Trotter, David/Bozzi, Andrea/Fairon, Cédrick (éd.) (2016). *Actes du XXVII^e Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013). Section 16 : Projets en cours ; ressources et outils nouveaux*. Nancy, ATILF :
http://www.atilf.fr/cilpr2013/actes/section-16.html.

Détection, extraction automatique et analyse des lexèmes construits par la préfixation en *non-* en français

1. Introduction

Ce travail vise à mettre à contribution linguistique théorique et Traitement Automatique des Langues (TAL), dans le cadre d'une thèse consacrée à l'étude du préfixe[1] de négation *non-*. En français contemporain, le préfixe *non-* peut s'adjoindre à des bases nominales (p.ex. NON-CONFORMITÉ, NON-RESPECT) et adjectivales (p.ex. NON-FERREUX, NON-PRODUCTIF), ce qu'illustrent (1) et (2) :

(1) Rien n'est plus facile que de constater la conformité de l'écriture d'un texte, ou sa non-conformité, avec
l'orthographe légale. (*TLFi, s.v.* 'non-conformité')
(2) Le luxe consiste essentiellement dans les dépenses non-productives, quelle que soit d'ailleurs la nature de ces
dépenses. (*TLFi, s.v.* 'non(-)')

J'expose dans cet article la méthodologie employée pour détecter, extraire automatiquement et analyser les lexèmes construits sur base nominale et adjectivale à l'aide du préfixe *non-* en français dans un corpus ouvert, à savoir les pages en français indexées par le moteur de recherche *Google*™. Un de mes objectifs est de comparer ces données avec les lexèmes préfixés par *non-* attestés dans le *Trésor de la Langue Française informatisé* (*TLFi*)[2], l'hypothèse sous-jacente étant que les dictionnaires ne reflètent que très partiellement l'état actuel du français (cf. notamment Dal & Namer 2012). En effet, dans le cas qui m'intéresse ici, il semble que les dictionnaires ne recensent qu'une infime partie des lexèmes préfixés par *non-* effectivement utilisés par les locuteurs et la préfixation par *non-* ne paraît soumise qu'à peu de contraintes (concernant par exemple les bases nominales, voir Dugas 2012). D'autre part, ce travail me permet de disposer d'un lexique aussi exhaustif que possible des lexèmes formés par la préfixation en *non-*, dans la perspective de l'examen des caractéristiques de cette préfixation et des lexèmes qu'elle permet de produire.

L'article est organisé en trois temps. Dans un premier temps, je décris la façon dont a été constitué mon corpus et comment s'est faite la détection des formes can-

[1] Par commodité, je parlerai dans cet article du « préfixe » non-, même si son statut préfixal est sujet à questionnement lorsqu'il s'adjoint à des adjectifs (Dugas 2013).

[2] ‹http://atilf.atilf.fr/›

didates. La seconde partie propose une analyse des fréquences obtenues. En conclusion, je rappelle les principaux points de ce travail et j'esquisse quelques perspectives de recherche.

2. Constitution du corpus et détection des formes

2.1. Constitution du lexique de référence

Mon lexique de référence est *Morphalou*[3], lexique ouvert des formes fléchies du français (Romary *et al.* 2006), construit à partir de la nomenclature du *TLFi*. J'ai extrait de ce lexique les formes nominales et adjectivales, soit 82 067 formes : 59 334 formes attestées comme noms et 22 733 formes attestées comme adjectifs. Il faut noter ici que parmi ces 82 067 formes, 6197 possèdent une double catégorisation et sont attestées la fois comme noms et comme adjectifs. Enfin, le lexique *Morphalou* contient 190 lexèmes déjà préfixés par *non-* : 147 noms (p.ex. NON-ACCEPTATION, NON-CONFORMITÉ, NON-INGÉRENCE), 30 adjectifs (p.ex. NON-PRÉDICATIF, NON-REMBOURSABLE, NON-VÉNÉNEUX), et 13 lexèmes avec la double catégorisation nom et adjectif (p.ex. NON-COMMUNISTE, NON-JUIF, NON-LIBÉRAL).

