SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100251

La préfixation française à travers les genres et les domaines : étude de corpus[1]

Lefer, Marie-Aude

Université catholique de Louvain et Institut libre Marie Haps
marie-aude.lefer@uclouvain.be

1 Introduction

La linguistique de corpus a amorcé une véritable révolution en morphologie, en particulier dans le domaine de la formation des mots (voir Hathout et al., 2009). Elle a notamment permis d'établir que la productivité des affixes dérivationnels varie considérablement à travers les registres (p. ex. langue écrite vs langue orale), les genres (p. ex. éditoriaux de presse vs articles d'information) et les domaines (p. ex. médecine vs physique). Ainsi, dans leur grammaire basée sur corpus, Biber et al. (1999) ont montré que les affixes anglais sont généralement plus productifs en prose académique et journalistique qu'en langue orale.

Bien que cet aspect de la morphologie soit maintenant largement reconnu, pour l'anglais surtout (cf. Plag et al., 1999), peu d'études à grande échelle ont été menées jusqu'à présent sur la langue française (voir Grabar et al., 2006). L'étude dont il est question ici se penche sur la variation de la productivité de la préfixation française à travers trois genres écrits (articles scientifiques, éditoriaux de presse, romans) et trois domaines académiques (médecine, économie, linguistique). Notre étude est intégralement basée sur corpus et se donne pour objectif de brosser un premier portrait de la préfixation française dans son ensemble en examinant la productivité réalisée (Baayen, 2008) de près de 100 préfixes[2].

2 État de la question et objectifs de l'étude

2.1 Tour d'horizon des études variationistes

Cette section a pour but de dresser un **état de la question** des études variationistes[3], en particulier celles qui se sont intéressées à la morphologie lexicale. Cet état de la question aborde l'anglais, puisque les études pionnières en la matière ont principalement examiné cette langue, et le français, la langue qui nous occupe ici.

Dans la *Longman Grammar of Spoken and Written English*, Biber et al. (1999 : 24) soulignent que

> Beaucoup de descriptions globales de l'anglais général sont incomplètes et peuvent même induire en erreur ou se révéler inexactes. Les descriptions de l'anglais général, qu'on base sur la moyenne des patrons observés à travers les textes et les registres, masquent souvent des différences importantes et généralement, en fait, ne représentent correctement les patrons d'aucun registre. [Notre traduction]

La morphologie a longtemps été le parent pauvre des études variationistes, comme Chafe (1982), Chafe et Danielewicz (1987) ou encore Biber (1988). Par exemple, dans Biber (1988), seules les nominalisations anglaises en *-ment*, *-ness*, *-ity* et *-ation* entrent en ligne de compte. En réaction à cet état de fait, Baayen (1994) a montré que les textes peuvent être classés en fonction de la productivité (de la rentabilité, pour

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100251

être plus précis - la notion de productivité sera brièvement définie ci-dessous) des affixes qu'ils contiennent. Son étude indique également qu' « une théorie de la productivité morphologique qui ne tient pas compte des facteurs stylistiques est incomplète » (notre traduction). Vers la fin des années 90 et le début des années 2000, d'autres travaux ont abordé cet aspect de la variation morphologique à travers les registres et les genres. Plag et al. (1999) ont étudié la productivité de quinze suffixes anglais dans les sous-corpus « écrit » et « oral » du *British National Corpus*. Il ressort que la suffixation anglaise est généralement plus productive dans les textes écrits qu'oraux et que certains suffixes sont plus sensibles que d'autres à la variation à travers les registres. L'étude des préfixes verbaux anglais de Schröder (2008) abonde également dans ce sens. Malgré ce nouvel intérêt pour la morphologie dans différents registres, la plupart des études se sont limitées aux suffixes. Seuls Biber et al. (1999) ont étudié les préfixes anglais de manière relativement exhaustive : 40 préfixes ont été examinés dans quatre registres (langue orale, fiction, prose journalistique et académique). Leurs résultats indiquent également une plus grande productivité de la préfixation dans les registres écrits, en particulier la prose académique.

Peytard (1975) a mis en lumière l'influence du registre sur la préfixation française en comparant les préfixés en langue orale (émissions radio et conversations) et écrite (romans). Néanmoins, son étude incluait des dérivés opaques (comme *présenter*, *prévenir* et *préserver*), qui ne sont généralement pas pris en compte dans les études de productivité « à la Baayen ». De plus, Peytard reconnaît le caractère quelque peu imprécis de ses conclusions, l'analyse étant tout à fait manuelle, et donc potentiellement erronée. Plus récemment, Grabar et al. (2006) ont étudié la productivité des suffixes *-ité* et *-able* dans un corpus journalistique de 25 millions de mots (*Le Monde*). Cette étude constitue, à notre connaissance, la première analyse variationiste et basée sur corpus de l'affixation française. Leur approche est variationiste dans la mesure où la productivité des deux suffixes a été calculée dans les différentes rubriques du journal (international, France, culture, sports, programmes radio et télé, etc.). Ces rubriques ne correspondent pas *stricto sensu* à des genres ou sous-genres mais aux différentes sections du journal, qui représentent chacune une ou plusieurs thématiques. Grabar et al. montrent que *-able* est plus productif dans les rubriques « société » et « livres ». La combinaison de suffixes *-ibilité/-abilité*, quant à elle, est la plus productive dans la rubrique « international » (p. ex. *crédibilité*, *brevetabilité*, *inviolabilité*, *inéligibilité*, *manœuvrabilité*, *opérabilité*, *rentabilité*). Dans la même veine, Grabar et Zweigenbaum (2003) se sont penchés sur la variation à travers les registres (langue générale vs langue spécialisée : la langue médicale), les genres (rapports hospitaliers vs articles scientifiques en langue médicale) et les domaines médicaux (cardiologie, neurologie, etc.). Leur étude porte sur les suffixes qui forment des adjectifs dénominaux, à savoir *-aire*, *-al*, *-el*, *-ien*, *-eux*, *-é, -ique*, *-in*, *-ais* et *-ois*, et indique que la productivité de ces suffixes varie à travers les registres, les genres et les domaines (voir également Condamines, 2003 et L'Homme, 2004). Tout comme Baayen (1994), Clavier (2006) met en avant le potentiel de l'analyse variationiste de la morphologie pour le profilage automatique des textes : « il nous est apparu que la plupart des noms et adjectifs étaient des mots dérivés et/ou composés extrêmement réguliers, ce qui permettrait d'envisager un profilage purement morphologique des textes ».

Ce tour d'horizon de la littérature nous permet de formuler les trois observations suivantes :

- L'approche variationiste de la morphologie s'est principalement penchée sur la variation à travers les registres (écrit vs oral ; cf. Plag et al., 1999), la micro-variation (p. ex. les différentes rubriques du journal *Le Monde* ; cf. Grabar et al., 2006) et la langue médicale (Grabar et Zweigenbaum, 2003) ;
- La plupart des études variationistes se sont intéressées à la suffixation, au détriment de la préfixation et des autres procédés de formation des mots comme la composition ;
- La préfixation française n'a pas encore fait l'objet d'une étude variationiste systématique basée sur corpus. A dire vrai, la remarque de Peytard (1975 : 5) semble encore valable aujourd'hui : « les chercheurs ne s'aventurent dans ce domaine [la préfixation] qu'avec une certaine réserve : on n'étudie pas la "préfixation", mais un préfixe dont on fait la monographie, rarement un ensemble de préfixes ».

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100251

2.2 But de notre étude

Notre étude a pour **objectif** principal d'examiner la productivité de la préfixation française à travers trois genres écrits (articles scientifiques, éditoriaux de presse, romans) et trois domaines académiques (médecine, économie, linguistique). Le choix de ces genres et domaines a été dicté par un critère très pragmatique : les ressources de corpus à notre disposition dans le cadre d'une étude contrastive anglais-français (Lefer, 2009).

Les principales similitudes et différences entre notre étude et celle présentée par Biber et al. (1999) dans la grammaire Longman sont décrites dans le Tableau 1. Tout comme Biber et al. (1999), notre approche se veut la plus **exhaustive** possible. Nous étudions près de 100 préfixes français (contre 40 préfixes anglais dans Biber et al., 1999). De plus, nous adoptons la même définition de productivité. La notion de **productivité** est particulièrement complexe (voir par exemple Plag, 1999 ; Bauer, 2001 et Dal, 2003). En effet, le terme « productivité » recouvre plusieurs aspects d'un phénomène aux multiples facettes. Dans notre étude, nous nous limitons à un seul aspect de la productivité, appelé 'realised productivity' ou 'extent of use' par Baayen (voir Baayen, 1993 et Baayen, 2008). Il s'agit d'une mesure basée sur corpus qui comptabilise le **nombre total de lexèmes différents formés à l'aide d'un affixe donné** (d'où le terme de « productivité réalisée »). Nous n'étudions pas les autres aspects de la productivité, tels que la rentabilité (angl. 'profitability') ou la disponibilité (angl. 'availability'). La raison principale de ce choix est la taille relativement réduite des corpus utilisés (1 million de mots par genre), qui ne permet pas d'avoir recours à des mesures de rentabilité basées sur le nombre total d'*hapax legomena* (dans les corpus de taille réduite, les *hapax* correspondent rarement à des néologismes). Plag et al. (1999) se basent également sur cette mesure de productivité réalisée pour une partie de leur étude. Comme l'indiquent Baayen et Renouf (1996 : 92), « l'inférence statistique à partir de corpus de taille relativement réduite permet déjà de faire ressortir les principales tendances de la productivité [au sens large] sur base des mots attestés du lexique » (notre traduction). Il est donc utile de souligner que dans la suite de cet article, par « productivité », on entend « **productivité réalisée** ».

