Productivité quantitative des suffixations par *-ité* et *-Able* dans un corpus journalistique moderne

Natalia Grabar[1], Delphine Tribout[2], Georgette Dal[3], Bernard Fradin[2], Nabil Hathout[4], Stéphanie Lignon[4,5], Fiammetta Namer[6], Clément Plancq[2], François Yvon[7], Pierre Zweigenbaum[1,8]

[1]Université Paris Descartes, Faculté de Médecine
Inserm, U729, SPIM
[2]Université Paris 7, CNRS, UMR 7110 LLF
[3]Universités Lille 3 et Lille 1, CNRS, UMR 8163 STL
[4]CNRS & Université Toulouse 2, UMR 5610 ERSS
[5]Université de Haute-Alsace
[6]Université Nancy 2, CNRS, ATILF
[7]GET/ENST, CNRS/LTCI
[8]INaLCO/Paris, DSI, AP-HP

Résumé

Dans ce travail, nous étudions en corpus la productivité quantitative des suffixations par *-Able* et par *-ité* du français, d'abord indépendamment l'une de l'autre, puis lorsqu'elles s'enchaînent dérivationnellement (la suffixation en *-ité* s'applique à des bases en *-Able* dans environ 15 % des cas). Nous estimons la productivité de ces suffixations au moyen de mesures statistiques dont nous suivons l'évolution par rapport à la taille du corpus. Ces deux suffixations sont productives en français moderne : elles forment de nouveaux lexèmes tout au long des corpus étudiés sans qu'on n'observe de saturation, leurs indices de productivité montrent une évolution stable bien qu'étant dépendante des calculs qui leur sont appliqués. On note cependant que, de façon générale, de ces deux suffixations, c'est la suffixation par *-ité* qui est la plus fréquente en corpus journalistique, sauf précisément quand *-ité* s'applique à un adjectif en *-Able*. Étant entendu qu'un adjectif en *-Able* et le nom en *-ité* correspondant expriment la même propriété, ce résultat indique que la complexité de la base est un paramètre à prendre en considération dans la formation du lexique possible.

Mots-clés : morphologie, corpus journalistique, *-Able*, *-ité*, productivité morphologique quantitative.

Abstract

In this paper, we study the quantitative productivity of French suffixes *-Able* and *-ité* in corpora. We first analyze them independently and then when they belong to the same derivational chain (the suffix *-ité* chooses bases with *-Able* in about 15 % of its formations). The productivity of these suffixations is statisticallly estimated in relation to the corpus size. Both these affixes are productive in modern French : they continue to form new lexemes throughout the corpora and are not saturated, their productivity indexes show a stable evolution although this is dependent on applied calculations. Nevertheless, *-ité* is more frequent in the journalistic corpora except when it is applied to adjectives suffixed with *-Able*. Knowing that both *-Able* adjective and its *-ité* noun convey the same property, this finding indicates that the base complexity has to be taken into account in the formation of possible lexicon.

Keywords: morphology, newspaper corpus, *-Able*, *-ité*, quantitative morphological productivity.

1. Introduction

La productivité des règles de construction de lexèmes (désormais RCL) doit être vue sous ses deux aspects complémentaires : la disponibilité de ces règles et leur rentabilité (Corbin, 1987). La *disponibilité*, qui cerne la productivité d'un point de vue qualitatif, définit l'aptitude d'une RCL à former de nouveaux lexèmes, de façon non intentionnelle (pour une discussion, cf. (Dal, 2003)). La disponibilité concerne leur caractère sémantique et, par conséquent, leur capacité de s'appliquer à des bases qui satisfont l'ensemble de contraintes impliquées dans les opérations morphologiques. La *rentabilité* des RCL, qui cerne la notion du point de vue quantitatif, permet de dénombrer l'aptitude effective des RCL à former de nouveaux lexèmes. Diverses méthodes de calcul ont été proposées pour tenter de quantifier cette aptitude. La plupart sont fondées sur des relevés dictionnairiques. À cet égard, (Baayen, 2001) se démarque, puisque les mesures qu'il préconise se calculent en corpus. Notre objectif étant le calcul de la productivité quantitative, nous retenons, dans la suite de ce travail, les propositions de (Baayen, 2001).

