Productivité quantitative de la suffixation par *-Able* dans un corpus journalistique du français

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

[1]Université Paris Descartes, Faculté de Médecine, Inserm U729, Paris 75006 France ;
[2] UMR 8163 STL, CNRS, Universités Lille 3 et Lille 1 ;
[3]UMR 7110 LLF, CNRS, Université Paris 7 ;
[4]UMR 5610 ERSS, CNRS et Université Toulouse 2 ;
[5]Université de Mulhouse ;
[6]ATILF, CNRS, Université de Nancy 2 ;
[7]GET/ENST, CNRS/LTCI, Paris ;
[8]INaLCO/Paris, DSI, AP-HP

Abstract

In this paper, we present our work on quantitative productivity of morphological units of contemporary French. We study especially the suffix *-Able* as it occurs in the journalistic corpus *Le Monde* during 1995. We combine the usage of automatic tools and of human analysis and we point out their complementary nature, especially when working on textual data. We assess as well the importance of *a priori* decisions on the material. We discuss the differences observed with this suffix in different studied corpora and bring out its productive and non productive zones in *Le Monde* articles during 1995.

Résumé

Dans cet article, nous présentons les travaux effectués autour de l'étude quantitative de la productivité des procédés morphologiques du français contemporain. Nous nous consacrons en particulier à la suffixation par *-Able* telle qu'elle apparaît dans le corpus journalistique *Le Monde* de l'année 1995. Dans ce travail, nous combinons les approches automatique et manuelle et nous montrons leur complémentarité, car chacune a un rôle à assurer dans le travail avec des données textuelles. Nous montrons également l'influence effectuée par les décisions *a priori* sur le matériel. Nous discutons les différences de comportement de la suffixation par *-Able* selon les corpus et dégageons ses zones de productivité et de non productivité dans les articles du *Monde* en 1995.

Mots-clés : Productivité quantitative, morphologie, corpus journalistique

1. Introduction

Nous présentons dans ce papier les premiers résultats quantitatifs portant sur le travail sur la productivité morphologique mené par l'action 1 du GdR 2220 du CNRS « Description et modélisation en morphologie »[i] . Nous nous donnons pour objectif d'observer en corpus les occurrences des lexèmes porteurs du suffixe *-Able* et de tirer des conclusions quant à la productivité de la règle à laquelle il est associé en français journalistique contemporain. Les résultats du travail sur la productivité des règles de construction de lexèmes (désormais, RCL) sont, en soi, des indications quant à l'état actuel de la langue. Ils permettent par exemple de

[i]Le directeur du GdR 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.

relever les règles qui sont le plus activement et avec le plus de constance utilisées par les locuteurs. Les implications de ce travail peuvent varier en fonction des objectifs : (1) en TAL, le nombre de lexèmes « inconnus » qui peuvent nécessiter une analyse morphologique sont plus nombreux pour les règles productives. Ces dernières demandent de ce fait à être décrites d'une manière plus détaillée ; (2) en linguistique, ces données permettent par exemple d'observer les contraintes morphologiques, sémantiques ou autres lors de l'application des règles observées ; (3) en linguistique de corpus et typologie de textes, de telles indications permettent de distinguer les 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 certains psycholinguistes font l'hypothèse que le traitement en mémoire est différent selon que les lexèmes construits résultent d'une RCL productive ou non productive (décomposition, dans certaines conditions (selon Hay (2001), dans le premier cas, les lexèmes sont traités par décomposition; dans le second, ils sont stockés en bloc).

La notion de productivité peut être abordée sous un angle qualitatif ou quantitatif. D'un point de vue qualitatif, on s'accorde à définir la productivité comme l'aptitude d'une RCL à former de nouveaux lexèmes, de façon non intentionnelle (pour une discussion, cf. (Dal, 2003)). D'un point de vue quantitatif, diverses méthodes de calcul ont été proposées pour tenter de quantifier cette aptitude. La plupart sont fondées sur des relevés dictionnairiques. A cet égard, (Baayen, 2001) se démarque, puisque les mesures qu'il préconise se calculent en corpus. C'est, dans la suite de ce travail, ses propositions que nous retiendrons. Nous y présentons d'abord les méthodes de calcul de la productivité en corpus (sec. 2) et la préparation du matériel requis (sec. 3). Nous donnons une description linguistique de la suffixation *-Able* (sec. 4). Nous discutons ensuite les résultats obtenus sur la productivité de cette règle (sec. 5). Nous terminons avec une conclusion et des perspectives (sec. 6).

2. Méthodes du calcul de la productivité

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

C	Un corpus
N	Nombre d'occurrences dans C
$V(N)$	Nombre total de types dans C
$V(m,N)$	Nombre de types apparaissant m fois dans C
$V(1,N)$	Nombre de hapax dans 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 : P et P^*.

L'indice de productivité P 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). P est donc assimilé à la probabilité de tirer après N^*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 P^* 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*. P^* est une mesure inconditionnelle de productivité, qui estime la probabilité que le prochain mot soit un hapax de catégorie c. Dans la pratique, on calcule :

$$P = \frac{V(1,Nc)}{Nc} \text{ et } P^* = \frac{V(1,Nc)}{V(1,N)}$$

Ces deux mesures sont complémentaires (Baayen, 2001) :

« In sum, category conditioned degree of productivity provides insight into the long run potentiality of a morphological category, while hapax conditioned degree of productivity measures its short term immediate productivity. »

Dans ce travail, nous appliquons les deux mesures. Nous suivons leur évolution en fonction de la taille du corpus, en mesurant leur valeur tous les *n* tokens. Dans les expériences présentées ici, *n* est fixé à 10 000.

