Edwige DUGAS
Section 16 – Projets en cours ; ressources et outils nouveaux

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

Ce poster présente un travail en cours qui vise à mettre à contribution linguistique théorique et Traitement Automatique des Langues (TAL), dans le cadre d'une thèse consacrée à l'étude synchronique et diachronique du préfixe[1] de négation *non-* et sa relation aux autres préfixes composant le paradigme de la préfixation négative : *a-, anti-, contre-, dé(s)-, dis-, in-* (et ses variantes *il-, im-, ir-* etc.), *mal-, mé(s)-*, *non-*. J'expose ici la méthodologie employée pour comparer les lexèmes construits sur base nominale et adjectivale[2] à l'aide du préfixe *non-* (p.ex. NON VIOLENCE, NON VOYANT, NON REMBOURSABLE) attestés dans le *Trésor de la Langue Française informatisé* (*TLFi*), et les formes attestées dans un corpus ouvert (en l'occurrence, les pages indexées par le moteur de recherche *Google*^{TM}). L'hypothèse sous-jacente est que les dictionnaires ne reflètent que très partiellement l'état actuel du français (cf. Dal et Namer 2012). Mon objectif est, à terme, de créer un outil d'extraction automatique des mots construits à l'aide du préfixe *non-* et à ceux construits avec les autres préfixes de négation.

1. Constitution du corpus et détection des formes

La méthodologie proposée est la suivante :
(1) Constitution du lexique de référence, à savoir les 22 790 formes attestées comme adjectifs et des 60 940 formes attestées comme noms contenus dans le lexique Morphalou[3], constitué à partir du *TLFi.* Parmi ces formes, on compte 173 lexèmes préfixés par *non-* (138 noms, 20 adjectifs, et 15 lexèmes présentés comme possédant une double catégorisation nom et adjectif).
(2) À partir de ce lexique, génération de tous les lexèmes préfixés par *non(-)* potentiels avec deux orthographes (p.ex. *nonprogrès* et *non progrès*[4]) ;
(3) Vérification manuelle et automatique de l'existence des lexèmes ainsi construits par des requêtes réalisées entre le 16/11 et le 29/11/2012 sur les pages en français indexées par le moteur de recherche *Google*^{TM} ;
(4) Examen et analyse des fréquences obtenues ;
(5) Interprétation des résultats.

2. Vérification et analyse des données

Les difficultés techniques sont de plusieurs sortes. Les requêtes réalisées à l'étape (3) génèrent beaucoup de faux positifs (cf. Fradin et *al.* 2008 et 2009 pour une discussion générale sur l'apprêt des données issues de la Toile) : (i) *non* a des emplois en syntaxe que le robot d'extraction ne peut pas discriminer ; (ii) l'espace entre les "" vaut pour tout caractère non alphanumérique, dont la ponctuation ; (iii) le serveur ne reconnaît pas les accents (p.ex. une requête avec "non épouvante" récolte aussi *non épouvanté*) ; (iv) malgré des requêtes sur les pages en français, on récolte parfois des pages en anglais ; (v) les coquilles et les fautes d'orthographe sont nombreuses. Ceci explique pourquoi le choix a été fait de regrouper les formes détectées dans des ensembles de fréquences (entre 10 millions et 1 million, entre 1 million et 100 000, etc.) et de comparer le nombre de formes dans chaque ensemble. Il a en effet été montré dans un travail précédent (Dugas 2012) consacré aux *non*-N (du type NON VIOLENCE) et basé sur un corpus de 1 601 562 mots issu de la Toile (74 432 occurrences de *non*, 2626 tokens, 464 types), qu'une requête sur le mot *non* génère un bruit important : seules 3.53% des occurrences de *non* correspondaient à des noms préfixés par *non-*.

---

[1] Par commodité, je parlerai dans cette proposition du « préfixe » *non-*, même si seule une analyse morphologique approfondie me permettra de décider du statut (préfixal ou non) de *non-*.
[2] Dans de très rares cas, *non-* s'adjoint à une base verbale. J'ai décidé de ne pas tester cette possibilité dans le cadre de ce poster.
[3] http://www.cnrtl.fr/lexiques/morphalou/
[4] Une requête avec « non progrès » permet d'obtenir à la fois les occurrences de *non progrès* et de *non-progrès*.