Les informations relatives à la productivité des RCL reflètent l'état de la langue des corpus dans lesquels s'effectuent les calculs. Elles permettent par exemple de relever les règles qui sont le plus activement utilisées par les locuteurs. L'implication de ce travail varie en fonction des domaines de la linguistique et des applications : (1) en TAL, parmi les lexèmes « inconnus » la grande majorité sont des lexèmes construits, résultant de l'application des RCL productives. Ces RCL demandent donc d'être décrites d'une manière plus détaillée afin de permettre une analyse morphologique automatique ; (2) en linguistique, l'étude de la productivité en corpus peut permettre de confirmer, d'infirmer ou de nuancer les hypothèses formulées sur les contraintes morphologiques, sémantiques ou autres pesant sur les RCL ; (3) en linguistique de corpus et typologie de textes, des indications sur la productivité permettent de distinguer des types de textes différents en fonction de leurs thématiques, genres, etc. ; (4) en psycholinguistique, il est utile d'avoir des informations sur la productivité des RCL, dans la mesure où il est probable que le traitement en mémoire soit différent selon que les lexèmes sont construits par une RCL productive ou non productive (décomposition, dans certaines conditions (*cf.* (Hay, 2001), ou stockage en bloc respectivement) ; (5) en terminologie, on observe que certaines opérations morphologiques sont plus fréquentes que d'autres, par exemple, les adjectifs dénominaux (*cutané*, *musculaire*, *aortique*) dans le domaine médical (Daille, 1999 ; L'Homme, 2004). L'analyse montre que ces adjectifs sont, dans la majorité des cas, formés sur des bases nominales qui désignent les parties du corps. Du point de vue sémantique, cela veut dire que les infections et les diagnostics sont ancrés par rapport à leurs localisations topographiques (*lymphocytome cutané bénin*, *affection musculaire*, *sténose aortique*). La topographie correspond donc à un paramètre important dans la formation des termes médicaux. Cette constatation implique que, lors de la structuration de termes, la détection de relations de localisation peut s'avérer primordiale. Dans cette tâche, les RCL constituent des indicateurs importants.

Le travail sur la productivité morphologique présenté ici est mené dans le cadre de l'activité de l'action 1 du GDR 2220 du CNRS « Description et modélisation en morphologie »[1]. L'objectif, à terme, consiste à dresser une cartographie de la productivité des RCL du français. Ici, nous nous intéressons plus particulièrement à l'étude des suffixations *-Able*[2] et *-ité* (ou des RCL dont les exposants sont les suffixes). La présentation qualitative de ces deux RCL est faite dans la section

[1] Dont le directeur est Bernard Fradin. L'action 1, « Aspects de la productivité morphologique », est pilotée par Georgette Dal. L'originalité de cette action est de réunir des compétences variées : morphologues théoriciens, talistes, statisticiens et informaticiens.

[2] *-Able* englobe les suffixes *-able*, *-ible* et *-uble*, voir plus loin, section 2.

2. Ces deux suffixes peuvent faire partie de la même chaîne dérivationnelle : *-Able* s'applique aux bases verbales (*calculer*) pour former des adjectifs (*calculable*) ; *-ité* s'applique aux bases adjectivales pour former des noms (*calculabilité*). C'est cette chaîne dérivationnelle qui nous intéresse, avec l'idée de vérifier l'intuition qu'en corpus journalistique, les nominalisations sont des formations morphologiques préférées.

Dans la suite de cet article, nous présentons d'abord les caractéristiques des RCL étudiées, les suffixations par *-Able* et *-ité* (section 2), ainsi que le corpus dans lequel les observations de productivité sont faites (section 3). Nous présentons ensuite les méthodes de calcul de la productivité quantitative en corpus (section 4). Les résultats sont présentés et discutés ensuite (section 5). Nous terminons avec une conclusion et des perspectives (section 6).

2. Présentation des suffixations

Dans la présentation « qualitative » des suffixations étudiées, nous nous appuyons sur les travaux linguistiques antérieurs. Si la suffixation par *-Able* a une riche bibliographie, les travaux autour de *-ité* restent pauvres.

Suffixation en *-Able*. Dans la mesure où les adjectifs suffixés en *-ible* ou *-uble* se distinguent de ceux en *-able* uniquement par le fait qu'ils sont formés sur un radical savant (issu du supin latin cf. (Plénat, 1988)) mais nullement par leur comportement sémantique, nous considérons ces trois suffixes comme des allomorphes et les noterons *-Able*. La suffixation en *-Able* construit des adjectifs à partir de verbes essentiellement {*laver*, *lavable*} et parfois de noms dénotant un statut social {*ministre*, *ministrable*}. Dans le cas des bases verbales, pour que cette RCL puisse s'appliquer, l'argument distingué du verbe, qui correspond au nom recteur de la construction dans laquelle figure l'adjectif dérivé en *-Able*, doit s'interpréter comme un Patient (*calculer la productivité* $\Longrightarrow$ *productivité calculable*, *les denrées périssent* $\Longrightarrow$ *denrées périssables*) ou comme un Site, locatif ou temporel (*naviguer sur la rivière* $\Longrightarrow$ *rivière navigable*, *pêcher pendant cette période* $\Longrightarrow$ *période impêchable*) (Fradin, 2003 ; Hathout *et al*., 2003). Dans le premier cas, le verbe servant de base au dérivé est normalement transitif direct, mais il peut parfois être transitif indirect (*se fier à un procédé* $\Longrightarrow$ *procédé fiable*). Les adjectifs construits en *-Able* expriment l'idée que le référent du N qu'il modifient peut être affecté par l'action dénotée par le V base : *calculable* = 'qui peut être calculé'. Lorsque l'interprétation attestée ne manifeste pas cette idée de potentialité (*remarquable* = 'qu'on remarque'), elle peut néanmoins être dérivée de cette dernière par des mécanismes pragmatiques généraux (Fradin, 2003).