3. Préparation du matériel

Nous utilisons un corpus écrit 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 humaine. Il est en effet difficile d'effectuer une analyse systématique de corpus aussi volumineux manuellement. Nous appliquons donc des outils de traitement automatique des langues pour le nettoyage et formatage de corpus, leur étiquetage morphosyntaxique et, enfin, pour la détection de lexèmes construits et leur analyse morphologique. En revanche, 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. Nous décrivons brièvement ci-dessous les différentes étapes du traitement.

	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

TAB. 1: Taille des corpus par rubrique du Monde en 1995.

Comme c'est l'étiquetage morphosyntaxique qui apporte les informations principales pour l'analyse morphologique, c'est autour de cette couche de traitement que les efforts se concentrent. Avant l'étiquetage, nous effectuons un nettoyage du format du matériel source : suppression des fins de lignes au milieu des mots et des phrases, découpage en paragraphes, ventilation des articles en fonction des rubriques du *Monde*. Le tableau 1 présente les corpus obtenus suite à l'ensemble de ces traitements. Les deux premières colonnes nomment 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 compter les signes de ponctuation. C'est par rapport aux occurrences autres que la ponctuation que les prélèvements sont faits. Chaque rubrique, de par sa thématique que nous supposons homogène, constitue un corpus. Les traitements décrits dans la suite de ce paragraphe sont appliqués à chaque corpus. L'étape suivante consiste à segmenter les corpus en mots et à effectuer un préétiquetage avec un lexique généraliste du français. Ce préétiquetage est indispensable, dans la mesure où l'étiqueteur morphosyntaxique TREETAGGER (Schmid, 1994), que nous utilisons, ne prévoit pas d'autre possibilité pour enrichir son lexique interne. La prise en compte de lexiques externes assez volumineux,

environ 400 000 formes provenant essentiellement du corpus Frantext[ii] , permet de contourner cette limite. Au terme de cet étiquetage, les sorties de TREETAGGER peuvent contenir des erreurs d'étiquetage ou des lemmes non connus (<UNKNOWN>). Nous utilisons alors le lemmatiseur du français FLEMM (Namer, 2000) pour corriger certaines de ces erreurs et pour proposer les lemmes non reconnus par TREETAGGER. Nous appliquons ensuite l'analyseur morphologique DERIF (Hathout et al., 2001, Namer, 2002) qui, sur la base de l'étiquetage morphosyntaxique, des règles de construction de lexèmes en français et des listes d'exceptions, détecte les lexèmes construits et propose leur analyse morphologique : décomposition en unités morphologiques répertoriées et ordre des opérations morphologiques. Les résultats de l'analyse morphologique des corpus sont chargés dans la base de données MYSQL. Nous extrayons ensuite les lexèmes suffixés en *-Able*, de même que leur analyse morphologique, et les soumettons à une validation manuelle. L'objectif de la validation manuelle consiste à éliminer les lexèmes qui ne sont pas analysables comme construits en français. Par exemple, parmi les lexèmes en *-Able*, nous ne prenons pas en compte *formidable* et *perméable*.

Les adjectifs *-Able* validés constituent notre matériel de travail. Nous les observons sur l'ensemble de la période 1995 et contrastons les résultats obtenus sur chaque corpus thématique.

4. Suffixation par *-Able* : fiche linguistique

Sous le suffixe *-Able*, nous regroupons aussi *-ible* et *-uble* (c'est ce que note la majuscule dans *-Able*). La RCL dont l'exposant est *-Able* (désormais la RCL^{Able}) est traditionnellement caractérisée par les deux propriétés suivantes.

Elle forme des adjectifs, et s'applique essentiellement à des verbes dont l'argument distingué, correspondant au nom recteur de l'adjectif, s'interprète comme un Patient (*laver un pull* => *un pull lavable*, *les denrées périssent* => *les denrées périssables*), ou un Site, locatif (*courir sur une surface* => *surface courable*) ou temporel (*skier pendant une période* => *période skiable*). On note que, pour un même verbe, les trois possibilités sont offertes, comme l'indiquent les exemples suivants, relevés sur Google au 22/01/2006 :

- *« En Aveyron, le brochet est pêchable du 1er janvier au 27 janvier et du 10 mai au 31 décembre inclus »*
- *« Je ne pense pas que le Rhône sera pêchable cette année »*
- *« Pour éviter cette situation discriminante, je propose de décaler les dates de la période pêchable »*

Elle peut aussi s'appliquer à des noms (*corvéable*, *ministrable*). L'ensemble des questions liées à l'aspect catégoriel de *-Able* sont discutées dans (Plénat, 1988 ; Fradin, 2003 ; Hathout et al., 2003).

D'un point de vue sémantique, les adjectifs que construit la RCL^{Able} indiquent que le référent de leur nom recteur peut être affecté par le procès que désigne le lexème-base, quand il s'agit d'un verbe, ou mis en jeu par lui, quand il s'agit d'un nom. Les discussions liées à la sémantique des adjectifs en *-Able* sont présentées dans les travaux déjà cités ainsi que dans (Leeman, 1992 ; Anscombre & Leeman, 1994).

5. Productivité de la suffixation par *-Able* : résultats et discussion

La productivité est un phénomène difficile à observer. Nous nous basons ici sur ses différents aspects présentés par l'ensemble des figures que nous discuterons : la croissance du

[ii]Ce lexique peut être téléchargé depuis le site www.lexique.org

vocabulaire tout au long des corpus et les indices de productivité. Nous tâchons de croiser ces différents résultats et de statuer sur la productivité de la RCL^{Able}.