Trois **différences** majeures entre notre approche et celle de Biber et al. (1999) doivent être mentionnées : (1) nous examinons la variation à travers les domaines académiques ; (2) notre approche combine l'étude des préfixes individuels et l'étude de catégories sémantiques de préfixes (position locative et temporelle, quantification, négation, etc. ; voir Section 3.5) ; (3) nous ne distinguons pas les différentes catégories grammaticales des lexèmes préfixés. Concernant le premier point, il est intéressant de souligner que près d'un quart du corpus académique utilisé par Biber et al. (1999) est constitué de textes médicaux. Cependant, les auteurs ne mentionnent pas l'influence éventuelle de la langue médicale sur les tendances générales mises au jour pour l'anglais académique. C'est ce que nous tentons de faire ici en comparant la médecine (science dure), l'économie (science sociale) et la linguistique (science humaine). Enfin, notons que deux raisons principales justifient le fait que notre étude n'ait pas recours à l'étiquetage grammatical : tout d'abord, l'étiquetage introduit nécessairement un biais dans les données, n'étant jamais exempt d'erreurs, en particulier pour les mots nouveaux (voir Plag et al., 1999 sur ce point) ; ensuite, il faut garder à l'esprit que l'étude de la préfixation française présentée ici fait partie intégrante d'une étude contrastive à grande échelle (Lefer, 2009). Il est bien connu qu'il est extrêmement difficile de trouver un étiqueteur qui puisse correctement annoter deux langues différentes avec des taux de rappel et de précision comparables (Poudat, 2004). Nous avons donc décidé de ne pas étiqueter les données.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100251

Tableau 1 : Similitudes et différences entre notre étude et Biber et al. (1999)

		Biber et al. (1999) [anglais]	Notre étude [français]
similitudes	**Approche générale**	approche variationiste basée sur corpus	
	Notion de productivité	productivité réalisée (‘realised productivity’) (+ Biber et al. : nombre de dérivés peu fréquents, qui peuvent représenter des néologismes - cette deuxième mesure de productivité n’a pas été systématiquement exploitée par Biber et al.)	
différences	**Définition du concept de « dérivé »**	définition implicite	critères définitoires explicites
	Registres /genres /domaines	4 registres : langue orale (conversations) fiction (romans et nouvelles) prose journalistique (grande variété de rubriques) prose académique (monographies et articles scientifiques, 16 domaines)	3 genres : romans éditoriaux de presse articles scientifiques (3 domaines)
	Corpus	5 millions de mots par registre	1 million de mots par genre
	Étiquetage morpho-syntaxique	Corpus étiqueté (résultats séparés pour chaque catégorie grammaticale)	×
	Étendue de l’étude	Étude relativement complète de la préfixation anglaise (mais, par exemple, la préfixation négative n’est pas abordée) 40 préfixes	Etude exhaustive de la préfixation française 88 préfixes
	Influence du domaine académique	×	√
	Approche sémantique de la préfixation	×	√

2.3 Hypothèses de recherche

L’état de la question présenté dans la Section 2.1, ainsi que des études contrastives menées dans le domaine de la langue académique (p. ex. Fløttum et al., 2006), mènent à la formulation des trois hypothèses suivantes, que nous proposons de tester dans cette étude :

- La productivité de la préfixation française varie à travers les genres et les domaines ;

- La préfixation est plus productive dans les articles scientifiques, suivis des éditoriaux de presse et enfin, des romans. En effet, Biber et al. (1999 : 16) ont montré que la fiction se situe entre les registres écrits et les registres oraux, notamment de par la présence de dialogues, proches d'un style oral ;
- En prose académique, la préfixation est plus productive en médecine qu'en linguistique, alors que l'économie se situe entre ces deux domaines. Cette hypothèse trouve sa source dans l'étude contrastive anglais-français-norvégien de Fløttum et al. (2006), qui indique que la médecine et la linguistique se différencient à bien des égards (par exemple dans l'utilisation de la première personne du singulier ou de la négation), alors que l'économie occupe une position intermédiaire, tantôt proche de la médecine, tantôt similaire à la linguistique. Il nous est possible d'examiner cette hypothèse parce que nous utilisons le sous-corpus français du corpus KIAP, sur lequel se basent Fløttum et al. (2006).

3 Données et méthodologie

3.1 Inventaire des préfixes français étudiés

La première étape de notre travail consiste à dresser un inventaire des préfixes français qui soit aussi exhaustif que possible. Dans un premier temps, nous avons adopté une approche relativement pragmatique et relevé les éléments de formation des mots considérés comme préfixes dans au moins un des travaux suivants (ils sont classés par ordre alphabétique des auteurs) : Amiot (2004, 2005a, 2005b), Apothéloz (2002), Béchade (1992), Corbin (1987), Dubois (1962), Dubois et Dubois-Charlier (1999), Fradin et Montermini (2009), Gaudin et Guespin (2000), Grevisse (1993), Haensch et Lallemand-Riekötter (1972), Huot (2006), Lehmann et Martin-Berthet (2007), Niklas-Salminen (1997) et Thiele (1987). Il s'agit de grammaires du français, de monographies dédiées à la lexicologie, à la morphologie ou à la dérivation et d'études consacrées à un ensemble de préfixes.

Il apparaît rapidement qu'un élément classé comme « préfixe » par un auteur n'est pas nécessairement listé comme préfixe par d'autres. Souvent, des éléments ne sont pas mentionnés ou d'autres étiquettes sont utilisées, à savoir « préfixe savant », « préfixoïde » ou partie de composé (« quasi-lexème » dans le cas des composés néoclassiques). En effet, il est bien connu que de nombreux éléments se situent à la frontière entre la dérivation et la composition (Amiot, 2004), ce qui rend leur définition difficile. Comme le souligne si justement Lieber (2004 : 14), les affixes sont « de petites choses insaisissables » (notre traduction). Comme illustré dans le Tableau 2, qui présente un échantillon de la base de données des préfixes français que nous avons compilée, un élement tel que *après-* est à la fois classé comme préfixe (p. ex. dans Amiot, 2004) et comme élément utilisé dans la formation de mots composés (p. ex. dans Grevisse, 1993).

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100251

Tableau 2 : Échantillon de l'inventaire des préfixes français dans les ouvrages de référence sélectionnés

Élements	Exemples	Préfixes	Préfixes savants	Préfixoïdes	Quasi-lexèmes (éléments de formation des composés néoclassiques)	Prépositions ou adverbes utilisés dans des composés
anté- (anti-)	antéposition (N) antépénultième antéprédicatif (A) antidater (V)	Amiot04 Apothéloz02 Béchade92 Haensch72 Lehmann07 Niklas97 Thiele87		Grevisse93		
anti-	antidopage antihéros (N) antiatomique antirabique (A)	Amiot04 Apothéloz02 Béchade92 Corbin87 Lehmann07 Niklas97 Thiele87			Grevisse93 Huot06	
après-	après-ski (N) après-demain (Adv)	Amiot04 Béchade92 Corbin87 Haensch72 Lehmann07 Thiele87				Grevisse93 Haensch72
arch-	archevêque (N)	Béchade92 Thiele87				
archi-	archiduc archiprêtre (N) archichouette archifaux (A)	Apothéloz02 Béchade92 Corbin87 Haensch72 Lehmann07 Montermini09 Thiele87		Grevisse93	Niklas97	
arrière-	arrière-pays arrière-plan (N)	Apothéloz02 Béchade92 Haensch72 Lehmann07 Thiele87				Grevisse93 Niklas97

auto-	autocensure (N) autosuffisant (A) autodétruire (V)	Apothéloz02 Lehmann07		Grevisse93 (*auto-analyse*) Haensch72	Béchade92 Grevisse93 (*autogène*) Niklas97 Thiele87	

La base de données obtenue contient 119 éléments classés au moins par un auteur comme préfixes, préfixes savants ou préfixoïdes. Nous avons effectué un premier nettoyage en éliminant les éléments exclusivement classés comme préfixes savants ou préfixoïdes (p. ex. *hydro-*, *kiné-*, *proto-*, *simili-*), ainsi que les éléments classés comme préfixes par Thiele (1987) ou Béchade (1992) uniquement. En effet, ces inventaires de préfixes (surtout celui de Thiele) se singularisent de par le fait qu'ils contiennent certains éléments qu'ils sont les seuls à considérer comme préfixes (p. ex. *amphi-*, *hors-*, *presqu', tré-*). Ce premier tri a réduit la liste initiale à **88 préfixes**, à savoir : *a- (ac-, ad-, af-, al-, ap-, ar-, as-, at-), a- (an-), ab- (a, abs-), anté-, anti-, après-, arch-, archi-, arrière-, auto-, avant-, bi- (bis-), bien-, circon-, circum-, cis-, co- (col-, com-, con-, cor-), contra-, contre-, crypto-, dé-¹ (dés-, des-), dé-², demi-, di- (dis-), dis-, dys-, é- (ex-, ef-, es-), en- (em-), entre-, épi-, ex-, exo-, extra-, hémi-, hyper-, hypo-, in-¹, in-² (il-, im-, ir-), infra-, inter-, intra-, intro-, juxta-, macro-, mal-, maxi-, mé- (més-), méga-, méta-, mi-, micro-, mini-, mono-, multi-, néo-, non-, omni-, outre-, pan-, par-, para- (par-), pén- (péné-), per-, péri-, pluri-, plus-, poly-, post-, pour-, pré-, pro-, pseudo-, quasi-, re- (ré-, r-, ra-), rétro-, sans-, semi-, sous- (sou-), sub-, super-, supra-, sur-, sus-, télé-, trans-, ultra-, uni-, vice- (vi-)*.