Suffixation en *-ité*. La suffixation par *-ité* forme exclusivement des noms. Elle s'applique essentiellement à des adjectifs (*fragilité*, *obscurité*, *productivité*) mais parfois aussi à des noms (*bouddhéité*, *domesticité*, *matissité*). Néanmoins son application aux bases nominales reste marginale (Corbin, 1987 ; Dal *et al*., 1999). D'un point de vue sémantique, *-ité* construit des noms de propriété dont le sens est globalement paraphrasable par « fait, qualité d'être A » : *productivité* = « fait d'être productif » (Corbin, 1987). Sa spécificité est de présenter les propriétés qu'expriment les noms dérivés comme objectives, ou du moins objectivables cf. *sensibilité* vs. *sensiblerie*. Étant donné la sémantique des noms que forme *-ité*, il opère de façon optimale sur les adjectifs qualificatifs, ceux-ci ayant pour caractéristique d'exprimer une propriété. Toutefois d'un point de vue sémantique la suffixation peut s'appliquer à d'autres types catégoriels de bases à condition que ces dernières puissent exprimer une propriété. Ainsi, quand elle opère sur un nom, celui-ci ne doit pas être considéré comme référant à une entité mais comme référant à une propriété saillante de cette entité (Dal, 1997 ; Dal *et al*., 1999). C'est le cas par exemple

dans *bouddhéité*, dans lequel le nom base *bouddha* ne réfère pas à une entité mais à une manière d'être, une philosophie.

3. Présentation et préparation du corpus journalistique

Dans le travail présenté ici, notre matériel journalistique est constitué des articles parus en 1995 dans le journal *Le Monde* : 47 640 articles, contenant plus de 25 millions d'occurrences. Notre approche combine les traitements automatiques et l'analyse manuelle. Les outils de traitement automatique des langues permettent d'assurer une analyse systématique de gros volumes de données. Nous les utilisons pour le nettoyage et formatage de corpus, leur étiquetage morphosyntaxique et pour la détection de lexèmes construits et leur analyse morphologique. Toutefois, ces outils ne permettent pas d'obtenir des résultats suffisamment précis et fiables (Evert & Lüdeling, 2001). Les propositions de l'analyse morphologique sont donc validées manuellement. Les principaux efforts du traitement automatique sont concentrés autour de l'étiquetage morphosyntaxique, car c'est à ce niveau que se trouvent les informations fondamentales pour l'analyse morphologique. Nous effectuons ainsi un nettoyage du format du matériel source et la ventilation des articles en fonction des rubriques du *Monde*. Le tableau 1 présente les corpus ainsi obtenus. Les deux premières colonnes notent et explicitent les rubriques utilisées, les deux suivantes indiquent le nombre d'articles et d'occurrences par rubrique. La dernière colonne *occ-ponc* correspond au nombre d'occurrences sans les signes de ponctuation. C'est par rapport aux occurrences autres que la ponctuation que les prélèvements sont faits.

	Rubrique	nbArt	occ.	occ-ponc.
AGE	Agenda	1 213	605 432	490 663
ART	Événements culturels	4 242	2 154 164	1 801 044
FRA	France	6 331	3 870 389	2 704 350
INT	International	9 276	3 661 400	3 065 884
LIV	Livres	1 949	1 624 924	1 350 540
RTV	Programme TV et radio	1 217	883 060	718 586
SOC	Société	4 009	2 020 013	1 678 573
SPO	Événements sportifs	2 362	1 177 669	894 648
TOT	8 rubriques	30 599	15 997 051	12 704 286

Tableau 1. Taille des corpus par rubrique du Monde en 1995

Afin d'améliorer la qualité de l'étiquetage de TREETAGGER (Schmid, 1994), nous effectuons un préétiquetage avec un lexique généraliste du français. Les sorties de TREETAGGER sont ensuite traitées avec le lemmatiseur du français FLEMM (Namer, 2000) pour corriger certaines des erreurs et pour proposer les lemmes non reconnus par TREETAGGER. Nous appliquons ensuite l'analyseur morphologique DERIF (Namer, 2002) qui, sur la base de l'étiquetage morphosyntaxique, de règles de construction de lexèmes en français et de listes d'exceptions, détecte les lexèmes construits et propose une analyse morphologique de ces derniers. Les résultats de l'analyse morphologique des corpus sont validés manuellement. La validation manuelle vise à éliminer un certain nombre de lexèmes parasites, étant donné l'objectif visé, par exemple : (1) Lexèmes dans lesquels l'affixation étudiée n'est pas la dernière opération constructionnelle : *dissemblable*, *irrésistible*, *coresponsable* ; (2) Lexèmes difficilement analysables

comme construits en français : *affable*, *impeccable*, *alacrité* ; (3) Lexèmes comportant une suite graphique accidentellement identique aux suffixes étudiés : *faible*, *double*, *deshérité* ; (4) Lexèmes polysémiques : le sens attesté n'est probablement pas celui qui nous intéresse : *impayable*, *majorité* ; (5) Des erreurs et fautes d'orthographe.

Les adjectifs en *-Able* et les noms en *-ité* validés constituent notre matériel de travail.