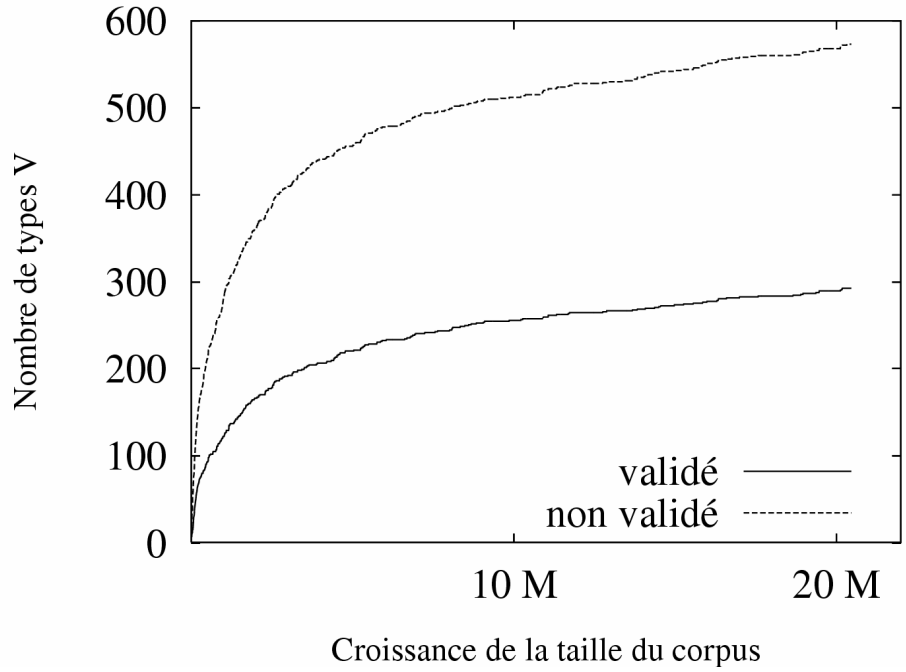

FIG. 1: Croissance du vocabulaire due à la suffixation par -Able : la totalité des articles en 1995

5.1 Validation humaine et croissance du vocabulaire

La figure 1 montre l'influence de la validation humaine sur la croissance du vocabulaire et, par là même, sur la productivité. La croissance du vocabulaire reflète la capacité d'une RCL à produire de nouveaux lexèmes dans le corpus étudié. Comme indiqué auparavant, nous prélevons le nombre de types $V_{Able}(N)$ en *-Able* toutes les 10 000 occurrences. Sur la figure 1, les prélèvements sont faits sur l'ensemble des articles parus en 1995. Lors de l'interprétation des résultats, on se tient de manière générale aux principes suivants :

– si la courbe s'aplatit la RCL tend à atteindre une saturation et devient non productive,
– si, au contraire, la courbe continue de grimper la RCL se montre productive.

Cette figure montre ainsi que la RCL^{Able} est productive car, tout au long des corpus, elle continue de former de nouveaux lexèmes. Mais les deux courbes, *non validé* et *validé*, présentent des différences évidentes qui sont dues à l'intervention humaine. Elles peuvent être observées à deux moments : le nombre de types *V* et la forme des courbes.

Le nombre de types $V_{Able}(N)$ est environ deux fois plus élevé avec les données non validées. La validation humaine a en effet éliminé 48 % de lexèmes supposés construits à l'issue des traitements automatiques. Pendant les validations, nous éliminons les lexèmes de types suivants :

1. Lexèmes dans lesquels la suffixation par *-Able* n'est pas la dernière opération constructionnelle : *dissemblable*, *irrésistible*, *coresponsable*, etc.
2. Lexèmes non analysables ou difficilement analysables comme construits en français. Il s'agit souvent d'emprunts : *affable*, *impeccable*, etc.
3. Lexèmes non suffixés. Dans ce cas, la chaîne de caractères a une ressemblance accidentelle avec *-Able* mais ne peut pas être considérée comme un affixe : *faible*, *double*.

4. Lexèmes ayant un sens lexicalisé différent du sens attendu : *remarquable*, *aimable*.
5. De nombreuses erreurs : *comprable*, *ineffaÇable*, etc.
6. Erreurs de catégorisation syntaxique, comme par exemple les noms *responsable* et *portable*, formés à partir des adjectifs par conversion.

Suite à l'élimination de ces erreurs, nous gardons uniquement les lexèmes construits par la suffixation par *-Able*. Notons que certains de ces types d'erreurs (1, 3, 5, 6) ont déjà été distingués dans (Evert & Lüdeling, 2001).