Tous ces éléments ont été extraits automatiquement des corpus (voir Section 3.2), sans aucun a priori théorique sur leur statut de préfixe en français. Notons aussi que *tri-* n'a pas été repris, bien que listé dans l'inventaire initial, car nous nous sommes limités aux préfixes de quantité nombrée correspondant à « un » et « deux ».

3.2 Corpus utilisés

Comme mentionné précédemment, les données ont été extraites de trois corpus d'un million de mots chacun et composés de romans, d'éditoriaux de presse (tirés des quotidiens *Le Monde*, *Libération*, *Le Figaro*) et d'articles publiés dans des revues scientifiques dans trois domaines (médecine, économie et linguistique). Le corpus de prose académique contient des textes issus de revues scientifiques françaises, canadiennes et belges (contrairement aux deux autres corpus analysés dans cette étude et repris dans le Tableau 3). Néanmoins, nous ne nous sommes pas penchés plus en détail sur l'effet de cette variable.

Tableau 3 : Corpus utilisés

Romans (tirés de *Frantext*)	1.027.036
Éditoriaux de presse (Mult-Ed-FR *Multilingual Editorial Corpus*[4])	993.849
Articles scientifiques (KIAP-FR *Cultural Identity in Academic Prose*[5])	906.341
Médecine	183.412
Linguistique	318.752
Économie	404.177
Nombre total de mots	2.927.226

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100251

3.3 Extraction automatique des données brutes

Dans un premier temps, les données ont été extraites automatiquement sur base d'un programme *perl* qui a permis d'obtenir tous les mots commençant par les séquences de lettres qui correspondent aux préfixes listés ci-dessus, ainsi que leur fréquence dans les différents sous-corpus. Il s'agit donc d'une méthode d'extraction purement formelle. Ces données brutes ont ensuite été soumises à une sélection manuelle afin d'éliminer le bruit (c.-à-d. les mots qui commencent par les séquences de lettres précitées mais ne contiennent en fait pas le préfixe en question ; p. ex. *été* pour *é-*) et les dérivés opacifiés, qu'il n'est pas judicieux de conserver dans les données finales (Plag et al., 1999). Le tri des données brutes passe par l'adoption de critères définitoires explicites des concepts de « mot préfixé » et « préfixe », que nous exposons dans la Section 3.4.

3.4 Tri manuel et critères de sélection des données

Les données brutes ont été soumises à différents filtres de sélection manuelle. L'étape de tri est essentielle en morphologie, comme souligné par Evert et Lüdeling (2001) et Fradin et al. (2003). Les critères de sélection sont décrits dans les Sections 3.4.1 (notion de « mot préfixé ») et 3.4.2 (notion de « préfixe »).

3.4.1 Critères définitoires du concept de « préfixé » et leur application dans la sélection des données

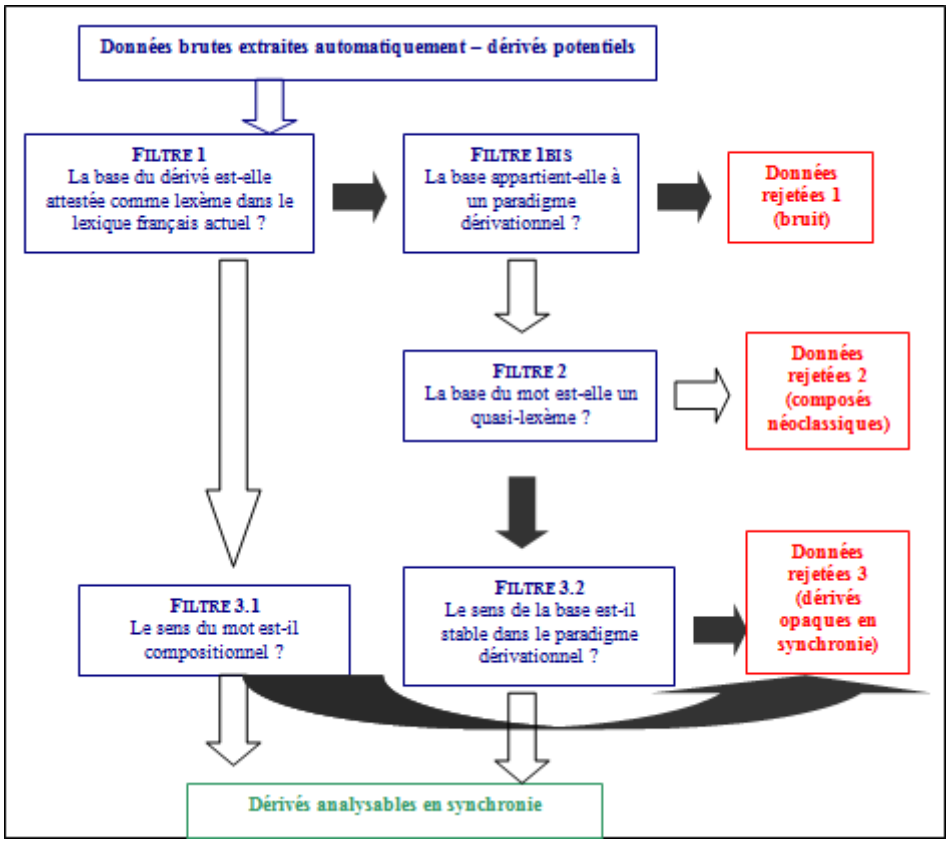
Les critères adoptés dans cette étude pour déterminer si un mot extrait automatiquement sur base de nos requêtes formelles devait être considéré comme un lexème préfixé sont les suivants :

- **Analysabilité formelle** : le préfixé est clairement reconnaissable comme étant constitué d'une base (libre ou liée) et d'un préfixe (voir la Section 3.4.2 pour la définition de « préfixe ») ;
- **Appartenance à un paradigme dérivationnel** : la base du préfixé, si elle ne correspond pas à un lexème du lexique français actuel, se retrouve dans d'autres dérivés où elle exprime un sens similaire (voir Apothéloz, 2002 sur cette notion) ;
- **Transparence sémantique** : le sens du préfixé est (au moins partiellement) dérivable d'un des sens de sa base et d'un des sens du préfixe qu'il contient. Le préfixe et la base collaborent tous deux au sens global du préfixé[6].

Ces critères ont été opérationnalisés dans la procédure de tri des données brutes illustrée dans le Graphique 1. Les flèches blanches représentent une réponse affirmative à la question posée dans les filtres et les flèches noires correspondent à une réponse négative. À chaque étape du tri, une certaine quantité de données est rejetée : (filtres 1) le bruit, comme *<u>é</u>té* ; (filtre 2) les composés néoclassiques, comme *<u>poly</u>graphe* ; et (filtres 3) les dérivés opaques en synchronie, comme *<u>pré</u>parer*.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100251

Graphique 1 : Procédure de filtrage des données brutes



Pour les filtres 1 et 1bis, nous avons consulté deux dictionnaires monolingues, à savoir le *Petit Robert* (sur CD-ROM) et le *Trésor de la Langue Française informatisé*. Des dictionnaires médicaux en ligne ont également été nécessaires pour classer les données issues du sous-corpus médical de KIAP. Pour le filtre 2, nous nous sommes basés sur l'inventaire des éléments de formation des mots décrit dans la Section 3.1, qui reprend les éléments classés comme quasi-lexèmes dans la littérature sur la formation des mots en français. Pour reprendre l'exemple mentionné ci-dessus, *polygraphe* est rejeté par le filtre 2 car *-graphe* est listé comme quasi-lexème formant des composés néoclassiques dans les ouvrages de référence. En d'autres termes, seuls des mots comme *polydépendance* ou *poly-usage* sont classés comme préfixés. Les filtres 3.1 et 3.2 se sont révélés plus complexes, notamment en raison du caractère non dichotomique de la transparence sémantique. Nous avons été confrontés aux cas suivants, qui représentent environ 10% des données à traiter (90% des préfixés potentiels ont pu être classés sans équivoque comme étant (au moins en partie) transparents ou totalement opaques sémantiquement) :

- Cas n°1 : le préfixé est transparent dans certains de ses sens uniquement (p. ex. *représenter* dans *représenter un examen*). Ces cas ont requis une désambiguisation en contexte à l'aide d'un concordancier et de l'interface d'interrogation du corpus Frantext ;
- Cas n°2 : au vu du fait que les données n'ont pas été étiquetées, des formes comme *informe* ont dû être désambiguisées (adjectif ou forme verbale).