4. Méthodes de calcul de la productivité en corpus

Dans ce travail, nous utilisons les mesures de productivité proposées dans (Baayen, 2001). Les notations suivantes sont employées :

$\mathcal{C}$	Un corpus
N	Nombre d'occurrences, ou nombre total de lemmes, dans $\mathcal{C}$
N_c	Nombre d'occurrences de la catégorie c dans $\mathcal{C}$
$V(N)$	Nombre total de types, ou de lemmes distincts, dans $\mathcal{C}$
$V(m, N)$	Nombre de types apparaissant m fois dans $\mathcal{C}$
$V(1, N)$	Nombre d'hapax, ou de lemmes qui occurrent une seule fois, dans $\mathcal{C}$

Les mesures de productivité morphologique de (Baayen, 2001) sont basées sur la *potentialité*, dérivée d'une estimation du nombre total d'éléments dans le vocabulaire des lexèmes construits ; cette estimation est éventuellement ramenée au nombre de types de la catégorie c, par exemple les formations en *-Able*. Deux indices de productivité sont proposés : $\mathcal{P}$ et $\mathcal{P}^*$.

L'indice de productivité $\mathcal{P} = \frac{E(V(1,N\times p_c)|Z_c...)}{N\times p_c}$ reflète le rythme de croissance du vocabulaire de catégorie c (p_c est la probabilité de la catégorie c, Z_c un des paramètres de la distribution des types dans la catégorie c). $\mathcal{P}$ est donc assimilé à la probabilité de tirer après $N \times p_c$ tirages un type nouveau, sachant que ce nouveau type est de catégorie c. (Baayen, 2001) l'appelle *category conditioned degree of productivity* ou *category internal growth rate*. L'indice $\mathcal{P}^* = \frac{E(V(1,N\times p_c)|Z_c...)}{E(V(1,N))}$ estime la probabilité de tirer un lexème de catégorie c, sachant que c'est un lexème nouveau. Cette mesure est appelée l'*hapax-conditioned degree of productivity*. $\mathcal{P}^*$ est une mesure inconditionnelle de productivité, qui estime la probabilité que le prochain hapax appartienne à la catégorie c. Dans la pratique, $\mathcal{P}$ et $\mathcal{P}^*$ sont estimés par :

$$\mathcal{P} = \frac{V(1, N_c)}{N_c}; \mathcal{P}^* = \frac{V(1, N_c)}{V(1, N)}$$

où $V(1, N_c)$, N_c et $V(1, N_c)$ ont été validés manuellement, tandis que $V(1, N)$, qui correspond au nombre d'hapax total à chaque prélèvement, n'est pas validé.

Dans ce travail, nous appliquons ces deux mesures qui sont considérées comme complémentaires par leur auteur (Baayen, 2001, 158). Nous suivons leur évolution en fonction de la taille du corpus, en mesurant leur valeur toutes les n occurrences (*occ-ponc* dans le tab. 1). Dans les expériences présentées ici, n est fixé à 10 000.

5. Présentation et discussion des résultats

Nous présentons ici les résultats du calcul de la productivité quantitative des deux suffixations analysées : la RCL_{Able} qui s'applique essentiellement aux bases verbales et forme des adjectifs et la RCL_{ite} qui s'applique aux bases adjectivales, et en particulier aux bases adjectivales en *-Able*, pour former des noms. Pour avoir une vue plus complète du comportement de ces suffixations, nous examinons d'abord la croissance du vocabulaire qu'elles forment (section 5.1).

Nous analysons ensuite leurs indices de productivité $\mathcal{P}$ et $\mathcal{P}^*$ (section 5.2). Pour rappel, l'indice $\mathcal{P}$ est dépendant de la taille du corpus. Comme son dénominateur correspond au nombre total d'occurrences N_C d'une RCL, plus un corpus est grand plus N_C sera important et $\mathcal{P}$ petit. Dans la même logique, les valeurs de $\mathcal{P}$ décroissent au fur et à mesure de la lecture du corpus. Quant à l'indice $\mathcal{P}^*$, il est indirectement dépendant de la taille du corpus, mais dépendant du nombre total d'hapax dans ce corpus, tous lexèmes confondus. Le dénominateur de $\mathcal{P}^*$ correspond au nombre des hapax d'un corpus. Si le vocabulaire d'une RCL est enrichi avec le même rythme que le vocabulaire du corpus entier, les valeurs de $\mathcal{P}^*$ restent stables. La RCL est alors considérée comme productive. Par contre, si la croissance du vocabulaire d'une RCL est moins importante que la croissance du vocabulaire du corpus en général, les valeurs de $\mathcal{P}^*$ vont aller en diminuant et on aura tendance à considérer la RCL comme non productive.

5.1. Croissance du vocabulaire

La figure 1 présente la croissance du vocabulaire pour les suffixations étudiées. Nous pouvons observer, en particulier sur la figure 1(a), où les trois courbes sont mises en parallèle, que la croissance des vocabulaires dans les articles du *Monde* 1995 diffère dans les trois cas. La suffixation par *-ité* présente ainsi les effectifs les plus importants (pas loin de 600 types) et sa courbe continue de croître même à la fin du corpus. La suffixation par *-Able* atteint 300 types et *-Abilité* dépasse 100 types. La forme des deux dernières courbes s'aplatit, mais continue de croître. Cet aplatissement graphique est dû en grande partie à l'échelle de la figure : une projection sur une échelle logarithmique écraserait beaucoup moins les courbes *-Able* et *-Abilité*. Ceci veut dire que nous pouvons considérer ces suffixations comme productives, et les figures suivantes vont le montrer plus clairement.