Quant aux courbes, leur forme ne marque pas de tendance vers une asymptote. Mais la courbe *validée* est beaucoup plus aplatie. On peut néanmoins considérer que la suffixation par *-Able* se montre productive dans les articles du *Monde* parus en 1995 : vers la fin du corpus, de plus de 20 millions d'occurrences, même avec les données validées la RCL^{Able} continue de former de nouveaux lexèmes non encore rencontrés. La différence entre les courbes *non validé* et *validé* est pourtant significative. Un tel effet de la validation humaine sur la croissance du vocabulaire et la productivité a déjà été remarqué dans (Evert & Lüdeling, 2001). Dans le travail cité et dans les résultats que nous obtenons, l'analyse automatique se montre ainsi assez déviante de la validation humaine. Dans notre travail, nous effectuons pourtant des pré-traitements assez importants qui visent à améliorer la reconnaissance des occurrences et leur étiquetage morpho-syntaxique. L'analyseur morphologique DERIF, que nous utilisons, est spécifiquement dédié à cette tâche. Notons que l'analyse qu'il propose est assez fine : décomposition en unités morphologiques, établissement de l'ordre des opérations morphologiques et glose sémantique de lexèmes construits. L'ensemble de ces traitements améliorent les résultats et soulagent d'autant la validation humaine. Il est important de remarquer par ailleurs que les décisions prises lors de la validation humaine sur le statut de tel ou tel lexème influencent également les résultats. Il va de soi que la validation humaine est différente selon les personnes qui interviennent dans ce processus. Il en est de même lorsque différents outils automatiques sont utilisés (Evert & Lüdeling, 2001).

Ces résultats montrent que les outils automatiques, tout en présentant une aide très importante lors des traitements de données textuelles, ne suffisent pas en soi. Les sorties de ces outils peuvent indiquer des tendances approximatives, mais pour avoir des données plus précises, une intervention humaine s'impose. La performance des outils automatiques de même que les décisions prises lors des validations des données effectuent donc une influence certaine sur les résultats.

5.2 Croissance de vocabulaire selon les corpus thématiques

La figure 1 présente la croissance du vocabulaire sur la totalité des articles parus en 1995 dans *Le Monde*. Mais si on regarde les corpus thématiques, figures 2 et 3, construits en fonction des rubriques du *Monde*, on voit que le comportement de la RCL^{Able} en fonction des rubriques est différent, et qu'elle a un comportement spécifique dans chaque corpus, qu'il s'agisse des données brutes ou des données validées. (La signification des noms de corpus et leur taille sont indiquées plus haut, dans le tableau 1.) Nous voyons donc que, dans les trois cas, la déviation entre les traitements automatiques et la validation humaine est assez importante.

Avec les données validées de la figure 3, la RCL^{Able} continue de former de nouveaux lexèmes dans tous les corpus. Mais le rythme de découverte de ces nouveaux lexèmes s'atténue vers la fin des parcours. Au vue de l'ensemble de ces courbes, on peut supposer un lien entre la taille des corpus et le rythme de croissance de leurs vocabulaires, et remarquer que le rythme de croissance du vocabulaire est inversement proportionnel à la taille du corpus. Il en est ainsi pour les corpus INT, FRA, ART et SOC, donnés ici dans l'ordre décroissant de leur taille : la croissance de leurs vocabulaires est la plus importante dans SOC, et la moins importante dans INT. Seul le corpus LIV semble ne pas respecter cette tendance : ayant une taille inférieure à

celle du corpus SOC, le rythme de croissance de son vocabulaire est également inférieur. Par ailleurs, c'est le corpus SOC qui atteint un nombre de types le plus élevé.

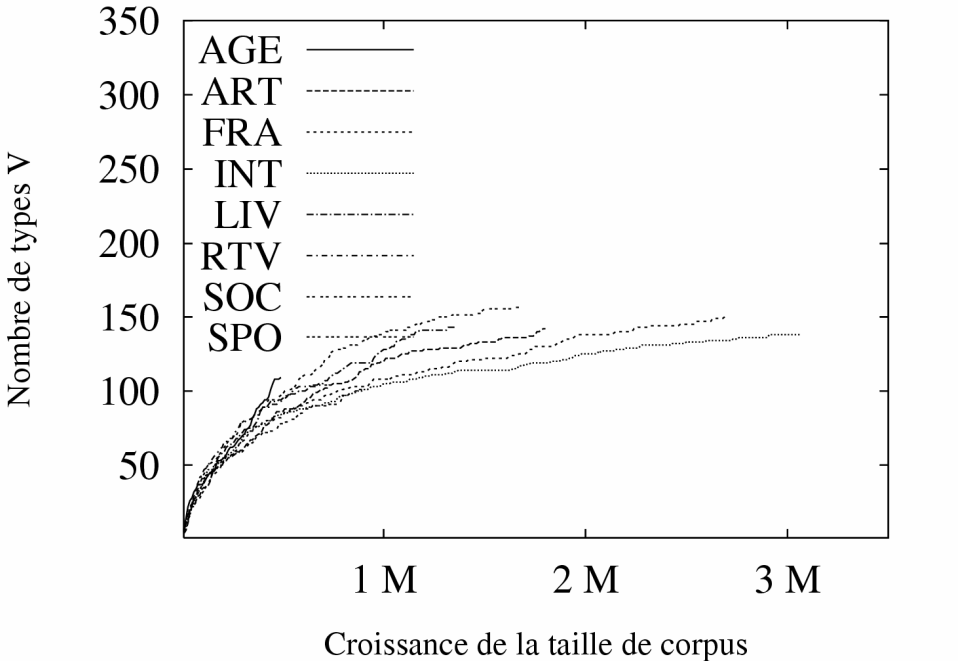

FIG. 2 : Croissance du vocabulaire due à la suffixation par -Able : données non validées