2.2. Génération des lexèmes potentiels préfixés par non- à partir du lexique de référence

L'exploitation du lexique *Morphalou* m'a permis de générer une liste de formes candidates au moyen du composant de prédiction morphologique « non ». Le préfixe *non-* a l'avantage de ne pas posséder d'allomorphes en français contemporain, contrairement à d'autres préfixes : c'est le cas par exemple de *in-,* très productif, qui a pour allomorphes *il-, im-, ir* : LÉGAL > ILLÉGAL, MANGEABLE > IMMANGEABLE, RATIONNEL > IRRATIONNEL ; c'est aussi le cas du préfixe *a-*, qui possède l'allomorphe *an-* devant voyelle : ORGANIQUE > ANORGANIQUE, ENCÉPHALE > ANENCÉPHALE. Ainsi, le préfixe *non-* a une forme stable, quelle que soit le lexème auquel il s'adjoint. De plus, il n'entraîne pas de modification morphologique de sa base. La génération des formes candidates est donc facilitée, puisque la chaîne de caractères à ajouter en début de mot est invariable, et qu'il n'est pas besoin d'adapter la chaîne de caractères composant la forme de base.

Toutefois, la question s'est posée de l'orthographe à choisir pour la génération des lexèmes potentiels : *non-* est-il soudé graphiquement à sa base, ou doit-il en être séparé par un trait d'union, ou par une espace typographique ? La plupart des préfixes du français sont graphiquement soudés à leur base (*a-* : AMORAL ; *in-* : IMPOSSIBLE ; *pré-* : PRÉCOLOMBIEN, etc.) le préfixe *non-* est particulier puisque dans certaines attestations il est séparé de sa base par un trait d'union ou par une espace. Le *TLFi* n'a pas une position claire sur la présence du trait d'union dans les formes en *non-* : certes, « le trait d'union est quasi systématique pour les substantifs; dans le corpus

[3] ‹http://www.cnrtl.fr/lexiques/morphalou/›

littéraire du *TLF*, proportion de l'ordre de 1 ou 2 % d'exceptions » ; toutefois, pour les adjectifs « la règle est l'absence du trait d'union, mais il y a des hésitations. Souvent aussi, notamment dans le discours philosophique, le trait d'union marque une liaison conceptuelle » (*TLFi, s.v.* 'non(-)').

Dans la pratique, si l'on observe de plus près les formes attestées, on rencontre trois cas de figure : (i) présence d'un trait d'union ; (ii) *non-* détaché de la séquence droite, mais sans trait d'union ; (iii) *non-* soudé graphiquement à la séquence droite. On remarquera que le trait d'union est plus systématique devant les bases nominales (p.ex. NON-LINGUISTIQUE, NON-REMBOURSEMENT) que devant les bases adjectivales (p.ex. NON-PRÉDICATIF, NON-SYNDICAL). Au sein de la classe adjectivale, on trouve plus souvent un trait d'union devant les adjectifs typiquement adjectivaux, c'est-à-dire les adjectifs simples et les adjectifs construits par suffixation (p.ex. NON-COMMERCIAL, NON-FERREUX), que devant les formes apparentées à des participes passés (p.ex. NON-ADAPTÉ, NON-DÉFINI). Il est possible que les locuteurs perçoivent les noms précédés de *non-* comme des unités lexicales, autrement dit comme des mots à part entière, où *non-* est un véritable préfixe. L'orthographe plus variable (présence moins systématique du trait d'union) des adjectifs précédés de *non-* témoignerait alors du fait que le statut de ces formes est peu clair pour le locuteur, *a fortiori* lorsque l'adjectif est apparenté à un participe passé, où la frontière entre catégorie adjectivale et verbale et entre lexique et syntaxe paraît brouillée.