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100251

3.4.2 Critères définitoires du concept de « préfixe »

Les requêtes formelles ont permis d'extraire les dérivés potentiels sur base de notre inventaire des préfixes français. Il ressort de la phase de tri que certains de ces éléments n'apparaissent pas dans les données, et par corollaire, ne remplissent sans doute pas les critères suivants, qui ont été pris comme point de départ de la définition du concept de « préfixe » dans notre étude (p. ex. *a-* (*ab-*, *ac-*, etc.), *ab-*, *anté-*, *arch-*, *circum-*, *exo-*, *intro-*) :

- Les préfixes sont typiquement des morphèmes lexicaux liés. Cependant, un certain nombre de formes grammaticales libres peuvent, suivant un procédé de grammaticalisation, développer le statut de préfixe (voir Amiot, 2004/2005a) ;
- Les préfixes ont un contenu sémantique stable. Leur contenu référentiel est plus faible que celui des formes libres. Les éléments de formation de mots qui correspondent par ailleurs à des formes libres (p. ex. prépositions) sont considérés comme préfixes (1) s'ils ont au moins un sens différent de la forme libre et/ou (2) s'ils désignent un ensemble plus réduit de sens que la forme libre. Nous avons également suivi les critères définitoires proposés par Iacobini (1998) pour distinguer les préfixes des quasi-lexèmes initiaux (sens plus général ou abstrait, correspondance à un préfixe d'un point de vue paradigmatique)[7] ;
- En synchronie, les préfixes s'attachent typiquement à des bases libres ;
- Les préfixes sont utilisés de manière répétitive dans la formation de préfixés (fréquence).

3.5 Classification sémantique des données

Nous adoptons - moyennant quelques adaptations mineures - la classification sémantique des préfixes proposée par Cartoni (2008) (voir également Lefer et Cartoni, 2011). En voici un bref aperçu :

- Préfixes inchoatifs (ou de changement d'état) (p. ex. *en-*);
- Préfixes évaluatifs (p. ex. *mini-*) ;
- Préfixes de position (locative et temporelle) (p. ex. *pré-*) ;
- Préfixes négatifs (contraire/contradiction/privation et réversion[8]) (p. ex. *in-*) ;
- Préfixes quantitatifs (quantité unique et nombrée, pluralité indéterminée, totalité) (p. ex. *multi-*) ;
- Préfixes modaux (réitérativité, réflexivité, union et réciprocité) (p. ex. *co-*) ;
- Préfixes de soutien partisan et d'opposition[9] (p. ex. *pro-*) ;
- Préfixes intensificateurs (p. ex. *dé-*)[10].

Certaines de ces catégories sont divisées en sous-catégories, comme les préfixes de position locative (p. ex. « devant », « derrière », « à côté », « dedans », « dehors », « dessus », « dessous », « au-delà » et « relations hiérarchiques ») ou temporelle (« avant », « après », « nouveauté » et « ancienneté »). Les préfixes évaluatifs, quant à eux, sont « subdivisés en fonction de la valeur positive ou négative qu'ils dénotent » ainsi qu' « en fonction des aspects qualitatif ou de grandeur sur lesquels ils portent » (Cartoni 2008 : 132). La classe des préfixes évaluatifs comprend également les préfixes d'approximation (p. ex. *quasi-*) et les préfixes d'atténuation (p. ex. *semi-*).

Bien sûr, cette catégorisation sémantique présente quelques difficultés. Citons par exemple le fait que certains préfixes appartiennent à plusieurs catégories. C'est le cas du préfixe *extra-*, qui peut être à la fois locatif (p. ex. *extra-atmosphérique*) et évaluatif (p. ex. *extrafrais*) (voir Fradin et Montermini, 2009 sur la morphologie évaluative en français et sur les glissements sémantiques entre les préfixes spatio-temporels

et les préfixes évaluatifs). En outre, certaines catégories sémantiques présentent des frontières extrêmement floues, comme les sous-classes négatives « privation », « contraire » et « contradiction », que nous avons décidé de regrouper dans cette étude (voir Cartoni et Lefer, 2011 sur ce point). Il a donc été nécessaire d'examiner les données contextualisées (sous forme de concordances) afin de désambiguiser certains préfixés et leur attribuer une étiquette sémantique appropriée (catégorie et sous-catégorie).

3.6 Test statistique

Nous nous sommes basés sur le test du chi-carré de Pearson (χ^2) afin de déterminer si les tendances mises au jour dans l'étude sont statistiquement significatives (Field, 2005). Dans notre cas, il s'agit de savoir si la productivité (1) de la préfixation et (2) des diverses (sous-)catégories sémantiques varie significativement à travers les genres écrits et les domaines scientifiques. Pour ce faire, nous avons entré deux types de mesures dans nos tables de contingence : (1) le nombre total de préfixés et le nombre total de lemmes non préfixés dans chaque genre/domaine[11] et (2) le nombre total de préfixés d'une catégorie donnée et le nombre total de préfixés dans chaque genre/domaine[12]. Le seuil de signification de $p \leq 0.05$ a été utilisé. Si la valeur du χ^2 indique que la différence observée entre les genres ou les domaines est significative, les résidus ajustés et leurs p ont été calculés afin de déterminer quel genre/domaine est caractérisé par une productivité significativement plus haute ou plus basse (Everitt, 1992). Les signes suivants sont utilisés afin de faciliter l'interprétation des résultats dans certains des tableaux présentés dans la Section 4[13] :

+	si	$0.01 < p \leq 0.05$	→ la productivité de telle ou telle classe/de tel ou tel préfixe est significativement plus haute dans tel ou tel genre ou domaine
++	si	$0.001 < p \leq 0.01$	
+++	si	$p \leq 0.001$	
-	si	$0.01 < p \leq 0.05$	→ la productivité de telle ou telle classe/de tel ou tel préfixe est significativement plus basse dans tel ou tel genre ou domaine
--	si	$0.001 < p \leq 0.01$	
---	si	$p \leq 0.001$	

4 Résultats

L'approche de corpus adoptée dans cette étude a généré un nombre considérable de résultats, qu'il ne nous est pas possible de présenter de manière exhaustive ici. Nous nous contenterons donc de présenter, d'une part, quelques résultats généraux et, d'autre part, un premier aperçu des tendances révélées par l'approche par catégories sémantiques. Nous nous pencherons également sur la sensibilité au genre de quelques paires ou groupes de préfixes qui appartiennent à la même catégorie sémantique, à savoir *a-*, *in-* et *non-*, *après-* et *post-*, *avant-* et *pré-*, *micro-* et *mini-*, et enfin *macro-* et *super-*. En effet, notre analyse révèle que leur productivité dépend fortement du genre dans lequel ils sont utilisés.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100251

4.1 Aperçu général de la productivité de la préfixation en français

Les résultats montrent clairement que la productivité de la préfixation est hautement sensible à la variation à travers les genres et les domaines, ce qui confirme la première hypothèse abordée dans cette étude. Plus précisément, il ressort que la préfixation est un peu plus productive dans les éditoriaux de presse (723 préfixés sur 10.000 lemmes[14]) que dans les articles scientifiques (654 préfixés sur 10.000 lemmes). C'est dans les romans qu'elle est la moins productive (496 préfixés sur 10.000 lemmes) (χ^2 = 119.51, $p < 0.001$). Ces résultats confirment la productivité réduite de la préfixation dans les romans, mais contrairement à l'une de nos hypothèses de départ, il ressort que la productivité de la préfixation est (légèrement) plus haute dans les éditoriaux de presse que dans la prose académique (cf. Tableau 4).

De même, la productivité de la préfixation est la plus faible en économie, suivie de la linguistique et enfin de la médecine, où elle est la plus haute (voir Tableau 5) (χ^2 = 31.73, $p < 0.001$). Les résultats indiquent que l'économie ne se situe pas entre la linguistique et la médecine, ce qui va à l'encontre des patrons identifiés par Fløttum et al. (2006).

Tableau 4 : Productivité réalisée de la préfixation à travers les genres écrits

	Romans		*Articles scientifiques*		*Éditoriaux*	
	Prod. obs.	Prod. rel.	Prod. obs.	Prod. rel.	Prod. obs.	Prod. rel.
Productivité de la préfixation	1.276 (---)	496	1.656 (+)	654	1.910 (+++)	723
Nombre total de lemmes	25.713	10.000	25.315	10.000	26.426	10.000

Tableau 5 : Productivité réalisée de la préfixation à travers les domaines académiques

	Économie		*Linguistique*		*Médecine*	
	Prod. obs.	Prod. rel.	Prod. obs.	Prod. rel.	Prod. obs.	Prod. rel.
Productivité de la préfixation	585 (---)	539	810	585	727 (+++)	718
Nombre total de lemmes	10.852	10.000	13.843	10.000	10.123	10.000

4.2 Productivité des classes sémantiques à travers les genres et les domaines

4.2.1 Les genres écrits

Le Tableau 6 présente la productivité des différentes catégories et sous-catégories sémantiques dans notre corpus (tous genres et domaines confondus). Il ressort que les préfixes les plus productifs sont les préfixes négatifs (28% des lemmes préfixés), les préfixes de position locative et temporelle (21%), les préfixes évaluatifs (17%) et les préfixes modaux (16%). Dans la suite de cette section, nous nous concentrons sur les divergences de productivité entre les trois genres.