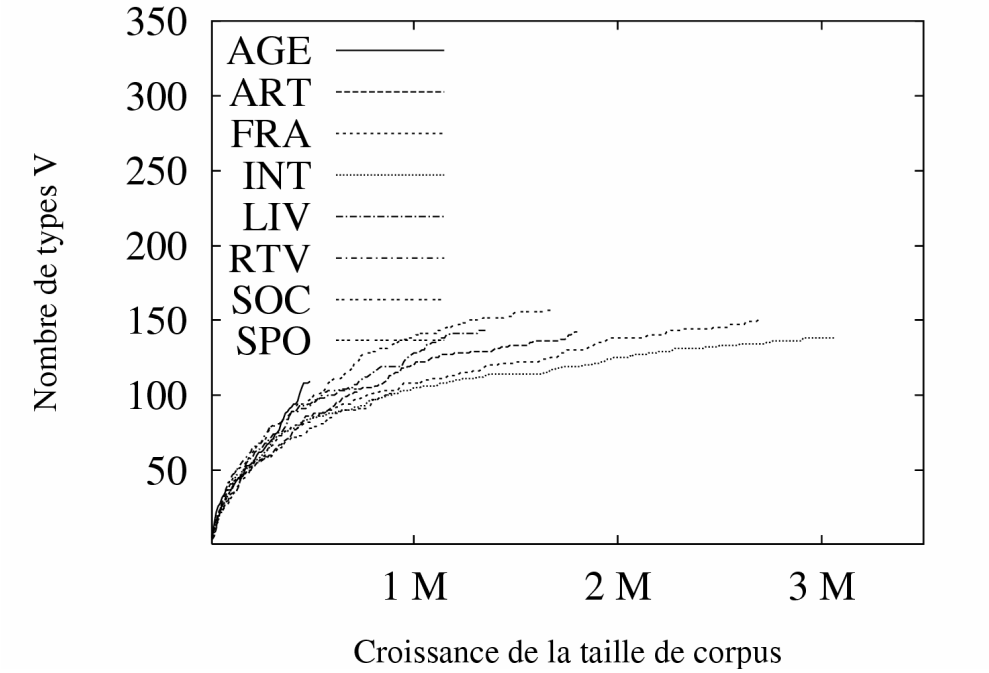

FIG. 3 : Croissance du vocabulaire due à la suffixation par -Able : données validées.

En généralisant ces remarques, on peut supposer qu'un corpus court est « plus pressé de montrer » son vocabulaire (AGE), alors qu'un corpus plus grand dispose de plus de temps pour le faire (INT). On pourrait croire aussi que les vocabulaires eux-mêmes sont comparables. La figure 3 montre pourtant que chaque corpus contient des nombres de types

différents, qui sont en réalité plus ou moins comparables entre eux. Nous savons par ailleurs que, s'il existe des lexèmes communs à tous les corpus ou à certains d'entre eux, il existe également des lexèmes spécifiques à chaque corpus. Voilà quelques adjectifs en *-Able* qui n'apparaissent que dans un des corpus :

- ART : *regardable*, *écoutable*, *oubliable*, *audimatisable*, *avionnable*
- FRA : *taxable*, *liquidable*, *déclinable*
- LIV : *surpassable*
- SPO : *sélectionnable*, *opérable*, *médaillable*
- SOC : *éducable*

En réalité, chaque nouveau corpus apporte des lexèmes « endémiques » en plus.

*5.3 Deux indices de productivité P et P**

Les fig 4 et 5 montrent l'évolution des deux indices de productivité de la suffixation par *-Able* par rapport au nombre de types *V*. Comme auparavant, les prélèvements sont faits toutes les 10 000 occurrences. Notons que l'axe sur lequel sont projetés les *indices de productivité P* et *P** est à l'échelle logarithmique. Pour rappel, l'indice *P* est dit dépendant de la taille de corpus. Comme son dénominateur correspond au nombre total des lexèmes construits par une RCL, plus un corpus est grand, plus N_c sera important et *P* petit. Dans la même logique, les valeurs de *P* décroissent au fur et à mesure de la lecture du corpus, à moins bien sûr de ne rencontrer que des hapax. Mais ce dernier cas est peu probable. Quant à l'indice *P**, il est dit indépendant de la taille du corpus, mais dépendant du nombre total de hapax dans ce corpus, tous lexèmes confondus. Le dénominateur de *P** correspond donc 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 *P** restent stables, et la règle peut alors être considérée comme productive. Par contre, si la croissance du vocabulaire formé par une règle est moins importante que la croissance du vocabulaire du corpus en général, les valeurs de *P** vont aller en diminuant et la règle sera considérée comme non productive.

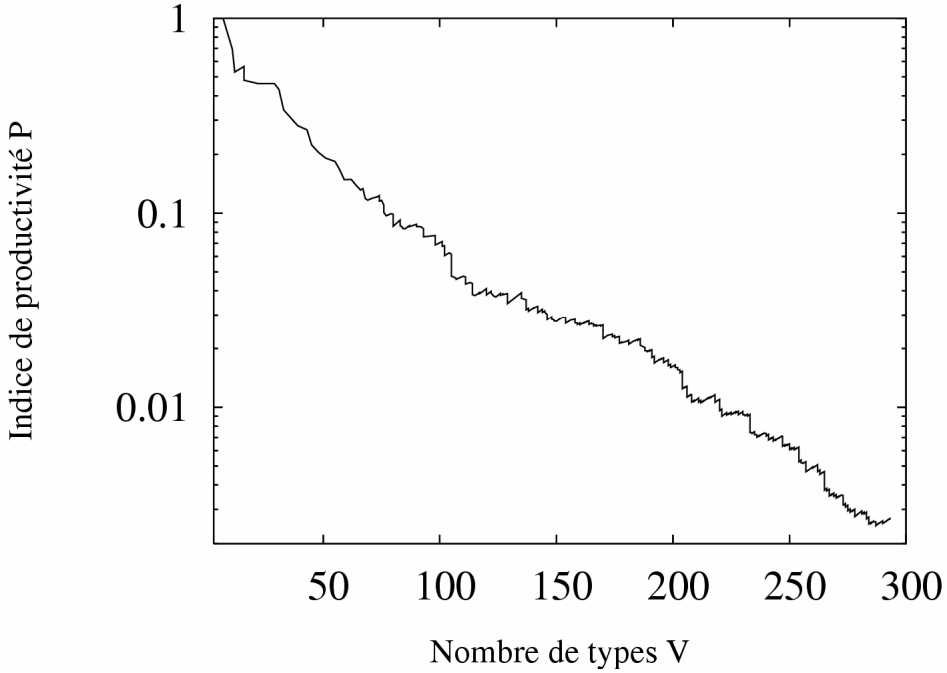

FIG. 4 : Indice de productivité P en 1995.