Les trois figures, 1(b), 1(c) et 1(d), montrent que l'évolution de chaque cas est différent selon les rubriques. Ainsi, pour *-Able*, figure 1(b), c'est la rubrique SOC qui contient le plus de types. Viennent ensuite FRA, ART, LIV et INT. Ces rubriques sont donc les plus productives en adjectifs qui signifient des propriétés possibles. Notons que SOC et LIV sont les plus petites de ces rubriques. SOC a donc vraiment un comportement remarquable dans l'utilisation des lexèmes en *-Able*. En ce qui concerne la suffixation par *-ité*, figure 1(c), c'est la rubrique LIV qui rassemble le plus de lexèmes. Cette rubrique se distingue assez nettement par la fréquence avec laquelle cette suffixation y est employée. La rubrique ART la suit d'assez près, de même que INT, FRA et SOC. Et enfin les noms en *-Abilité*, figure 1(d), sont les plus fréquents dans INT. Viennent ensuite LIV, FRA et SOC. Nous pouvons donc noter que globalement, chaque rubrique a des préférences quant à l'utilisation des RCL pour la formation de lexèmes. *-Able* est préféré dans SOC, *-ité* dans LIV et *-Abilité* se distingue dans INT. De manière générale, ces courbes ne marquent pas de tendance vers une asymptote. Leur forme indique donc que chaque suffixation continue de former de nouveaux lexèmes et peut être considérée comme productive.

Dans les rubriques étudiées, la RCL_{ite} opère surtout sur les bases adjectivales (97 %). Les bases nominales représentent seulement 1,5 % ({*domestique*, *domesticité*}, {*vassal*, *vassalité*}, {*dieu*, *déité*}, {*islam*, *islamité*}). Cela confirme les caractéristiques catégorielles du suffixe présentées dans la section 2. Les 1,5 % restant correspondent à des cas discutables. Parmi les bases adjectivales, 45 % sont des adjectifs non construits ({*rapide*, *rapidité*}, {*opaque*, *opacité*}, {*tenace*, *tenacité*}), les bases adjectivales construites en *-Able* occupant la deuxième place avec 15 % d'effectifs. Les bases en *-Able* sont donc les plus fréquentes parmi les bases construites, où nous trouvons également des bases en *-al* ({*frontal*, *frontalité*}), *-ique* ({*scientifique*, *scientificité*}), *-eux* ({*verbeux*, *verbosité*}). D'autres types de bases sont possibles, mais dans une moindre

proportion (moins de 6 %).

Il est intéressant de remarquer que les noms en *-Abilité* sont dans la majorité des cas moins fréquents que leurs bases en *-Able* dans le corpus *Le Monde* 1995. La complexité morphotactique des lexèmes (*-Abilité* étant plus complexe que *-ité*) doit y jouer un rôle. On peut trouver cependant quelques exceptions dans *Le Monde* 1995 : *faisabilité* est ainsi plus fréquent que *faisable* et *flexibilité* est plus fréquent que *flexible*. Il en est de même pour *crédibilité*, *brevetabilité*, *inviolabilité*, *inéligibilité*, *manoeuvrabilité*, *opérabilité*, *rentabilité*. Par ailleurs, quasiment toutes les bases en *-Able* apparaissent dans le corpus. Seulement cinq noms en *-Abilité* n'y ont pas de base correspondante : *digestibilité*, *figurabilité*, *immutabilité*, *mutabilité* et *traçabilité*. Le comportement de *-ité* est en cela très différent de celui de *in-* qui opère également sur les bases en *-Able* : les bases en *-Able* des *inXAble* étant souvent absentes des corpus (Dal *et al*., 2006).