Tableau 6 : Vue d'ensemble de la productivité des classes et sous-classes sémantiques

		Productivité	
Rang	*Classes sémantiques*	*no.*	*%*
1	**Négation**	**923**	**27,78**
	• Contradiction, contraire, privation *(a-, anti-, dé-, dis-, in-, mal-, més- non-, sans-)*	645	
	• Réversion *(dé-, dis-)*	278	
2	**Position**	**707**	**21,28**
	• Position locative	399	
	Dedans (*en-, in-, intra-*)	80	
	Dessous (*hypo-, infra-, sous-, sub-*)	77	
	Au-delà (*méta-, outre-, per-, trans-, ultra-*)	49	
	Dehors (*ex-, extra-*)	36	
	Dessus (*épi-, super-, supra-, sur-, sus-*)	27	
	Derrière (*arrière-, post-, rétro-*)	25	
	Autour (*circon-, péri-*)	20	
	Relations hiérarchiques (*archi-, pro-, sous-, vice-*)	18	
	Devant (*anti-, avant-, pré-*)	17	
	Entre (*entre-, inter-*)	15	
	À distance (*télé-*)	11	
	À côté (*contre-, juxta-, para-*)	9	
	Milieu (*mi-*)	9	
	Relations familiales (*arrière-*)	5	
	De ce côté (*cis-*)	1	
	• Position temporelle	308	
	Ancienneté (*ex-*)	96	
	Avant (*avant-, pré-, rétro-*)	76	
	Après (*après-, arrière-, post-*)	64	
	Nouveauté (*néo-*)	52	
	Milieu (*mi-*)	13	
	Entre (*entre-, inter-*)	4	
	À travers (*per-*)	3	

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100251

3	**Évaluation**	**578**	**17,39**
	• Évaluation positive - grandeur/quantité et qualité *(hyper-, macro-, maxi-, méga-, super-; archi-, bien-, extra-, hyper-, per-, plus-, super-, sur-, ultra-)*	225	
	• Évaluation négative - grandeur/quantité et qualité *(micro-, mini-; dys-, hypo-, mal-, més-, para-, sous-, sub-)*	178	
	• Atténuation *(demi-, entre-, mi-, semi-)*	89	
	• Approximation *(crypto-, pén-, pseudo-, quasi-)*	86	
4	**Modalité**	**531**	**15,98**
	• Réitérativité *(re-)*	331	
	• Union et réciprocité *(co-, entre-, inter-)*	119	
	• Réflexivité *(auto-)*	81	
5	**Soutien partisan et opposition**	**265**	**7,97**
	• Opposition *(anti-, contra-, contre-, para-)*	244	
	• Soutien partisan *(pro-)*	21	
6	**Quantité**	**178**	**5,36**
	• Quantité unique ou nombrée *(bi-, demi-, di-, hémi-, mono-, semi-, uni-)*	86	
	• Pluralité indéterminée *(multi-, pluri-, poly-)*	75	
	• Totalité *(omni-, pan-)*	17	
7	**Inchoativité** (*a-, é-, en-*)	**121**	**3,64**
8	**Intensification** (*dé-, par-, pour-*)	**20**	**0,60**
TOTAL PREFIXATION		3.323	100

Le Tableau 7 reprend les classes sémantiques qui sont significativement plus productives dans un genre particulier par rapport aux deux autres genres. De nombreuses variations sont à mettre en évidence. Les résultats indiquent, par exemple, que (1) les préfixes négatifs, réitératifs et inchoatifs sont plus productifs dans les romans que dans les deux autres genres ; (2) les préfixes de position locative ainsi que les préfixes évaluatifs et quantitatifs sont plus productifs dans les articles scientifiques ; et enfin, (3) les préfixes temporels et les préfixes de soutien partisan et d'opposition sont plus productifs dans les éditoriaux de presse, qui ressortent donc clairement comme des textes d'opinion (cf. Le, 2009). En d'autres termes, la productivité des classes sémantiques varie fortement d'un genre à l'autre[15]. Des exemples contextualisés sont fournis en guise d'illustration dans le Tableau 8.

Tableau 7 : Vue d'ensemble des différences de productivité des classes sémantiques à travers les genres écrits

Classes et sous-classes sémantiques	**Ex. de préfixes**	**Prod. rel. romans**[16]	**Prod. rel. articles scientifiques**	**Prod. rel. éditoriaux**	**χ^2**
*Classes sémantiques plus productives dans le sous-corpus de **romans** que dans les deux autres sous-corpus*					
Négation	*dé-*, *in-*	378	296	321	$\chi^2 = 22{,}51$ $p < 0.001$
Derrière	*arrière-*	10	6	3	$\chi^2 = 6{,}38$ $p < 0.05$
Relations hiérarchiques	*vice-*	8	0	6	$\chi^2 = 11{,}90$ $p < 0.01$
Atténuation	*demi-*, *mi-*	34	12	17	$\chi^2 = 19{,}65$ $p < 0.001$
Réitérativité	*re-*	172	111	118	$\chi^2 = 27{,}71$ $p < 0.001$
Inchoativité	*a-*, *é-*, *en-*	82	32	48	$\chi^2 = 37{,}99$ $p < 0.001$
*Classes sémantiques plus productives dans le sous-corpus d'**articles scientifiques** que dans les deux autres sous-corpus*					
Position (en particulier : position locative, « autour », « dessous », « au-delà », « devant », « dedans », « dehors »)	*extra-*, *péri-*	141	206	181	$\chi^2 = 20{,}71$ $p < 0.001$
Évaluation (en particulier : évaluation négative, approximation)	*hypo-*, *sous-*, *pseudo-*	121	158	138	$\chi^2 = 8{,}30$ $p < 0.05$
Union et réciprocité	*co-*	24	48	34	$\chi^2 = 12{,}56$ $p < 0.01$
Réflexivité	*auto-*	2	27	24	$\chi^2 = 28{,}87$ $p < 0.001$
Quantité (en particulier : pluralité indéterminée)	*mono-*, *multi-*	26	62	38	$\chi^2 = 25{,}27$ $p < 0.001$

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100251

*Classes sémantiques plus productives dans le sous-corpus d'**éditoriaux** que dans les deux autres sous-corpus*					
Position temporelle (en particulier : « après », « nouveauté », « ancienneté »)	*après-*, *ex-*, *néo-*	49	52	113	$\chi^2 = 63{,}49$ $p < 0.001$
Soutien partisan et opposition	*anti-*, *pro-*	39	51	91	$\chi^2 = 40{,}29$ $p < 0.001$

Tableau 8 : Exemples contextualisés

Romans	
Négation	Trop directement dites, les choses se ***dépoétisent***. (Frantext, LANZMANN Jacques, La Horde d'or, 1994, p. 187, Chapter VI) Peu à peu, cela devrait constituer un assez bel ensemble, ***impubliable***, bien sûr, mais destiné à être connu et apprécié des amateurs professionnels. (Frantext, SOLLERS Philippe, Le Secret, 1993, p. 248, III)
Derrière	Dès le lendemain de la Révolution, tout au long des guerres de l'Empire, tu marches sur les généraux de génie qui seraient restés sur leur fumier et dans leurs **arrière-boutiques** si les temps n'avaient pas changé. (Frantext, ORMESSON Jean d', La Douane de mer, 1993, p. 457)
Relations hiérarchiques	Ils avaient roi, reine, ***vice-roi***, drapeaux et cérémonies - et assuré pas peut-être un grand savoir magique. (Frantext, CHAMOISEAU Patrick, Texaco, 1992, p. 75)
Atténuation	Je ne connais rien de plus éprouvant que de balancer entre ce qui vous apparaît, selon l'humeur, tantôt comme une ***demi-satisfaction***, tantôt comme une déconvenue. (Frantext, BAZIN Hervé, L'école des pères, 1991, p. 291, 1982) La meute nous cernait toujours, ***mi-sidérée***, ***mi-menaçante***. (Frantext, ORSENNA Érik, Grand amour, 1993, p. 63)
Réitérativité	Furieux d'avoir laissé la place au silence d'août, les mots ***réaffirmaient*** leur empire, ils s'imposaient, ils s'insinuaient comme des cafards dans une chambre d'hôtel africain, la moindre minute d'emploi du temps se couvrait de toasts, allocutions, exordes... (Frantext, ORSENNA Érik, Grand amour, 1993, p. 151)
Inchoativité	Ou alors... Est-ce que, par hasard, elle aurait ***enlaidi*** sans s'en apercevoir ? (Frantext, DORMANN Geneviève, La Petite main, 1993, p. 172)