3. Résultats

Résultats intermédiaires. Au moment de la remise de cette proposition ont été réalisées 7532 requêtes par le moissonneur (1422 noms et 2344 adjectifs) et 7388 requêtes manuelles (845 noms et 2849 adjectifs), soit un total de 2267 noms et 5193 adjectifs. L'analyse de ce premier échantillon de données nous apprend que :

- les lexèmes générés apparaissent avec des fréquences variables, tout comme les 174 lexèmes préfixés par *non-* attestés dans le *TLFi* :

		Noms	Ratio/Noms	Adjectifs	Ratio/Adj
Fréquences total lexèmes générés	**> 1 million**	1	0.04%	9	0.17 %
	> 100 000	21	0.93%	93	1.79 %
	> 10 000	131	5.78%	393	7.57 %
	> 1000	245	10.81%	831	16.00 %
	> 100	292	12.88%	853	16.42 %
Fréquences lexèmes en *non-* attestés dans le *TLFi*	**> 1 million**	12	8.69%	3	8.57 %
	> 100 000	34	24.64%	12	34.28 %
	> 10 000	72	52.17%	16	45.71 %
	> 1000	20	14.49%	3	8.57 %
	> 100	0	0%	0	0 %

- environ 37% des noms et 29% des adjectifs ne renvoient aucun résultat, alors qu'une requête sur la forme non préfixée est positive, p.ex. *non épaississeur* vs. *épaississeur* (31800 occ.) ;
- la proportion de formes où *non* est graphiquement soudé à la base est négligeable ; il n'est donc pas nécessaire de faire deux requêtes pour chaque lexème généré.

Résultats attendus. Il a déjà été procédé à un essai de quantification de la répartition des types sémantiques des bases nominales (40.4% dénotent des actions, 20.1% des êtres humains, 14% des états et des abstractions, 8.1% des objets) (Dugas 2012). Les étapes (3) et (4) présentées plus haut seront poursuivies et je disposerai d'un corpus extensif de lexèmes préfixés par *non-* sur base nominale et adjectivale qui me permettra d'approfondir la description et l'analyse morpho-sémantique de ce type de préfixation négative.

4. Perspectives de travail

Mon objectif est de disposer d'un lexique aussi exhaustif que possible des lexèmes formés par la préfixation en *non-,* en synchronie mais aussi en diachronie. J'étendrai pour cela mon corpus à la base textuelle Frantext, et, pour les formes antérieures à 1600, aux textes de la Base de Français Médiéval (BFM) et du Corpus représentatif des premiers textes français (CoRPTeF)[5]. Enfin, la détection ciblée de formes construites en *non-* dans un corpus ouvert n'est pas la seule méthode d'investigation. J'envisagerai également (i) la dépréfixation, qui consiste à enlever au lexème la chaîne de caractères correspondant formellement à un préfixe de négation et à vérifier si ce qui subsiste après la dépréfixation est un mot existant dans le lexique ; (ii) l'analyse morphologique à l'aide de DériF[6], qui décompose le lexème complexe en constituants (base, affixes), mais qui reste encore à améliorer – le présent travail pourrait aussi y contribuer.

Bibliographie indicative

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

Dal, G. et *al*. 2004. « Morphologie constructionnelle et traitement automatique des langues : le projet MorTAL », *Lexique* 16, pp. 199-229.

Dugas E., 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 B. et *al*. 2008. « Remarques sur l'usage des corpus en morphologie ». *Langages* 171, pp. 34-59.

Hathout N. et *al*. 2009. « La collecte et l'utilisation des données en morphologie ». In Fradin B., Kerleroux F. et Plénat M. (éds.), *Aperçus de morphologie du français,* pp. 267–287.

[5] http://txm.bfm-corpus.org/bfm/ et http://corptef.ens-lyon.fr/

[6] *Dérivation en français* ; ressource accessible à partir du site http://www.cnrtl.fr/outils/DeriF/