D'autre part, les fluctuations orthographiques s'observent non seulement entre différentes formes, mais aussi pour une même forme : ainsi on pourra trouver à la fois dans les corpus des attestations de *non-progrès, non progrès* et *nonprogrès* (Dugas 2012). Toutefois, l'examen d'un échantillon de données a révélé que la fréquence des formes où *non-* est graphiquement soudé à sa base était négligeable (de l'ordre de quelques centaines d'occurrences) et par conséquent, il a été décidé que seules les formes où *non-* est lié à sa base par un trait d'union et celle où il est séparé de sa base par une espace typographique devaient faire l'objet d'une requête. Étant donné qu'une requête comportant un trait d'union (par exemple « non-linguiste ») permet aussi de recueillir des formes sans trait d'union (« non linguiste »), le format de requête retenu fut le suivant : « *non*-mot »; ceci est illustré dans le tableau 1 ci-après.

Formes attestées dans Morphalou	*Requêtes*	*Formes collectées par le moissonneur*
linguistique	non-linguistique	non-linguistique; non linguistique
lingule	non-lingule	non-lingule ; non lingule
lingère	non-lingère	non-lingère ; non lingère
linier	non-linier	non-linier ; non linier
liniment	non-liniment	non-liniment ; non liniment

Tableau 1. Format de requête et formes collectées.

2.3. Vérification automatique de l'attestation sur Google™ des lexèmes générés

Une fois les lexèmes générés à partir du corpus de référence, j'ai utilisé un moissonneur (Berland & Grabar 2001) qui a automatisé l'interrogation des pages en français indexées par le moteur de recherche *Google™*. L'interrogation des pages s'est faite en avril et mai 2013. Les requêtes ont été protégées avec des guillemets («») afin que ce soit l'ensemble « *non*-mot », et uniquement cet ensemble, qui soit l'objet de la requête (évitant ainsi que, par exemple, une requête de format « non-linguistique » recueille des formes contenant « linguistique » mais pas « non »). Le moissonneur m'a donc permis de repérer toutes les formes graphémiques commençant par la chaîne de caractères « non » générées à partir du corpus de référence, où « non » peut être suivi d'une espace ou d'un trait d'union.

Le moissonneur m'a fourni deux types de résultats: une sortie qui indiquant la fréquence ainsi que la première URL d'attestation (3), et une autre sortie n'indiquant que la fréquence (4) :

(3) google|lang_fr|2|»»»»non-linguistique»»» - Recherche Google»|Environ 35 000|
google|lang_fr|»»non-linguistique»»|1|http://eurofle.wordpress.com/2008/08/19/le-linguistique-et-le-non- linguistique-en-intercomprehension/
google|lang_fr|2|»»»»non-lingule»»» - Recherche Google»|Environ 0|
google|lang_fr|»»non-lingule»»|0

(4) non-linguistique|35 000
non-lingule|0

Ainsi, au moment de l'interrogation par le moissonneur, « non-linguistique » a recueilli 35 000 attestations tandis que « non-lingule » n'en a recueilli aucune.

2.4. Traitement du bruit

Il était nécessaire de procéder à un tri manuel afin de filtrer les formes candidates. En effet, une requête avec le mot « non » génère un bruit important. Un travail antérieur portant sur les lexèmes nominaux préfixés par *non-* a montré que sur un corpus de 1 601 562 mots issu de la Toile, seules 3.53% des occurrences de « non » correspondaient à des noms préfixés par *non-* (Dugas 2012). Je vais ici indiquer les principales causes de bruit (cf. Fradin *et al.* 2008 et Hathout *et al.* 2009 pour une discussion générale sur l'apprêt des données issues de la Toile).

Une des particularités du préfixe *non-* en français est qu'il est homographe de l'adverbe de phrase *non,* qui est une forme libre employée en syntaxe. D'autre part, comme indiqué plus haut, les requêtes ont été protégées avec des guillemets ; mais il a fallu prendre en compte le fait que l'espace entre les guillemets vaut pour tout caractère alphanumérique, dont la ponctuation. Par conséquent, le bruit a majoritairement été causé par les formes syntaxiques, que le robot d'extraction ne peut pas discriminer (5), et par la ponctuation (6) :

(5) La propagande a un effet de vérité mobilisateur, mais non émancipateur.
(6) Je m'attendais plutôt à mâcher un bout de poulet sec, mais non. Surprenant.