Articles scientifiques	
Position locative	Un garçon de 13 ans était hospitalisé pour l'exploration d'une masse abdominale découverte dans les suites d'un coup de pied violent dans la région ***sus-ombilicale***, 15 jours auparavant. (KIAP, frmed09) (…) elle a la responsabilité de répartir cette dotation au niveau ***infra-régional*** entre structures de soins et actions en faveur de la santé publique, selon une extension de la logique contractuelle qui préside déjà aux Contrats d'Objectifs et de Moyens pour l'hospitalisation. (KIAP, frecon03) La fonction sujet est exclue dans le cas de en quantitatif, c'est-à-dire lorsque le quantificateur se trouve en position ***préverbale*** (…). (KIAP, frling06)
Évaluation	Si l'équilibre concurrentiel stationnaire est caractérisé par une situation de ***sous-capitalisation*** par rapport au stock de capital de la règle d'or, la limite de l'économie avec oligopole lorsque le nombre de firmes de ce dernier tend vers l'infini est l'équilibre concurrentiel. (KIAP, frecon43) Il était d'une certaine logique d'inférer que la ***quasi-disparition*** des ventes illicites d'héroïne dans les rues des grandes villes de France était liée à ce déploiement. (KIAP, frmed19)
Union et réciprocité	En fait, la ***co-présence*** pour chaque émission d'un critique plus ou moins " alternatif "(Inrockuptibles), d'un critique " spécialisé " (Cahiers du Cinéma ou Positif), et de deux critiques plus ou moins " tous publics " n'est jamais problématique (…). (KIAP, frling34)
Réflexivité	Il restait à expliquer, du côté de la linguistique, en quoi la langue était à même d'***auto-organiser*** mentalement la signification sans qu'une conscience n'intervienne. (KIAP, frling21)
Quantité	Dans cette série de patients en partie traités avant 1975, 70,6 % des sujets ont reçu une ***bithérapie*** par INH + SM. (KIAP, frmed29) La fréquence des ***poly-consommations*** a également été soulignée. (KIAP, frmed21)
Éditoriaux de presse	
Soutien partisan et opposition	On a ainsi vu défiler pendant la campagne moult faits divers impliquant des maires (y compris de gauche) partis à la chasse des travailleurs polonais implantés dans leur commune, comme aux pires heures des bulldozers ***anti-immigrés*** du PCF. (Mult-Ed, <F-LI-300505-2>) On comprend dans ces conditions que les patrons britanniques, jusqu'ici massivement ***pro-euro***, aient changé d'avis (50% contre, 42% pour selon les derniers sondages). (Mult-Ed, <F-FI-230103-0>)
Position temporelle	Lors de celle de mercredi prochain, l'ultime gadget ***pré-estival***, vendu hier au Chambon-sur-Lignon, occultera aisément l'affaire Euralair. (Mult-ed, <F-LI-090704-1>) Le second attendit l'alternance de l'***après-Chirac*** avec bonne humeur. (Mult-Ed, <F-LI-130307-0>)

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100251

4.2.2 Les domaines académiques

La productivité de la plupart des catégories sémantiques en prose académique est sensible au domaine (voir le Tableau 9, qui reprend les principales catégories). À titre d'exemple, nous pouvons citer les préfixes négatifs (*dé-*, *in-* et *non-*), qui sont beaucoup plus productifs en linguistique qu'en médecine (p. ex. *désémantisé*, *désambiguisation*, *intraduisible*, *non-communicationnel*, *non-marqué*) ($\chi^2 = 59.38$, $p < 0.001$). Il est intéressant de noter que cette tendance peut être reliée à l'observation faite par Fløttum et al. (2006) que la négation syntaxique *ne ... pas* est significativement plus fréquente en linguistique qu'en médecine. Ce contraste vaut donc aussi pour la négation affixale. L'économie, elle, ne se distingue clairement des deux autres domaines que par la productivité des préfixes modaux, en particulier *inter-* (p. ex. *inter-firmes*, *inter-banques*, *inter-entreprises*) et *re-* (p. ex. *réajustement*, *réindexation*, *renormalisation*), qui sont peu productifs en médecine ($\chi^2 = 24.54$, $p < 0.001$).

En outre, les résultats révèlent que la médecine s'écarte très distinctement de la linguistique et de l'économie car les préfixes de position (comme *extra-*, *intra-*, *péri-* et *post-*), les préfixes évaluatifs (comme *dys-*, *hyper-* et *hypo-*) et les préfixes quantitatifs (comme *poly-*) y sont beaucoup plus productifs que dans les deux autres domaines (χ^2 respectifs : $\chi^2 = 38.07$, $p < 0.001$; $\chi^2 = 18.17$, $p < 0.001$; $\chi^2 = 27.68$, $p < 0.001$). Les préfixés médicaux incluent ainsi *dysrégulation*, *hypervascularisé*, *hypopigmenté*, *extravertébral*, *intra-osseux*, *postopératoire* et *poly-toxicomanie*. En fait, il apparaît que certaines des catégories sémantiques initialement identifiées comme typiques des articles scientifiques, comme les préfixes de position locative (voir Section 4.2.1), sont en réalité typiques d'un domaine uniquement (ici, la médecine). Nos résultats montrent donc qu'il est essentiel de prendre en compte la variation à travers les domaines dans les travaux de morphologie basés sur corpus afin de différencier les tendances de la langue académique générale et celles des domaines particuliers (cf. Grabar et Zweigenbaum, 2003).

Tableau 9 : Vue d'ensemble des différences de productivité des principales classes sémantiques à travers les domaines académiques

Classes sémantiques principales	*Économie*		*Linguistique*		*Médecine*	
	Prod. obs.	Prod. rel.	Prod. obs.	Prod. rel.	Prod. obs.	Prod. rel.
Négation	203 (+)	35	309 (+++)	38	150 (---)	21
Position (loc. et temp.)	74 (---)	13	133	16	184 (+++)	25
Modalité	156 (++)	27	193	24	116 (---)	16
Évaluation	74	13	81 (---)	10	126 (+++)	17
Quantité	32	5	20 (---)	2	62 (+++)	9
Total lemmes préfixés	**585**	**100**	**810**	**100**	**727**	**100**

Néanmoins, soulignons qu'un certain nombre de préfixes ont une productivité très stable à travers les domaines. Il s'agit des préfixes *auto-*, *co-*, *multi-*, *pré-*, *pseudo-* et *sous-*. Ces préfixes appartiennent plus que probablement à la langue académique générale et font certainement preuve d'une productivité très similaire dans d'autres domaines que la médecine, l'économie et la linguistique.

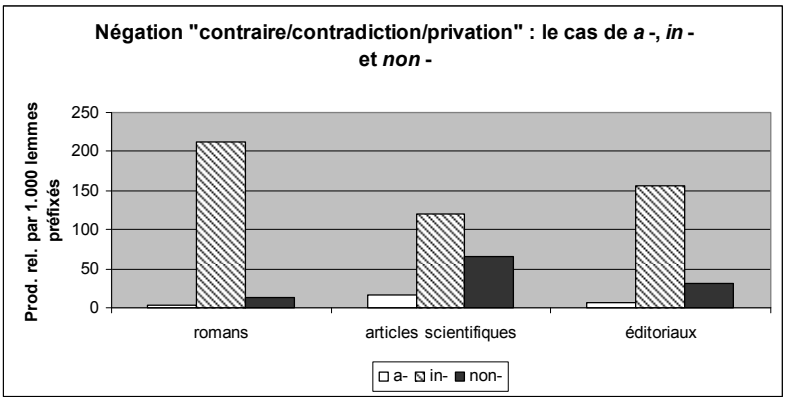
4.3 Sensibilité au genre de paires/groupes de préfixes de la même classe sémantique

Notre approche variationiste a également permis de mettre au jour des tendances intéressantes en ce qui concerne cinq groupes de préfixes appartenant aux mêmes sous-catégories sémantiques, à savoir *a-*, *in-* et

non- (« négation : contraire/contradiction/privation »), *après-* et *post-* (« position temporelle : après »), *avant-* et *pré-* (« position temporelle : avant »), *micro-* et *mini-* (« évaluation négative : grandeur »), et enfin *macro-* et *super-* (« évaluation positive : grandeur »).

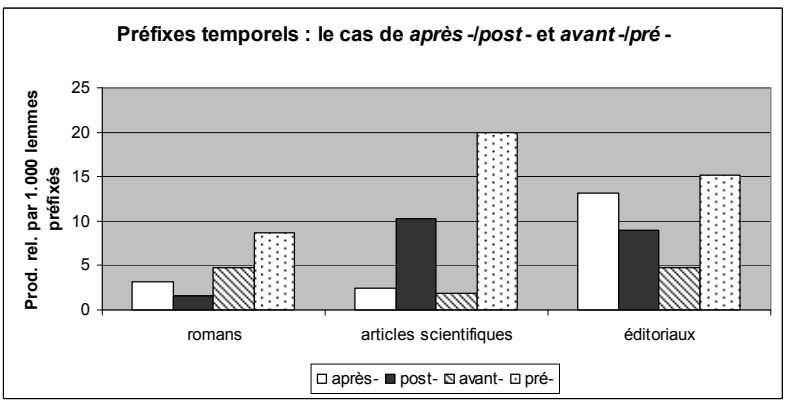
Comme indiqué dans la Section 4.2.1 (Tableau 7), les préfixes négatifs sont particulièrement productifs dans le sous-corpus littéraire et relativement peu productifs en prose académique. Cependant, deux préfixes négatifs, *a-* et *non-*, sont en fait plus productifs en prose académique que dans les deux autres genres (les chiffres de productivité réalisée et les valeurs du χ^2 des préfixes abordés dans cette section sont présentés dans le Tableau 10). Ces contrastes sont représentés dans le Graphique 2, où il apparaît clairement que le préfixe *in-* est largement favorisé dans les romans et que les préfixes *a-* et *non-* sont plus productifs en prose académique que dans les deux autres genres. Les éditoriaux, quant à eux, se situent entre les romans et les articles scientifiques.