5.2. Indices de productivité $\mathcal{P}$ et $\mathcal{P}^*$

Les figures 2 présentent les indices de productivité $\mathcal{P}$ et $\mathcal{P}^*$ des suffixations *-Able* et *-ité* dans chaque rubrique étudiée. Les courbes $\mathcal{P}$ évoluent de manière similaire dans les trois cas (figure 2(a), 2(c) et 2(e)). Elles forment un faisceau orienté vers le bas. Nous l'avons dit : une telle évolution de $\mathcal{P}$ est attendue, car avec l'accroissement du nombre d'occurrences les valeurs de $\mathcal{P}$ diminuent. Quant aux courbes $\mathcal{P}^*$ (figure 2(b), 2(d) et 2(f)), leur position initiale sur les graphiques dépend du nombre des hapax, et donc des effectifs, de chaque suffixation. La suffixation par *-ité* (figure 2(d)), étant la plus riche en effectifs, reçoit les valeurs $\mathcal{P}^*$ les plus élevées, les lexèmes en *-Abilité* (figure 2(f)) sont les moins nombreux et la suffixation par *-Able* (figure 2(b)) se trouve entre les deux. Par contre leur évolution montre la constance avec laquelle les trois catégories de lexèmes continuent de croître. On voit ainsi que les valeurs de $\mathcal{P}^*$ restent stables tout au long des corpus. Ces suffixations continuent donc d'enrichir le vocabulaire des rubriques d'une manière comparable à l'enrichissement global des vocabulaires de ces rubriques. Notons néanmoins que les courbes de la RCL_{ite} vont déclinant. Notons aussi que, pour l'indice $\mathcal{P}^*$, ces suffixations doivent faire face à l'ensemble des hapax dans les rubriques, ces hapax étant alimentés par différentes sources : formations morphologiques uniques, noms propres, coquilles et fautes d'orthographe, etc. La première catégorie, formations morphologiques, est certainement la plus pertinente pour pondérer l'apport d'une RCL, mais elle n'est pas distinguée au sein de l'ensemble des hapax. Malheureusement parmi ces hapax, ce sont les noms propres et les coquilles qui sont de loin les plus nombreux. Il serait intéressant, dans le futur, d'évaluer les mesures $\mathcal{P}$ et $\mathcal{P}^*$ uniquement par rapport aux formations morphologiques pertinentes. Dans l'étude actuelle, nous pouvons néanmoins considérer que les suffixations étudiées ici sont productives car elles continuent de former de nouveaux lexèmes et leurs indices de productivité résistent à l'augmentation des nombres d'occurrences et d'hapax.

6. Conclusion et perspectives

Nous avons analysé en corpus journalistique les RCL en *-Able* et *-ité* indépendamment l'une de l'autre et lorsqu'elles constituent une chaîne dérivationnelle. Cette étude a montré qu'elles peuvent être considérées comme productives dans le corpus analysé. Elle a montré aussi que la suffixation par *-ité* a une prédilection pour les bases adjectivales simples et que, parmi les bases adjectivales construites, les bases en *-Able* sont les plus fréquentes. Même si les résultats restent dépendants des corpus étudiés, nous nous attendons à ce qu'ils soient généralisables à

d'autres corpus du français moderne. Nous avons aussi remarqué qu'entre les nominalisations (*-ité*) et les adjectivations (*-Able*) étudiées, ce sont les premières qui sont plus fréquentes. Or, étant entendu qu'un adjectif en *-Able* et le nom en *-ité* correspondant expriment la même propriété, l'un à la manière d'un adjectif, l'autre à la manière d'un nom (selon (Croft, 1991), ils relèvent de la même classe sémantique mais ont des fonctions pragmatiques différentes), ce résultat indique que la complexité de la base est un paramètre à prendre en considération dans la formation du lexique possible. Le fait que les nominalisations soient plus fréquentes est un constat ui demande à être appuyée par l'étude d'autres RCL permettant de former des noms et des adjectifs. Il serait également intéressant d'observer les rapports entre ces catégories syntaxiques dans d'autres corpus. Les corpus non journalistiques présentent ainsi une complémentarité intéressante : corpus scientifiques, en particulier médicaux, et corpus oraux. Rappelons aussi que dans *Le Monde* 1995, nous n'avons pas étudié l'ensemble des rubriques. Selon le tab. 1, il reste environ 10 M d'occurrences distribuées entre une vingtaine d'autres rubriques mineures de l'année. Pour construire un tableau évolutif des affixes du français, d'autres années du *Monde* seront étudiées. Par ailleurs, ce travail permet de mettre en regard les deux aspects de la productivité des affixes : quantitatif et qualitatif. Nous comptons de cette manière compléter la description des affixations qui restent peu ou pas étudiées.

Références

BAAYEN H. (2001). *Word frequency distributions*, volume 18 of *Text, Speech and Language Technology*. Dordrecht, The Netherlands : Kluwer Academic Publishers.

CORBIN D. (1987). *Morphologie dérivationnelle et structuration du lexique*, volume 1. Lille : Presse universitaire de Lille.

CROFT W. (1991). *Syntactic categories and grammatical relations : the cognitive organization of the information*. Chicago/London : The University of Chicago Press.

DAILLE B. (1999). Identification des adjectifs relationnels en corpus. In P. AMSILI, Ed., *Traitement Automatique des Langues Naturelles (TALN)*, p. 105–114, Cargèse.

DAL G. (1997). Du principe d'unicité catégorielle au principe d'unicité sémantique : incidence sur la formalisation du lexique construit morphologiquement. *Bulag (numéro spécial)*, p. 105–115.

DAL G. (2003). Productivité morphologique : définitions et notions connexes. *Langue française* 140, 3–23.

DAL G., GRABAR N., LIGNON S., YVON F., TRIBOUT D. & PLANCQ C. (2006). Les adjectifs en inXable en français. In *Journées Romanes sur la négation*, Toulouse.

DAL G., NAMER F. & HATHOUT N. (1999). Construire un lexique dérivationnel : théorie et réalisations. In P. AMSILI, Ed., *Traitement Automatique des Langues Naturelles (TALN)*, p. 115–124, Cargèse.

EVERT S. & LÜDELING A. (2001). Measuring morphological productivity : Is automatic preprocessing sufficient ? In P. RAYSON, A. WILSON, T. MCENERY, A. HARDIE & S. KHOJA, Eds., *Proceedings of the Corpus Linguistics 2001 conference*, p. 167 – 175, Lancaster.