Deuxièmement, malgré une requête sur les pages françaises uniquement, les données recueillies par le robot d'indexation contenaient quelques occurrences de la forme *non* qui, sans être des emprunts, appartiennent à d'autres langues que le français. Il s'agit notamment de formes de l'anglais, de l'italien et du latin, où existe une forme *non* homographe du *non* français.

D'autre part, le serveur de reconnaît pas les accents. Ainsi, par exemple, une requête de forme « non sûr » a recueilli *non sûr* mais également *non sur*, où *sur* sans accent circonflexe est une préposition, et où *non* est un adverbe. De la même façon, les fréquences des formes obtenues avec la requête « non épouvante » (où *non-* s'adjoint au nom ÉPOUVANTE) comprennent également les occurrences de *non épouvanté* (où *non-* s'adjoint à l'adjectif ÉPOUVANTÉ).

Enfin, les productions langagières issues de la Toile peuvent contenir de nombreuses coquilles et fautes d'orthographe, ce dont il a également fallu tenir compte. Fait bien connu, cette fluctuation orthographique est particulièrement avérée dans les contextes non normés, où les scripteurs ont tendance à s'exprimer très rapidement, sans relecture : messages postés sur les forums de discussion, commentaires de lecteurs sur les sites de presse, etc. Par exemple, la forme *non épaule* dans un énoncé comme *Charles Aznavour chante Au creux de non épaule à l'Olympia* est un faux positif. Au contraire, les articles de journaux et les textes publiés sur des sites administratifs ou officiels (sites gouvernementaux, etc.) contiennent peu de coquilles. D'autre part, et pour faire le lien avec le point précédent, on note une forte tendance à l'omission des accents ou à leur emploi inapproprié (p.ex. *non-progres* ou *non-progrés* pour *non-progrès*), mais cela n'a aucune incidence sur le bruit étant donné que les accents ne sont pas reconnus.

Étant donné l'ampleur du bruit et l'impossibilité de le corriger, du fait de la quantité considérables de données, le choix a été fait de répartir les statistiques obtenues dans 12 ensembles de fréquences : < 500 occurrences ; 500 occ. > 1000 occ. ; 1000 occ. > 5000 occ. ; 5000 occ. > 10 000 occ. ; 10 000 occ. > 50 000 occ. ; 50 000 occ. > 100 000 occ. ; 100 000 occ. > 500 000 occ. ; 500 000 occ. > 1 million occ. ; 1 million occ. > 5 millions occ. ; 5 millions occ. > 10 millions occ. ; 10 millions occ. > 50 millions occ ; 50 millions occ. > 100 millions occ. ; > 100 millions occ. L'analyse des fréquences a ensuite reposé sur la comparaison des statistiques de chaque ensemble.

3. Analyse des fréquences

3.1. Les fréquences

Le tableau 2 indique les fréquences d'attestation sur *Google™* des lexèmes préfixés générés.

Fréquences	*Noms*	*Ratio/N*	*Adjectifs*	*Ratio/Adj*	*Total N+Adj*	*% N+Adj*
≥ 100M	7	0.01%	3	0.01%	10	0.01%
≥ 50M	9	0.01%	0	0%	9	0.01%
≥ 10M	38	0.06%	13	0.05%	51	0.06%
≥ 5M	45	0.07%	11	0.04%	56	0.06%
≥ 1M	213	0.36%	121	0.53%	334	0.41%
≥ 500 000	222	0.37%	116	0.51%	338	0.41%
≥ 100 000	1153	1.94%	617	2.71%	1770	2.16%
≥ 50 000	1020	1.71%	460	2.02%	1480	1.80%
≥ 10 000	3942	6.64%	1948	8.56%	5890	7.18%
≥ 5000	2227	3.75%	1121	4.93%	3348	4.08%
≥ 1000	5488	9.25%	2557	11.24%	8045	9.80%
≥ 500	2569	4.32%	1116	4.90%	3685	4.49%
< 500	*42 401*	*75.45%*	*14 650*	*64.4%*	*57051*	*69.52%*
Total	*59 334*	*100%*	*22 733*	*100%*	*82 067*	*100%*

Tableau 2. Fréquence sur *Google™* des lexèmes préfixés générés (12 ensembles).

