailler ci-dessous les différents cas de figure de gémination.

3.4.2 Règles graphiques de concaténation du suffixe –IER

À partir de l'analyse de notre corpus, nous avons posé huit contraintes graphiques

régissant la concaténation du suffixe –IER :

- Contrainte n°1 : lorsque la base se termine par une consonne muette (*abricot*), il y a concaténation simple (*abricotier*).
- Contrainte n°2 : lorsque la base se termine par un phonème consonantique et par une voyelle muette (*agence*), la voyelle finale est supprimée et le suffixe –IER se concatène au radical de la base (*agencier*).
- Contrainte n°3 : lorsque la base se termine par un phonème vocalique et par une voyelle (*moru*), la consonne épenthétique *–t-* s'ajoute suivie de la concaténation du suffixe –IER (*morutier*).
- Contrainte n°4 : lorsque la base se termine par *–os, -as, -is* ou *–us*, (*matelas*) la consonne *–s-* s'ajoute suivie de la concaténation du suffixe –IER (*matelassier*).
- Contrainte n°5 : lorsque la base se termine par *–on*, la consonne *–n-* s'ajoute suivie de la concaténation du suffixe –IER (*boutonnière*).
- Contrainte n°6 : lorsque la base se termine par *–an* ou *–in*, il y a concaténation simple du suffixe –IER (*rubanier*, *jardinière*).
- Contrainte n°7 : lorsque la base se termine par *–il*, la consonne *–l-* s'ajoute suivie de la concaténation du suffixe –IER (*œillère*).
- Contrainte n°8 : lorsque la base contient dans sa dernière syllabe un accent grave (*trèfle*), cet accent devient un accent aigu dans le dérivé suffixé par –IER (*tréflier*).

En ce qui concerne les règles n°4 et n°5, elles ont été élaborées sur la base de la fréquence des phénomènes étudiés car il existe des contre-exemples à ces règles : *tamis>tamisier, thon>thonier*.

4 Implémentation

Nous avons dégagé les contraintes de la suffixation par –IER déterminant ainsi les connaissances linguistiques nécessaires à la génération automatique de néologismes suffixés par –IER. Cependant, la génération automatique de néologismes formels nécessite également un lexique de référence (ou lexique d'exclusion) et un lexique de bases annoté.

Bien que l'utilisation d'un dictionnaire présente des limites lorsque l'on étudie la néologie (Sablayrolles, 2000 ; Sablayrolles, 2008), nous avons choisi comme lexique de référence le lexique *Morphalou* élaboré à partir de la nomenclature du *TLF*. L'objectif de ce lexique de référence est de retenir uniquement les formes nouvelles générées. Si une forme générée est attestée dans le lexique de référence, celle-ci ne sera pas retenue.

L'élaboration du lexique de bases annoté a été plus complexe. Ce lexique devait comprendre les informations linguistiques nécessaires à l'application des contraintes de la suffixation par –IER. Ainsi, ce lexique devait contenir des indications sur la classe référentielle, le nombre de syllabe et une description graphique de la base afin de déterminer quelle règle graphique à appliquer. Au vu de ces informations, la ressource *Lexique* 3. 1. était une ressource utile en ce qui concerne le nombre de syllabes et la description orthographique (*cf.* TABLE 4 ci-dessous). La description orthographique permet de savoir si la finale de la base est vocalique (*café*) ou consonantique (*cafard*) ce qui a son importance pour l'application des contraintes graphiques n°1 et n°3. La

description phonétique permet de savoir si la base se termine par une finale consonantique et une voyelle (*agence*) ce qui a de l'importance pour l'application de la contrainte n°2. Enfin, la description orthographique des syllabes permet d'appliquer la contrainte n°8 du changement d'accent.

Lemme	Description orthographique	Description phonétique	Nombre de syllabes	Description orthographique des syllabes
café	CVCV	CVCV	2	ca-fé
cafard	CVCVCC	CVCVC	2	ca-fard
agence	VCVCCV	VCVC	2	a-gen-ce

TABLE 4 – Informations contenues dans *Lexique* 3. 1. et utiles au lexique des bases

Nous avons choisi aléatoirement 268 lexèmes de *Lexique* 3. 1. pour constituer notre lexique de bases avec pour seul objectif d'obtenir une homogénéité de classe référentielle. Puis, nous avons procédé à l'annotation des classes référentielles de ces lexèmes-bases. Aucune ressource actuellement en français ne possédant ce type d'information, nous nous sommes appuyés sur les définitions du *TLF* pour établir les classes référentielles.

5 Résultats

Notre hypothèse en introduction était que les nouvelles unités lexicales qui satisfont les contraintes linguistiques liées à un procédé constructionnel sont prédictibles. Ici, il s'agissait de mettre en évidence les contraintes de la suffixation par –IER en s'attachant particulièrement à l'étude des noms d'artefacts.