FRADIN B. (2003). *Nouvelles approches en morphologie*. Paris : Presses universitaires de France (PUF).

HATHOUT N., PLÉNAT M. & TANGUY L. (2003). Enquête sur les dérivés en -able. *Cahiers de Grammaire* 28, 49–90.

HAY J. (2001). Lexical frequency in morphology : is everything relative ? *Linguistics* 39(6), 1041–1070.

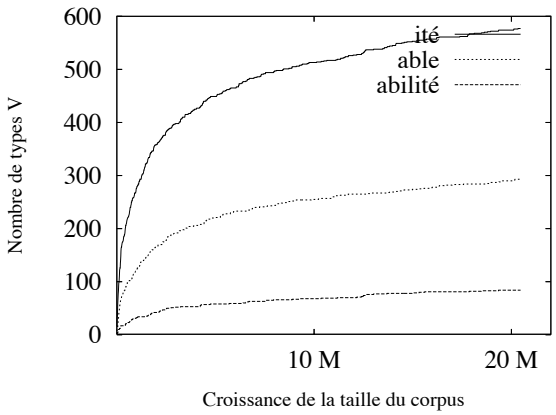
L'HOMME M.-C. (2004). Adjectifs dérivés sémantiques (ADS) dans la structuration des terminologies. In *Journées d'étude Terminologie, Ontologie et représentation des connaissances*, Lyon.

NAMER F. (2000). FLEMM : un analyseur flexionnel du français à base de règles. *Traitement Automatique des Langues (TAL)* 41(2), 523–547.

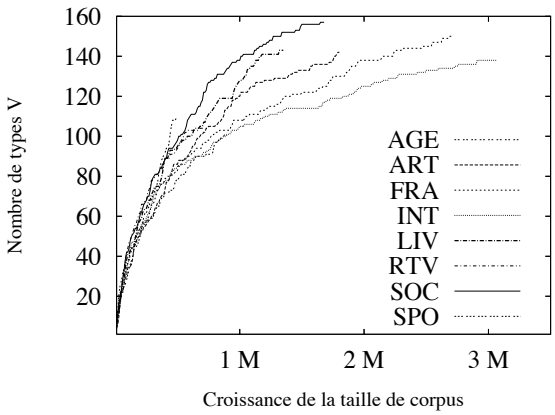
NAMER F. (2002). Acquisition automatique de sens à partir d'opérations morphologiques en français : étude de cas. In *Traitement Automatique de la Langue Naturelle (TALN)*, p. 235–244, Nancy.

PLÉNAT M. (1988). Morphologie des adjectifs en -able. *Cahiers de grammaire* 13, 101–132.

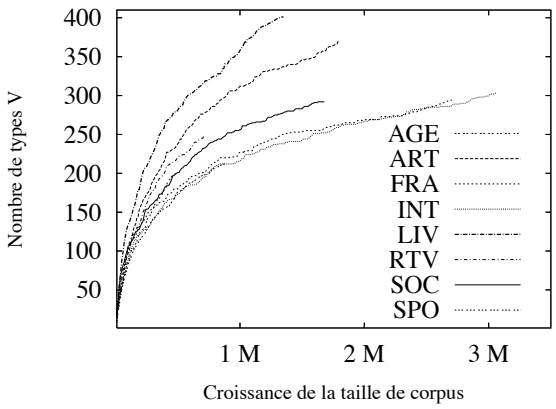
SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, p. 44–49, Manchester, UK.



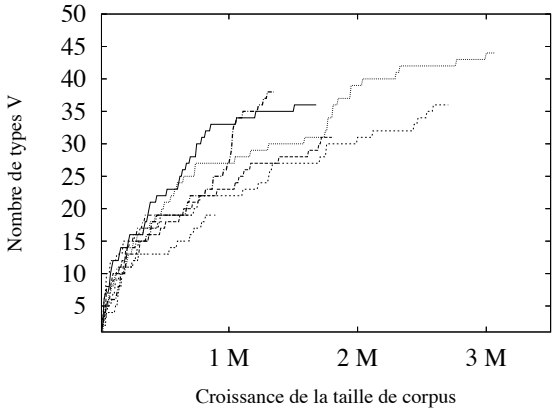
(a) Croissance du vocabulaire avec le suffixe *-ité*, indépendamment des bases.



(b) Croissance du vocabulaire avec le suffixe *-Able*.

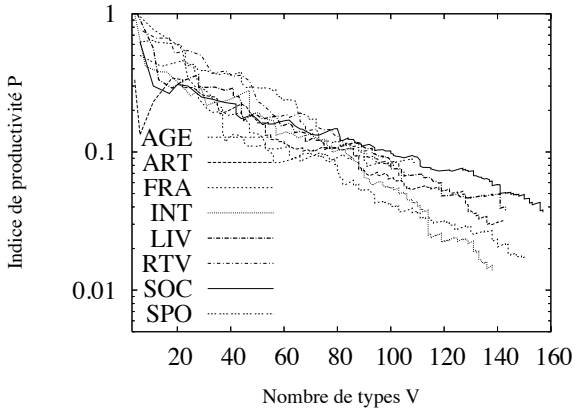


(c) Croissance du vocabulaire avec le suffixe *-ité*, indépendamment des bases.