Comme le montrent les figures 4 et 5, l'évolution de ces deux indices est comparable, même si on note une hésitation de la courbe *P** en début du corpus 1995 : leurs valeurs vont en diminuant. Ceci dit, le rythme de leur décroissance est très faible. De manière générale, la courbe *P** évolue par à-coups. Cette évolution dépend de l'ensemble des hapax du corpus, qui sont largement nourris par les noms propres et les fautes d'orthographe. La compétition est donc difficile pour la suffixation par *-Able*. Quant à la courbe *P*, sa forme décroissante est attendue : elle est dépendante de la croissance de la taille du corpus. Remarquons que les deux courbes terminent l'année en « remontant ». Ceci est surtout visible sur la figure 5.

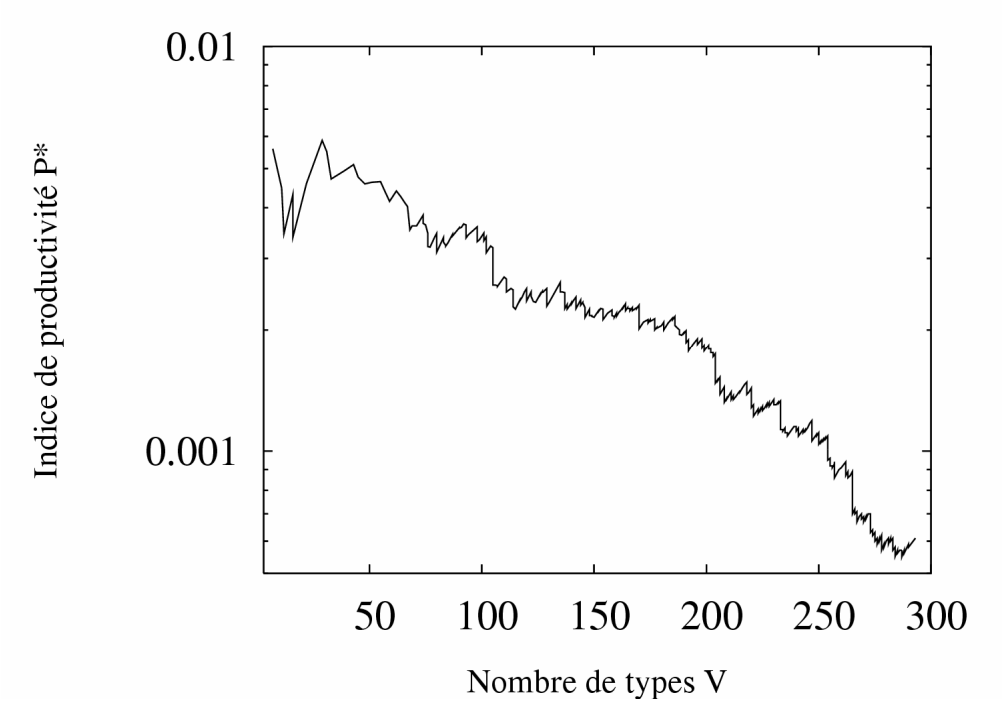

FIG. 5 : Indice de productivité P en 1995.*

5.4 Deux indices de productivité P et P selon les corpus*

Les figures 6 et 7 présentent également l'évolution des indices de productivité *P* et *P** mais cette fois pour chaque corpus thématique. De manière générale, *P* de chaque corpus évolue comme pour la totalité de la période, en montrant une décroissance, tandis que *P** a un comportement différent en fonction des corpus.

Sur la figure 6, nous pouvons voir que l'indice *P* est souvent égal à 1 au début des corpus : les premiers types n'occurrent qu'une seule fois lors des premiers prélèvements. Seul le corpus ART se différencie alors : parmi ses premiers types plusieurs occurrent plus d'une fois dès le début. Mais le comportement de la RCL^{Able} dans ART rejoint assez rapidement celui des autres corpus, où les courbes forment un couloir assez homogène. Son évolution dans le corpus AGE est également intéressante : l'indice de productivité décroît fortement dans le premier tiers du corpus, mais il reste à un niveau assez stable ensuite. Dans AGE, le vocabulaire des adjectifs en *-Able* est alors enrichi sans montrer énormément de répétitions. Les valeurs finales de l'indice *P* de tous les corpus, sauf peut-être celles de AGE, tendent vers le bas. Trois corpus peuvent néanmoins être distingués, SOC, LIV et ART : faisant partie des plus grands ils continuent de montrer l'indice *P* relativement plus élevé que dans deux autres grands corpus, INT et FRA.