Notre approche a permis de générer 544 nouvelles unités lexicales dont 222 dénotant des artefacts. Afin de vérifier notre hypothèse, nous avons interrogé le Web et déterminé quelles nouvelles unités lexicales étaient attestées. La FIGURE 4 ci-dessous dresse un bilan de cette vérification quantitative pour les néologismes dénotant des artefacts.

Ces résultats, dont notamment le nombre de néologismes formels générés attestés sur le Web (87 au total soit 39,19%), valident en partie notre hypothèse. Il apparaît possible de prédire les nouvelles unités lexicales qui satisfont les contraintes linguistiques liées à un procédé constructionnel. Il apparaît même possible de prédire la référence de ces nouvelles unités lexicales (18 au total parmi les attestés). Néanmoins, il apparaît également que notre système est en surgénération (135 non-attestés au total soit 60,81%).

Autant une attestation relevée sur le Web est sujette à caution (Kilgarriff & Grefenstette, 2003), autant l'absence d'attestation sur le Web est significative.

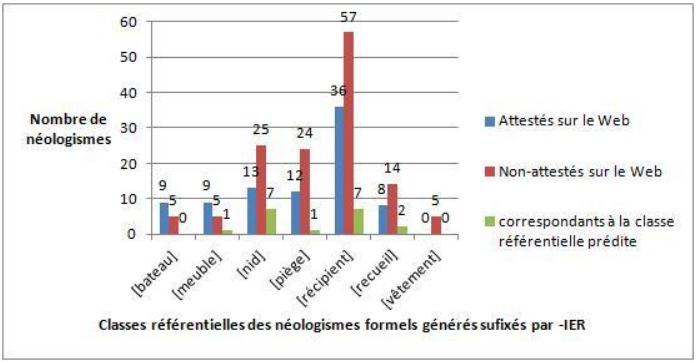

FIGURE 4 – Évaluation quantitative des résultats

Nous sommes face à trois types de résultats. Le premier concerne tous les cas de néologismes formels générés suffixés par –IER ne trouvant aucune attestation sur le Web. Cela provient-il d'une erreur de génération graphique ? Le néologisme généré *antilopière* par exemple n'est pas attesté sur le Web alors qu'il ne présente pas *a priori* d'erreur graphique. Cela provient-il alors d'une erreur de catégorisation de la classe référentielle de la base ? Cette hypothèse est plus probable dans la mesure où nous avons procédé sans tests formels mais à l'aide des définitions du *TLF*. Le second type de résultat conforte d'ailleurs cette hypothèse puisqu'il s'agit des néologismes formels générés suffixés par –IER attestés sur le Web mais pas dans la classe référentielle prédite. C'est ainsi le cas de *cochonnière* prédit dans la classe référentielle [piège] et attesté dans un contexte faisant référence à un véhicule[6]. Enfin, le dernier type de résultat concerne les attestations de néologismes générés suffixés par –IER dans les classes référentielles prédites tels que *choucroutière* prédit dans la classe [récipient]. D'autres hypothèses peuvent être avancées pour expliquer cette surgénération. Nous n'avons pas abordé la question de l'échangisme suffixal (Roché, 1997). La validation de notre hypothèse par le biais de la vérification des formes générées sur le Web est également discutable. Une autre alternative pourrait être une validation auprès d'un panel de locuteurs.

6 Conclusion : sur les pas des connaissances encyclopédiques

En conclusion, nous avons atteint nos deux objectifs qui étaient (i) de mettre en évidence les contraintes liées à la suffixation par –IER et (ii) de montrer qu'il est possible de prédire les néologismes formels par l'implémentation de ces contraintes au sein d'un système de génération morphologique.

Du point de vue applicatif, cette approche apporte des perspectives quant à l'incomplétude lexicale. Sur une thématique précise qui est l'étude des noms suffixés par –IER dénotant des artefacts, nous avons prédit 87 néologismes formels au total pouvant

[6] « René Thoré avec Kapi (mulassier) attelé à une cochonnière à 4 roues » (http://picasaweb.google.com/lh/photo/Pp3EyPQcQH1DaduXJK_9Mw)

directement être intégrés dans un lexique pour le TAL. Cette quantité est très insuffisante mais nous espérons améliorer le taux de prédiction prochainement. Du point de vue théorique, cette approche apporte de nouvelles perspectives de recherche en morphologie constructionnelle. Nous avons prédit la référence de 18 néologismes formels sur 87 attestés sur le Web grâce à une annotation en classe référentielle. Cela implique que notre approche de la prédiction par la notion de classe référentielle mériterait d'être approfondie. Tout d'abord, afin d'éviter l'utilisation du *TLF* pour la catégorisation référentielle des lexèmes, une première solution serait d'élaborer une série de tests formels en s'inspirant de travaux en sémantique-cognitive (notamment Fillmore, 1982 ; Langacker, 1987 ; Kleiber, 1990). Une seconde solution consisterait à utiliser des ontologies existantes telles que WOLF pour le français (Sagot & Fišer, 2008) ou WordNet (Fellbaum, 1998). Puis, à partir de là, il faudrait, par classe référentielle, faire émerger les contraintes extralinguistiques. Dans un travail de thèse que nous venons de débuter, nous tentons de démontrer l'hypothèse de la nécessité des connaissances encyclopédiques en morphologie constructionnelle (Aronoff, 1980 ; Clark & Clark, 1979). L'exemple du néologisme généré *cochonnière* met en lumière cette nouvelle hypothèse car si ce néologisme apparaît dans un contexte désignant un véhicule et non pas un piège c'est parce le cochon est un animal « domestique » que nous n'avons pas besoin de chasser. De la même manière, on relève *bananier* dont le référent désigne un nom de bateau formé sur une base dont le référent désigne un fruit mais pas *mirabellier* car la mirabelle ne se commercialise pas dans le monde entier et n'a donc pas besoin d'un transport en bateau. Les connaissances encyclopédiques permettraient ainsi de faire le tri entre ce qui est ce qui a une pertinence dénominative et ce qui n'en a pas.