Graphique 2 : *a-*, *in-* et *non-* à travers les genres écrits



Pour les paires de préfixes temporels *après-*/*post-* et *avant-*/*pré-*, on remarque, en dépit du fait que les chiffres sont relativement bas (et appelent de ce fait des études basées sur des corpus de plus grande taille), que les préfixes *post-* et *pré-* sont plus productifs dans les articles scientifiques que dans les deux autres genres (voir Tableau 10), au détriment des préfixes *après-* et *avant-* qui sont peu productifs dans ce genre. Dans les éditoriaux, par contre, *après-* est aussi productif que *post-* (voir Graphique 3) (p. ex. *après-11 septembre*, *après-Saddam*). Il est également intéressant de noter que dans les trois genres, *pré-* est plus productif qu'*avant-* (cf. Amiot, 1997 : 74-75).

Graphique 3 : *après-/post-* et *avant-/pré-* à travers les genres écrits



SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100251

Bien que les chiffres pour *micro-*/*mini-* et *macro-*/*super-* soient également relativement bas (voir Tableau 10), une tendance générale semble se dégager. Dans les articles scientifiques, *micro-* est plus productif que *mini-* (p. ex. *micro-ordinateur*, *micro-abcès*, *micro-contexte*, *micro-organisme*, *microdéficit*, *microperforation*) alors que dans les éditoriaux, *mini-* est plus productif (p. ex. *mini-État africain*, *mini-crise*, *mini-révolte*, *mini-Hitler*, *mini-recul*, *mini-micro*). Un patron similaire émerge pour *macro-* et *super-* : *macro-* est plus productif que *super-* en prose académique (il n'y a pas d'occurrence de *super-* dans le corpus KIAP) (p. ex. *macrostructure*, *macro-approche*) alors que *super-* est plus productif que *macro-* dans les éditoriaux (p. ex. *super-école primaire*, *super-État fédéral*, *superprofits*[17]) (voir Tableau 10). Bien sûr, ces premières observations sont de nature très préliminaire et devront être validées à l'aide de corpus de plus grande taille.

Tableau 10 : Groupes de préfixes de la même sous-catégorie sémantique à travers les genres écrits

Préfixes	**Prod. rel. romans**[18]	**Prod. rel. articles scientifiques**	**Prod. rel. éditoriaux**	χ^2
a-	4	16	7	$\chi^2 = 14{,}13$ $p < 0.001$
in-	213	121	158	$\chi^2 = 45{,}96$ $p < 0.001$
non-	13	66	31	$\chi^2 = 59{,}85$ $p < 0.001$
après-	3	2	13	$\chi^2 = 18{,}40$ $p < 0.001$
post-	2	12	9	$\chi^2 = 10{,}25$ $p < 0.01$
avant-	7	4	7	$\chi^2 = 2{,}05$ $p = 0.3589$
pré-	9	25	15	$\chi^2 = 12{,}71$ $p < 0.01$
micro-	5	14	3	$\chi^2 = 17{,}24$ $p < 0.001$
mini-	2	1	8	$\chi^2 = 14{,}36$ $p < 0.001$
macro-	0	6	1	$\chi^2 = 15{,}85$ $p < 0.001$
super-	5	3	7	$\chi^2 = 2{,}60$ $p = 0.2723$

5 Conclusion

Notre étude représente un premier essai de caractérisation de la productivité réalisée de la préfixation française dans son ensemble à travers les genres et les domaines. Les genres sont souvent différenciés et caractérisés sur base de variables morpho-syntaxiques (Malrieu et Rastier, 2001 ; Poudat, 2006) alors que les domaines sont régulièrement repérés à l'aide de variables lexicales (Marshman et al., 2009). Notre étude montre que la préfixation peut être utilisée pour faire ressortir les principales différences à la fois entre les genres et les domaines. La variabilité stylistique de la morphologie, en particulier de la dérivation et de la composition, et son utilisation potentielle pour la discrimination des genres et des domaines ont déjà été suggérées par un certain nombre d'auteurs, notamment en relation avec l'étude du français académique (Clavier, 2006 ; Condamines, 2003). L'étude présentée ici soutient cette suggestion.

D'un point de vue méthodologique, notre étude montre clairement que tout corpus de langue académique est fortement influencé par les domaines qu'il contient (cf. le rôle prépondérant de la médecine) et, plus largement, que toute étude morphologique qui se base sur des corpus se doit de prendre en compte le facteur de variation entre les genres et les domaines.

Bien sûr, il faut garder à l'esprit que notre étude ne s'est penchée que sur un seul aspect de la productivité et se base sur des corpus de taille relativement réduite. Elle devra notamment être élargie à des corpus de langue orale, à d'autres genres écrits et à d'autres domaines.

Références bibliographiques

Amiot, D. (1997). *L'antériorité temporelle dans la préfixation en français*. Villeneuve d'Ascq : Presses Universitaires du Septentrion.

Amiot, D. (2004). Préfixes ou prépositions? Le cas de *sur*(-), *sans*(-), *contre*(-) et les autres. In Corbin, D. (ed.), *La formation des mots: horizons actuels. Lexique 16*, Villeneuve-d'Ascq : Presses universitaires du Septentrion, 67-84.

Amiot, D. (2005a). Between compounding and derivation. Elements of word-formation corresponding to prepositions. In Dressler, W.U., D. Kastovsky, O.E. Pfeiffer et F. Rainer (eds), *Morphology and its demarcations*, Amsterdam : Benjamins, 183-195.

Amiot, D. (2005b). Plusieurs vs *poly*-, *pluri*- et *multi*-. In Flaux, N. et D. Amiot (eds), *La quantification côté déterminants et côté préfixes*, *Verbum*, *27(4)*, 403-417.

Apothéloz, D. (2002). *La Construction du Lexique en Français. Principes de Morphologie Dérivationnelle*. Paris : Ophrys.

Baayen, R. H. (1993). On frequency, transparency and productivity. In Booij, G. et J. Van Marle (eds), *Yearbook of morphology 1992*, Dordrecht : Kluwer, 181-208.

Baayen, R. H. (1994). Derivational productivity and text typology. *Journal of Quantitative Linguistics*, *1*, 16-34.

Baayen, R. H. (2008). Corpus linguistics in morphology: morphological productivity. In Lüdeling, A. et M. Kytö (eds), *Corpus Linguistics. An International Handbook*, Berlin : Mouton de Gruyter, 899-919.

Baayen, R. H. et A. Renouf (1996). Chronicling the Times: Productive Lexical Innovations in an English Newspaper. *Language*, *72(1)*, 69-96.

Bauer, L. (2001). *Morphological Productivity*. Cambridge : Cambridge University Press.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100251

Béchade, H. (1992). *Phonétique et morphologie du français moderne et contemporain*. Paris : Presses Universitaires de France.

Biber, D. (1988). *Variation across speech and writing*. Cambridge : Cambridge University Press.

Biber, D., S. Johansson, G. Leech, S. Conrad et E. Finegan (1999). *Longman Grammar of Spoken and Written English*. London : Longman.

Cartoni, B. (2008). *De l'incomplétude lexicale en traduction automatique : vers une approche morphosémantique multilingue*. Thèse de doctorat. Université de Genève : Genève. Accessible en ligne : <http://archive-ouverte.unige.ch/downloader/vital/pdf/tmp/iq62u1u7vghmcaogpsg9lon7f1/out.pdf>

Cartoni, B. et M.-A. Lefer (2011). Negation and lexical morphology across languages: insights from a trilingual translation corpus. *Poznan Studies in Contemporary Linguistics*, *47(4)*, 795-843.

Chafe, W. (1982). Integration and involvement in speaking, writing, and oral literature. In Tannen, D. (ed.), *Spoken and written language: Exploring orality and literacy*, Norwood/New Jersey : Ablex Publishing Corporation, 35-53.

Chafe, W. L. et J. Danielewicz (1987). Properties of written and spoken language. In Horowitz, R. et J. Samuels (eds), *Comprehending oral and written language*, San Diego : Academic Press, 83-113.

Clavier, V. (2006). Le genre comme point d'accès au document : analyse comparée de textes scientifiques en mécanique et linguistique. Papier présenté à la *Journée de l'ATALA Typologies de textes pour le traitement automatique*, 9 Décembre 2006, Paris.

Condamines, A. (2003). *Sémantique et corpus spécialisés: Constitution de bases de connaissances terminologiques*. Thèse d'habilitation, Université de Toulouse 2.

Corbin, D. (1987). *Morphologie dérivationnelle et structuration du lexique*. Tübingen : Niemeyer.

Dal, G. (2003). Productivité morphologique: définitions et notions connexes. *Langue Française*, *140(1)*, 3-23.

Dubois, J. (1962). *Étude sur la dérivation suffixale en Français moderne et contemporain*. Paris : Larousse.

Dubois, J. et F. Dubois-Charlier (1999). *La dérivation suffixale en français*. Paris : Nathan.

Everitt, B. S. (1992). *The Analysis of Contingency Tables*. London : Chapman & Hall/CRC.