Le tableau 3 donne un compte-rendu simplifié des résultats, regroupés en 4 grands ensembles de fréquences :

Fréquence	*Noms*	*Ratio/N*	*Adjectifs*	*Ratio/Adj*	*Total N+Adj*	*% N+Adj*
≥ 500 000	534	0.9%	264	1.2%	798	1%
≥ 50 000	2173	3.7%	1077	4.7%	3250	3.9%
≥ 5000	6169	10.4%	3069	13.5%	9238	11.3%
< 5000	*50 458*	*85%*	*18 323*	*80.6%*	*68 781*	*83.8%*
Total	*59 334*	*100%*	*22 733*	*100%*	*82 067*	*100%*

Tableau 3. Fréquence sur *Google™* des lexèmes préfixés générés (4 ensembles).

On constate que la plupart des formes générées sont très peu fréquentes : 69.52% des lexèmes générés recueillent moins de 500 occurrences (tableau 2), et 83.8% des lexèmes générés recueillent moins de 5000 occurrences (tableau 3). Une comparaison de ces statistiques avec celles que l'on obtient sur *Google™* pour les lexèmes déjà

préfixés par *non-* attestés dans le *TLFi* (tableau 4, ci-dessous) montre que ces derniers recueillent un plus grand nombre d'attestations sur la Toile que les formes préfixées générées : 43.7% des lexèmes préfixés par *non-* attestés dans le *TLFi* comptent plus de 5000 attestations, contre seulement 11.3% des lexèmes préfixés générés. Ceci va à l'encontre de l'hypothèse selon laquelle les dictionnaires ne reflètent pas l'état actuel de la langue. Toutefois, on peut objecter que l'écart de fréquence entre les lexèmes générés et les lexèmes déjà préfixés par *non-* dans le *TLFi* est relativement minime et que les lexèmes attestés dans le *TLFi* ne recueillent pas des fréquences très élevées.

Fréquence	*Noms*	*Ratio/N*	*Adjectifs*	*Ratio/Adj*	*Total N+Adj*	*% N+Adj*
≥ 500 000	20	13.6%	5	11.6%	25	13.2%
≥ 50 000	50	34%	*19*	*44.2%*	69	36.3%
≥ 5000	*66*	*44.9%*	17	39.5%	*83*	*43.7%*
< 5000	11	7.5%	2	4.7%	13	6.8%
Total	*147*	*100%*	*43*	*100%*	*190*	*100%*

Tableau 4. Fréquence sur *Google™* des *non*-N et des *non*-Adj attestés dans le *TLFi*.

D'autre part, 25% des *non*-N et des *non*-Adj générés ne sont pas attestés dans *Google™* au moment de la requête, alors que leur forme non préfixée est attestée. Il s'agit principalement de mots construits (≈ 95% du total des lexèmes générés qui ne recueillent aucune occurrence). Le tableau 5 donne des exemples de ces lexèmes générés qui ne recueillent aucune occurrence sur *Google™*, et indique la fréquence, sur ce même corpus, des bases de ces lexèmes. Ainsi, alors que NON MULTICOQUE ne recueille aucune attestation, MULTICOQUE est tout à fait fréquent puisqu'il recueille 1 670 000 occurrences.