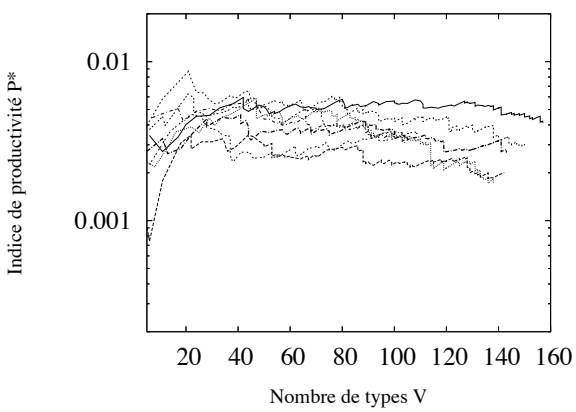


(d) Croissance du vocabulaire avec le suffixe *-ité* lorsqu'il est appliqué aux bases en *-Able*.

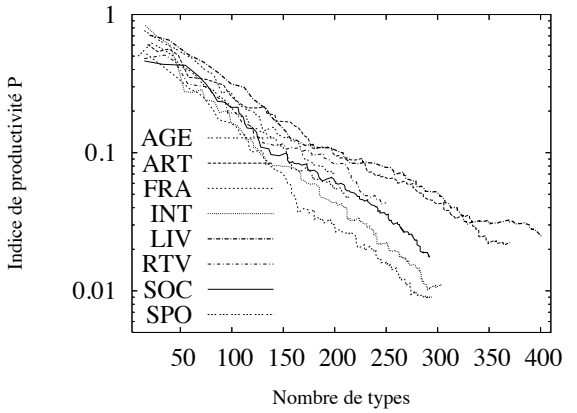
Figure 1. Croissance des vocabulaires de -Able, -ité *et* -Abilité *en fonction du nombre d'occurrences dans l'ensemble des rubriques du Monde 1995*



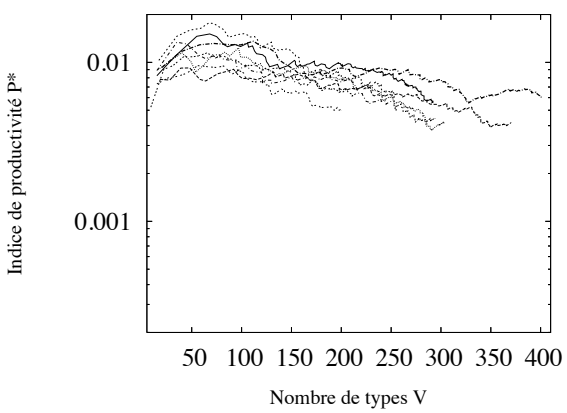
(a) Indice de productivité $\mathcal{P}$ de *-Able*.



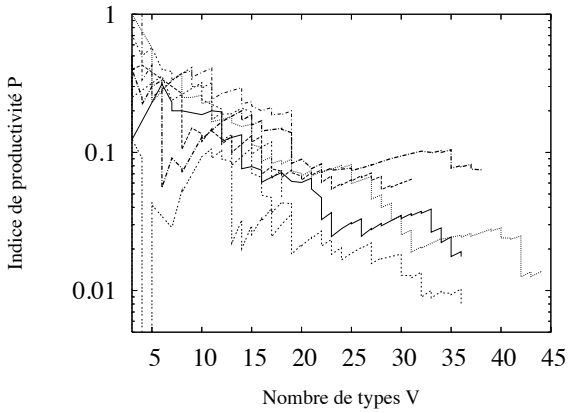
(b) Indice de productivité $\mathcal{P}^*$ de *-Able*.



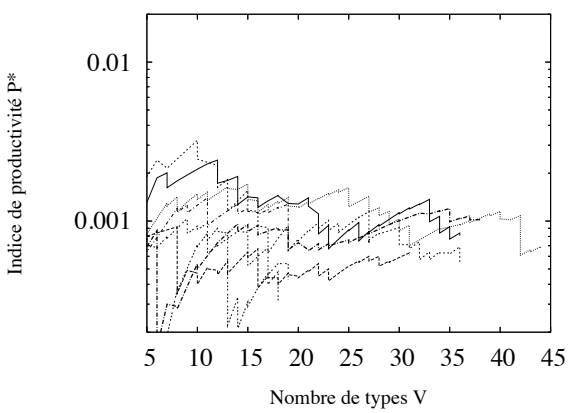
(c) Indice de productivité $\mathcal{P}$ de *-ité*.



(d) Indice de productivité $\mathcal{P}^*$ de *-ité*.



(e) Indice de productivité $\mathcal{P}$ de *-Abilité*.



(f) Indice de productivité $\mathcal{P}^*$ de *-Abilité*.

Figure 2. Indices de productivité $\mathcal{P}$ et $\mathcal{P}^*$ des catégories -Able, -ité *et* -Abilité *dans les rubriques du Monde 1995*