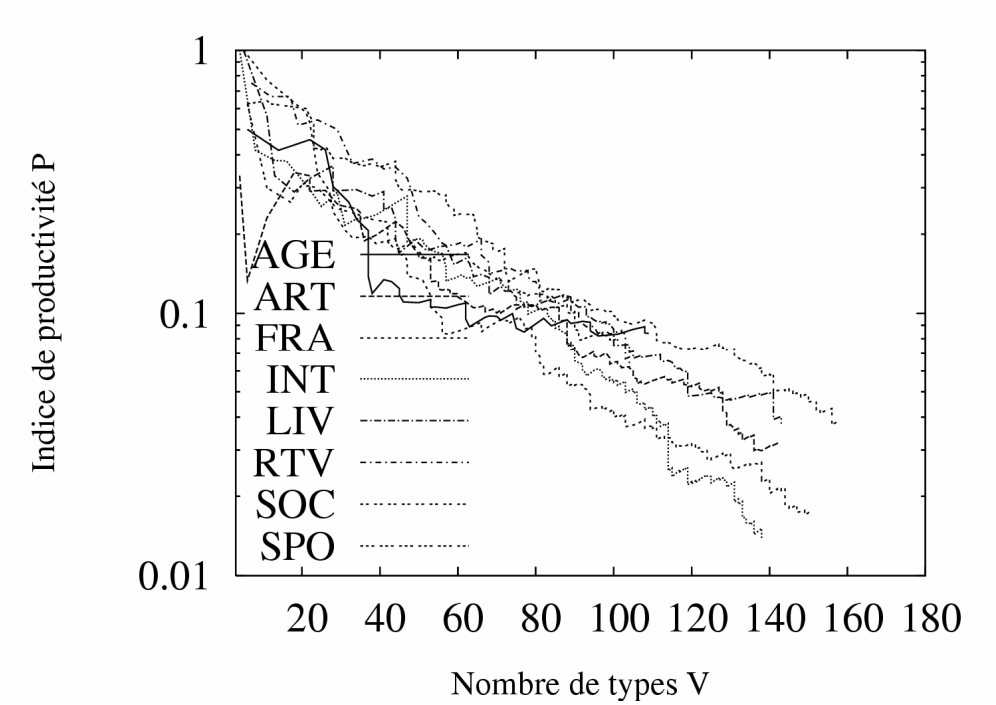

FIG. 6 : Indice de productivité P selon les corpus.

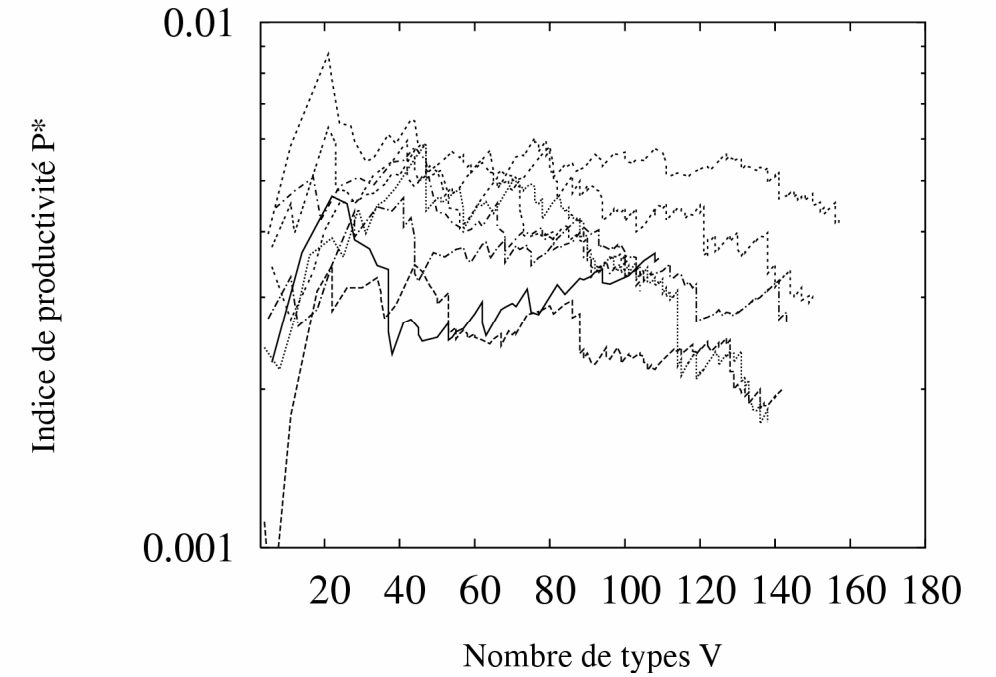

FIG. 7 : Indice de productivité P selon les corpus.*

La fig. 7 montre que *P** connaît un pic au début de la majorité des corpus : les hapax en *-Able* augmentent alors plus rapidement que l'ensemble des hapax de ces corpus. Les valeurs de *P** se stabilisent ensuite et décroissent légèrement vers la fin des corpus. C'est encore AGE qui se différentie : son indice *P**, contrairement à ce qui se passe dans d'autres corpus, croît. Il reste sur une pente ascendante, effet qui, avec *P** ne peut pas être dû à la petite taille de ce

corpus, mais est lié au rythme de découverte de nouveaux hapax en *-Able*. Comme avec *P*, SOC et LIV, mais également FRA, connaissent des valeurs de l'indice *P** relativement élevées, tandis que ART et INT terminent avec des valeurs de *P** moins importantes. Dans la section précédente, nous avons remarqué que les deux courbes « remontent » vers la fin. À l'examen des figures 6 et 7, il apparaît que c'est le corpus ART qui contribue à cette croissance des indices de productivité.