Lexème préfixé	*Fréquence sur* Google™	*Lexème-base*	*Fréquence sur* Google™
NON-MULTICOQUE	0	MULTICOQUE	1 670 000
NON-ORANGERAIE	0	ORANGERAIE	458 000
NON-ENFIÉVRÉ	0	ENFIÉVRÉ	197 000
NON-BALEINIER	0	BALEINIER	194 000
NON-INTERSYNDICAL	0	INTERSYNDICAL	166 000
NON-AFFECTUOSITÉ	0	AFFECTUOSITÉ	139 000
NON-DÉGRINGOLANT	0	DÉGRINGOLANT	125 000
NON-VIDÉOTRANSMISSION	0	VIDÉOTRANSMISSION	113 000

Tableau 5. Exemples de fréquences des formes générées vs. fréquences des bases.

Le cas des lexèmes générés ne recueillant aucune occurrence est tout à fait intéressant. La productivité du préfixe *non-*, entendue comme l'aptitude de la préfixation en *non-* à former de nouveaux lexèmes (cf. Corbin 1987) a été soulignée à plusieurs reprises dans la littérature (Jespersen 1917, Zimmer 1964, Gaatone 1971, Hamawand 2009, Dugas 2012 entre autres), suggérant, d'une part, qu'il n'y aurait pas de restriction à la préfixation en *non-*, c'est-à-dire que n'importe quel lexème (quelles que soient ses caractéristiques morphologiques et phonologiques) pourrait être l'input d'une préfixation par *non-* ; d'autre part, que plus un lexème est fréquent, plus sa forme négative en *non-* serait fréquente. Or la non-attestation de certaines formes en *non-* est un argument en faveur de l'existence de contraintes à la préfixation en *non-* en français contemporain. Tout d'abord, on peut légitimement supposer que les lexèmes les moins fréquents, parce qu'ils appartiennent soit à un domaine de langue spécialisé, soit à un dialecte régional, ou parce que leur référent n'est plus courant aujourd'hui, ont peu de probabilité d'apparaître avec le préfixe *non-*. Ainsi, ne recueillent aucune occurrence des lexèmes comme NON-ABERROGRAPHE, NON-GOUTTIER, NON-GRISOUMÈTRE. Toutefois, NON-ACIÉRISTE, NON-ACROBATISME, NON-AFRICANISME, NON-GIFLEUR, NON-MULTICOQUE, NON-ORANGERAIE, NON-PRÉFIGURATEUR, dont les lexèmes bases sont pourtant *a priori* plus fréquents que ne le sont ABERROGRAPHE, GOUTTIER et GRISOUMÈTRE, ne sont pas attestées dans *Google*^{TM,}. Se pose donc la question des facteurs qui entrent en jeu dans leur apparente incapacité à prendre la préfixation en *non-* - mais ceci dépasse les limites de cet article.

3.2. Tri manuel et examen des formes les plus fréquentes

Dans un second temps, je me suis focalisée sur les 798 formes recueillant plus de 500 000 occurrences et j'ai procédé à leur tri manuel. Les formes éliminées l'ont été dans les cas suivants :

(i) Lorsque le moissonneur a retenu le nombre d'occurrences de la forme suggérée par *Google*^{TM}. Il arrive en effet que le moteur de recherche suggère des formes dont la fréquence est beaucoup plus importante que la forme qui fait l'objet de la requête : dans mon cas, il s'est agi par exemple de *non-gage* pour *non-aiguage*, *non-adhésive* pour *non-athésie*, *non-fiction* pour *non-coction*. Parfois, la forme suggérée était un lexème négatif préfixé par *in-* : *incertitude* à la place de *non-certitude*, *inquiétude* à la place de *non-quiétude*, *invertébré* à la place de *non-vertébré*. Ceci est d'ailleurs intéressant du point de vue de la fréquence respective, pour une même base, d'un dérivé en *non-* et d'un dérivé en *in-*. D'autre part, le moteur de recherche a parfois renvoyé une forme dont la première chaîne de caractère était *in,* non pas le préfixe de négation, mais le préfixe homographe à valeur inchoative : par exemple, *incarnation* à la place de *non-carnation*, *incinération* pour *cinération*, *infloresence* pour *non-florescence*.