Evert, S. et A. Lüdeling (2001). Measuring morphological productivity: Is automatic preprocessing sufficient? In Rayson, P., A. Wilson, T. McEnery, A. Hardie et S. Khoja (eds), *Proceedings of the Corpus Linguistics 2001 conference*, Lancaster : UCREL, 167-175.

Field, A. P. (2005). *Discovering statistics using SPSS*. London : Sage publications.

Fløttum, K., T. Dahl et T. Kinn (2006). *Academic Voices – across languages and disciplines.* Amsterdam & Philadelphia : Benjamins.

Fradin, B. (2003). *Nouvelles approches en morphologie*. Presses Universitaires de France : Paris.

Fradin, B., N. Hathout et F. Meunier (2003). La suffixation en *–et* et la question de la productivité. *Langue Française*, *140(1)*, 56-78.

Fradin, B. et F. Montermini (2009). La morphologie évaluative. In Fradin, B., F. Kerleroux et M. Plénat (eds), *Aperçus de morphologie du français*, Paris : Presses Universitaires de Vincennes, 231-266.

Gaudin, F. et L. Guespin (2000). *Initiation à la lexicologie française*. Louvain-la-Neuve : Duculot.

Grabar, N. et P. Zweigenbaum (2003). Productivité à travers domaines et genres: dérivés adjectivaux et langue médicale. *Langue Française*, *140*, 102-125.

Grabar, N., D. Tribout, G. Dal, B. Fradin, N. Hathout, S. Lignon, F. Namer, C. Plancq, F. Yvon et P. Zweigenbaum (2006). Productivité quantitative des suffixations par *-ité* et *-able* dans un corpus journalistique moderne. Papier présenté à *TALN 2006* conference, Leuven : 10-13 Avril 2006.

Grevisse, M. (1993). *Le Bon Usage*. Treizième édition révisée par A. Goosse. Paris & Louvain-la-Neuve : De Boeck-Duculot.

Haensch, G. et A. Lallemand-Rietkötter (1972). *Wortbildungslehre des modernen Französisch*. München : Max Hueber Verlag.

Hathout, N., F. Namer, M. Plénat et L. Tanguy (2009). La collecte et l'utilisation des données en morphologie. In Fradin, B., F. Kerleroux et M. Plénat (eds), *Aperçus de morphologie du français*, Saint-Denis : Presses Universitaires de Vincennes.

Huot, H. (2006). *La morphologie. Forme et sens des mots du français*. Paris : Colin.

Iacobini, C. (1998). Distinguishing Derivational Prefixes from Initial Combining Forms. In Booij, G., A. Ralli et S. Scalise (eds), *Proceedings of the First Mediterranean Conference of Morphology*, Patras : University of Patras.

L'Homme, M.-C. (2004). Adjectifs dérivés sémantiques (ADS) dans la structuration des terminologies. In *Actes. Terminologie, ontologie et représentation des connaissances*, Université Jean-Moulin Lyon-3, 22-23 Janvier 2004.

Le, E. (2009). Editorials' genre and media roles: *Le Monde*'s editorials from 1999 to 2001. *Journal of Pragmatics*, *41(9)*, 1727-1748.

Lefer, M.-A. (2009) *Exploring lexical morphology across languages: A corpus-based study of prefixation in English and French writing*. Thèse de doctorat. Université catholique de Louvain: Louvain-la-Neuve.

Lefer, M.-A. et B. Cartoni (2011). Prefixes in contrast. Towards a meaning-based contrastive methodology for lexical morphology. *Languages in Contrast*, *11(1)*, 86-104.

Lehman, A. et F. Martin-Berthet (2007). *Introduction à la lexicologie: sémantique et morphologie*. Paris : Armand Colin.

Lieber, R. (2004). *Morphology and Lexical Semantics*. Cambridge : Cambridge University Press.

Malrieu, D. et F. Rastier (2001). Genres et variations morphosyntaxiques. *Traitement Automatique des Langues*, *42(2)*, 548-577.

Marshman, E., M.-C. L'Homme et V. Surtees (2009). Portability of cause-effect relation markers across specialised domains and text genres: a comparative evaluation. *Corpora*, *3(2)*, 141-172.

Niklas-Salminen, A. (1997). *La lexicologie*. Paris : Armand Colin.

Peytard, J. (1975). *Recherches sur la préfixation en français contemporain*. Paris : Diffusion Honoré Champion.

Plag, I. (1999). *Morphological Productivity. Structural Constraints in English Derivation*. Berlin & New York : Mouton de Gruyter.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100251

Plag, I., C. Dalton-Puffer et H. Baayen (1999). Morphological productivity across speech and writing. *English Language and Linguistics*, *3(2)*, 209-228.

Poudat, C. (2004). Recension et présentation comparative d'étiqueteurs pour le français et l'anglais. *Texto!*, *9(4)*.

Poudat, C. (2006). *Étude contrastive de l'article scientifique dans une perspective d'analyse des genres*. Thèse de doctorat. Université d'Orléans: Orléans. *Texto !*, *11(3/4)*.

Schröder, A. (2008). *On the productivity of Verbal Prefixation in English*. Habilitationsschrift. Halle : Martin-Luther-Universität Halle-Wittenberg.

Thiele, J. (1987). *La Formation des Mots en Français Moderne*. Traduction et Adaptation de A. Clas. Montréal : Les Presses de l'Université de Montréal.

---

[1] Nous remercions Bruno Cartoni et les deux relecteurs anonymes pour leurs commentaires avisés sur la première version de cet article.

[2] Notre étude étant fondée sur un dénombrement des préfixes français, elle repose sur un cadre théorique relativement traditionnel, à savoir celui de la morphologie morphématique.

3 Par « études variationistes », nous entendons les études qui se sont penchées sur la variation à travers les registres, les genres et les domaines.

[4] Le corpus Mult-Ed a été compilé au *Centre for English Corpus Linguistics* de l'Université de Louvain. Voir : <http://www.uclouvain.be/en-cecl-multed.html>

[5] Nous tenons à remercier K. Fløttum (Université de Bergen) de nous avoir donné accès au corpus KIAP.

[6] Dans l'approche de morphologie lexématique (Fradin, 2003), les affixes ne sont que les exposants de règles de construction des mots et n'ont donc pas de sens à proprement parler. En fait, selon ce cadre théorique, le sens est exprimé par les règles de formation des lexèmes, dont les affixes ne sont que les exposants.

[7] Cf. note 6.

[8] Notre sous-division de la classe des préfixes négatifs est sensiblement différente de celle proposée par Cartoni (2008).

[9] Contrairement à Cartoni (2008), nous incluons les préfixes de soutien partisan et d'opposition dans une même classe sémantique et non dans les catégories des préfixes de position locative et des préfixes négatifs, respectivement.

[10] Bien que peu productive, cette catégorie a été ajoutée à la classification de Cartoni (2008) dans un souci d'exhaustivité.

[11] Les corpus ont été lemmatisés à l'aide de *Cordial Analyseur 7*.

[12] Par « nombre total de préfixés », nous entendons le nombre total de préfixés **différents** (cf. notion de « productivité réalisée »).

[13] Il s'agit ici de la valeur p des résidus ajustés.

[14] Pour permettre une lecture plus aisée des résultats, la productivité réalisée des préfixes étudiés ici a été ramenée à une mesure relative (productivité réalisée sur 10.000 lemmes).

[15] La productivité réalisée des (sous-)catégories suivantes ne varie pas à travers les genres : évaluation positive : grandeur/quantité et qualité ; évaluation négative : grandeur/quantité ; position locative : dessus, entre ; position temporelle : avant, milieu ; modalité (lorsque toutes les sous-catégories sont considérées dans leur ensemble) ; quantité : quantité nombrée, totalité ; intensification. Nous n'avons pas pu identifier de tendances pour les (sous-)catégories suivantes, qui sont trop peu fréquentes dans le corpus : position locative : à distance, à côté, relations familiales, milieu, de ce côté ; location temporelle : entre, à travers.

[16] La productivité réalisée des différentes classes et sous-classes sémantiques est ramenée à une mesure relative (productivité réalisée sur 1.000 lemmes préfixés).

[17] Voici les contextes de ces préfixés (ils sont tous utilisés entre guillemets) :

- En dépit des allégations de certains irresponsables, restaurer le collège ne doit pas se résumer à le réduire à une « ***super-école primaire*** ». (Mult-Ed, <F-FI-251102-4>)

- Parions qu'il se trouvera quelques souverainistes aigris pour déceler dans les attendus d'hier une avancée menaçante du « ***super-Etat fédéral*** ». (Mult-Ed, <F-FI-140704-0>)

- L'an dernier, on s'en souvient, les grandes entreprises françaises avaient bouclé les meilleurs résultats de leur histoire : plus de 100 milliards de bénéfices cumulés. Ce qui, au passage, a permis à l'État d'engranger des rentrées fiscales records, preuve s'il en fallait que les « ***superprofits*** » si souvent décriés profitent directement à la collectivité.

[18] La productivité réalisée des différents préfixes est ramenée à une mesure relative (productivité réalisée sur 1.000 lemmes préfixés). Le tableau comprend les chiffres des données désambiguisées pour les préfixes appartenant à plusieurs catégories sémantiques (p. ex. *a-*).