Références

ARONOFF, M. (1980). Contextuals, *in Language*, 56, No. 4, pp. 744-758.

BLANCAFORT, S. J. H., RECOURCE, 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.

CARTONI, B. (2006). Constance et variabilité de l'incomplétude lexicale. *Noûs* (3), pp. 10–13.

CLARK, E. V. et CLARK, H. (1979). When Nouns Surface as Verbs, *in Language*, 55, No. 4, pp. 767-811.

CORBIN, D. et CORBIN, P. (1991). Un traitement unifié du suffixe -ier(e), *in Lexique* 10, pp. 61-145.

COURTIN, J., DUJARDIN, D., GENTHIAL, D. et KOWARSKI, I. (1994). Analyse et génération morphologique avec le système PILAF, *in TAL* « Morphologie computationnelle », vol. 35, n° 2.

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

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

FELLBAUM, C. (1998). *WordNet: An Electronic Lexical Database*, Cambridge : MIT Press.

FILLMORE, C. J. (1982). Frame semantics in T. L. S. of Korea (Ed.), Linguistics in the Morning Calm. Seoul: Hanshin Publishing Co.

KILGARRIFF, A et GREFENSTETTE, G. (2003). Introduction to the special issue on the Web as a corpus, *in Computational Linguistics, 29*(3), 333347.

KLEIBER, G. (1990). *La sémantique du prototype. Catégorie et sens lexical*, Paris : PUF.

LANGACKER, R.W. (1987). *Foundations of Cognitive Grammar*. Vol. I: Theoretical Prerequisites, Stanford: Stanford University Press.

MAUREL, D. (2004). Les mots inconnus sont-ils des noms propres ? *in Actes des JADT 2004.*

MERLO, A. (2011). *Élaboration d'un prototype de générateur automatique de néologismes formels : le cas des suffixés par –IER*. Mémoire de recherche en vue de l'obtention du Master professionnel « Lexicographie, Terminographie et Traitement Automatique de Corpus ». Université Charles de Gaulle, Lille 3.

NAMER, F. (2009). *Morphologie, lexique et Traitement Automatique des Langues – Le système DériF : TIC et Sciences cognitives*, London : Hermès Sciences Publishing.

PLENAT, M. (1997). Analyse morpho-phonologique d'un corpus d'adjectifs en –esque, *in Journal of French Language Studies*, 7 : 163–179.

ROCHE, M. (1997). Briard, bougeoir et camionneur : dérivés aberrants, dérivés possibles, *in* Corbin et al., éd. (1997), pp. 241–250.

ROCHE, M. (1998). *Deux études sur la dérivation en -ier(e)*, Toulouse, Carnets de grammaire (Rapports internes de l'ERSS, CNRS et Université de Toulouse-Le Mirail).

SAGOT, B. et FISER, D. (2008). Construction d'un wordnet libre du français à partir de ressources multilingues, in TALN 2008, Avignon, France.

TZOUKERMANN, E. et JACQUEMIN, C. (1997). Analyse automatique de la morphologie dérivationnelle et filtrage de mots possibles, *in Silexicales* « Mots possibles et mots existants », n° 1, pp. 251 – 260.

SABLAYROLLES, J.-F. (2000). *La néologie en français contemporain. Examen du concept et analyse de productions néologiques récentes*, Paris : Honoré Champion.

SABLAYROLLES, J.-F. (2008). Néologie et dictionnaire(s) comme corpus d'exclusion, *in Néologie et terminologie dans les dictionnaires*, douzième Journée des dictionnaires, Université de Cergy-Pontoise, 17 mars 2004, sous la direction de Sablayrolles Jean-François, Paris : H. Champion.

SILBERZTEIN, M. (1993). Dictionnaires électroniques et analyse de texte : le système INTEX, Masson : Paris.

SPROAT, R. (1992). *Morphology and Computation*, Cambridge : MIT Press.

TEMPLE, M. (1996). *Pour une sémantique des mots construits,* Villeneuve d'Asq : Presses Universitaires du Septentrion.