(ii) Les formes qui ont un homographe très fréquent en anglais, italien etc. Il s'agit notamment de *non-agricultural, non-business*, *non-living-room*, *non-credo, non-mi, non-perché*.

(iii) Les formes qui correspondent très souvent à des expressions en syntaxe, par exemple *non car, non-même, non merci, non-pour-soi, non-son*. Ces groupes de mots peuvent être employés en réponse à une question, où *non* est employé comme adverbe, par exemple *viens, on tombe amoureux ? non car je suis déjà amoureux de toi*.

À l'issue de ce tri manuel, 108 des 798 formes en *non-* recueillant plus de 500 000 occurrences ont été retenues. Elles ont été classées selon les caractéristiques morphologiques de leur base, i.e. selon que la base était (i) elle-même dérivée par suffixation sur base verbale (noms portant les suffixes *-ade, -age, -ance, -ée, -ment, -sion, -tion, -ure* ; adjectifs portant les suffixes *-eur, -able*, *-ible, -uble*, *-if, -oire*), (ii) dérivée sur base adjectivale (noms portant les suffixes *-ité*, *-té*, *-eté*, *-eur, -esse, -ise*, *-ice*, *-ion, -erie, -ie, -itude*, *-étude*), (iii) dérivée sur base nominale (noms portant les suffixes *-ade, -age, -ance, -aie, -aille, -at, -erie, -ier, -ure* ; adjectifs portant les suffixes *-aire, -al, -el, -esque, -eux, -ien, -ier, -ique, -u, -ais, -ain, -an, -oïde, -ois, -ote*), (iv) apparentée à un participe passé ou présent, (v) non construite en synchronie. Le tableau 6 donne un aperçu des résultats obtenus : les formes en *non-* les plus fréquentes possèdent soit une base apparentée à un participe, soit une base simple.

Morphologie	*Fréquences*	*Exemples*
apparentés à des participes	41	non-adapté, non-communiqué non-défini, non-renseigné
simplex	33	NON-FICTION, NON-RESPECT NON-DISPONIBLE, NON-TOXIQUE
déverbaux	19	NON-FUMEUR, NON-UTILISATION NON-NÉGLIGEABLE, NON-NÉGOCIABLE
dénominaux	6	non-commercial, non-contractuel, non-ferreux, non-professionnel
désadjectivaux	4	non-conformité, non-franchise, non-responsabilité, non-violence
autres	5	NON-HOMOLOGUE, NON-LOF

Tableau 6. Exemples de formes les plus fréquentes sur *Google*^{TM} (> 500 000 occ.) et leurs caractéristiques morphologiques.

Enfin, il m'a paru intéressant de comparer les fréquences des lexèmes générés avec la fréquence de leur base (c'est-à-dire sans le préfixe *non-*). Par exemple, comme l'indique le tableau 7, NON-CONTRE-INDICATION est plus fréquent que CONTRE-INDICATION puisque le rapport de la forme préfixée à la forme non préfixée est supérieur à 1. D'autres formes préfixées sont relativement fréquentes (quoique moins fréquentes que leur base), comme NON-CUMULABLE (rapport de 0.78), NON-FERREUX (rapport de 0.68).

[4] Livre des Origines Français, qui répertorie les origines des chiens de race français, ici utilisé avec valeur d'adjectif pour désigner un chien (non) répertorié au LOF.

Lexème préfixé (1)	*Lexème non préfixé (2)*	*Ratio (1)/(2)*
NON-CONTRE-INDICATION	CONTRE-INDICATION	1.75
NON-CUMULABLE	CUMULABLE	0.78
NON-LOF[1]	LOF	0.76
NON-FERREUX	FERREUX	0.68
NON-LUCRATIF	LUCRATIF	0.63
NON-FUMEUSE	FUMEUSE	0.61

Tableau 7. Ratio forme préfixée/forme non préfixée.

On note toutefois que la quasi-totalité (102 sur 108) des lexèmes préfixés ont une fréquence inferieure à 50% de celle de leur forme non préfixée (i.e. avec un ratio inférieur à 0,5).

3. Conclusions et perspectives de recherche

Dans cet article ont été retracées les différentes étapes d'un travail conjuguant méthodes de TAL et linguistique théorique et consistant à détecter, à extraire automatiquement et à analyser les lexèmes construits sur base nominale et adjectivale à l'aide du préfixe *non-* en français dans les pages en français indexées par *Google*™. Le principal apport de ce travail est d'illustrer le fait les données dictionnairiques ne reflètent qu'un état limité de la langue, puisque les lexèmes en *non-* effectivement employés par les locuteurs ne se limitent pas à ceux qui sont attestés dans le *TLFi*. Toutefois, les résultats obtenus suggèrent l'existence de limites à la capacité de la préfixation en *non-* à créer de nouveaux lexèmes, étant donné d'une part que 25% des formes générées n'étaient pas attestées dans le corpus examiné au moment de la requête, et que, d'autre part, les lexèmes générés apparaissent à des fréquences variables.

Je dispose à l'issue de ce travail d'un corpus extensif de lexèmes préfixés par *non-* sur base nominale et adjectivale, qui va me permettre de rendre compte des éventuelles contraintes morphologiques, sémantiques et phonologiques qui entrent en jeu dans leur formation, et de proposer une interprétation sémantique de ces dérivés. J'espère ainsi pouvoir approfondir la description et l'analyse de ce type de préfixation négative, qui n'a encore jamais fait l'objet d'un examen systématique en morphologie.

UMR STL 8163, Université Charles de Gaulle - Lille 3 Edwige DUGAS

Références bibliographiques

Berland, Sophie / Grabar, Natalia, 2001. « Construire un corpus Web pour l'acquisition terminologique », *Terminologie et intelligence artificielle* (TIA), 44-54.

Corbin, Danielle, 1987. *Morphologie dérivationnelle et structuration du lexique*, volume 1, Lille, Presses Universitaires de Lille.

Dal, Georgette / Namer, Fiametta, 2012. « Faut-il brûler les dictionnaires ? Ou comment les ressources numériques ont révolutionné les recherches en morphologie », *SHS Web of Conferences* 1, 1261-1276.

Dugas, Edwige, 2013. « [non(-)Adj] sequences in contemporary French: morphological negation, syntactic negation, or in between? », Communication faite lors du Symposium *Morphologie et ses interfaces*, Université de Lille 3, 12-13 septembre 2013.

Dugas, Edwige, 2012. *La négation en morphologie : le cas des formes nominales en* non- *en français*, Mémoire de Master 2 (non publié), Université de Lille 3.

Fradin, Bernard et al., 2008. « Remarques sur l'usage des corpus en morphologie », *Langages* 171, 34-59.

Gaatone, David, 1971. *Étude descriptive du système de la négation en français contemporain*, Genève, Librairie Droz.

Hamawand, Zeki, 2009. *The semantics of English negative prefixes*, Equinox Publications.

Hathout, Nabil et al., 2009. « La collecte et l'utilisation des données en morphologie », in : Fradin, Bernard / Kerleroux, Françoise / Plénat, Marc (ed.), *Aperçus de morphologie du français,* Presses universitaires de Vincennes, 267–287.

Jespersen, Otto, 1917. « Negation in English and other languages », in : *Selected Writings of Otto Jespersen*, Routledge, 3-151.

Muller, Claude, 1991. *La Négation en français : syntaxe, sémantique et éléments de comparaison avec les autres langues romanes*, Genève, Librairie Droz.

Romary, Laurent et al., 2004. « Standards going concrete: from LMF to Morphalou », in : Zock, Michael / Saint-Dizier, Patrick (ed), *COLING 2004 Enhancing and using electronic dictionaries*, Geneva, 22-28.

Zimmer, Karl, 1964, « Affixal negation in English and other Languages: an Investigation of Restricted Productivity », Supplement to *Word* 20 (2).