5.5 Productivité de la suffixation par -Able dans les articles du Monde parus en 1995

Afin de statuer sur la productivité de la suffixation par *-Able*, nous nous basons sur la discussion de l'ensemble des figures présentées. L'observation des figures qui présentent le comportement de la RCL à laquelle est associé *-Able* durant toute la période 1995 dans *Le Monde*, fig. 1, 4 et 5, montre que cette règle suit l'évolution attendue : les courbes suivent une pente descendante. Mais elle continue de former de nouveaux lexèmes et ne semble pas vouloir atteindre une saturation (fig. 1). De plus, vers la fin de l'année, la pente des courbes des indices de productivité remonte légèrement (fig. 4 et 5).

L'observation des corpus thématiques composés de la majorité de ces mêmes articles permet de confirmer la vivacité de la RCL^{Able} (fig. 3, 6 et 7). Cela permet surtout de distinguer des « zones » de productivité et de non-productivité à l'intérieur de toute la période 1995. Du point de vue de la croissance de vocabulaires, ce sont les corpus AGE et SOC qui montrent une RCL^{Able} la plus créative en nouveaux lexèmes. SOC est par ailleurs le corpus qui propose le plus de lexèmes de ce type. Quant aux indices de productivité, les corpus SOC et LIV montrent les indices les plus élevés. Ils sont rejoints par AGE et ART pour l'indice *P* et par FRA pour l'indice *P**. De manière générale, c'est dans les corpus SOC, AGE et LIV que la RCL^{Able} se montre la plus productive. Les courbes dans ART remontent légèrement à la fin de l'année. Cette règle est la moins productive dans le corpus INT, qui est le corpus le plus grand et des moins riches en types en *-Able*.

6. Conclusion et perspectives

Nous avons présenté un travail autour de la productivité des règles de construction de lexèmes en français contemporain. Nous nous sommes concentrés en particulier sur la RCL^{Able} à travers les articles du *Monde* parus en 1995. Du point de vue linguistique, elle forme des adjectifs à partir de verbes, et distingue surtout le Patient mis en jeu par le lexème-base. Les adjectifs qu'elle construit expriment le fait que le référent de leur nom recteur peut être affecté par le procès associé au lexème-base.

Dans nos analyses, nous avons croisé plusieurs indications fournies par (1) les courbes de croissances de vocabulaires, (2) l'indice de productivité *P*, dit dépendant du corpus, car il pondère la productivité d'un procédé par rapport à toutes ses occurrences, (3) et l'indice de productivité *P**, dit dépendant des hapax, car il pondère la productivité d'un procédé par rapport à tous les hapax d'un corpus. La suffixation par *-Able* se montre ainsi assez productive, du moins pas saturée, sur l'ensemble de la période. L'analyse de corpus thématiques de cette période permet en plus de voir que la zone de productivité de la RCL^{Able} englobe les corpus SOC (rubrique Société), AGE (Agenda) et LIV (Livres), tandis que la zone de non productivité englobe le corpus INT (International). Dans la zone de productivité, beaucoup d'entités et d'événements sont caractérisés par les actions qui leur sont éventuellement applicables. De la part des journalistes, il s'agit sans doute de suppositions et d'évaluations nuancées par le sens du « possible » véhiculé par la suffixation par *-Able*.

Nous montrons par ailleurs qu'un travail sur des données textuelles, surtout si les expériences portent sur de grands volumes de données et si elles sont redondantes, devrait relier des outils automatiques avec l'expertise humaine. Les outils automatiques permettent de pré-traiter et de préparer des données brutes, tandis que l'intervention humaine est indispensable pour statuer

là-dessus. Les outils automatiques indiquent des tendances générales, tandis que l'intervention humaine permet d'observer le phénomène avec plus de précision.

En parallèle, nous effectuons un travail similaire sur d'autres règles de construction de lexèmes (préfixation par *in-*, suffixations par *-ité*, *-ifier* et *-ion*), avec l'objectif de dresser un tableau de productivité comparative de la majorité des RCL du français. Ce travail sera appliqué à d'autres années du *Monde* et éventuellement à d'autres corpus, ce qui permettra de donner une vue évolutive de la morphologie du français actuel. Les courbes de la suffixation par *-Able* seront comparées avec les courbes de ces autres règles, ce qui nous permettra de les appréhender avec plus de recul et compréhension.

Les mesures probabilistes de mesure de la productivité s'appuient sur un modèle de génération de l'urne, selon lequel chaque occurrence est tirée (avec remise) parmi un ensemble de types possibles, chacun étant caractérisé par une probabilité d'apparition p_i. Il reste à évaluer l'impact de ce choix de modélisation, en étudiant d'autres modèles probabilistes des fréquences lexicales, *eg.* (Church & Gale, 1995, Katz, 1996).

Références

Anscombre, J. and Leeman, D. (1994). La dérivation des adjectifs en *-able* : morphologie ou sémantique ? *Langue Française*, 103, 32-44.

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

Church, K. W. and Gale, W. A. (1995). Poisson mixtures. 1(2), 163-190.

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

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

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

Hathout, N., Namer, F. and Dal, G. (2001). An experimental constructional database: the MorTAL project. In P. Boucher, Ed., *Morphology book*. Cambridge, MA: Cascadilla Press.

Hathout, N., Plénat, M. and 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.

Katz, S. M. (1996). Distribution of content words and phrases in text and language modelling. *Journal of Natural Language Engineering*, 2(1), 15-59.

Leeman, D. (1992). Deux classes d'adjectifs en *-ble*. *Langue Française*, 96, 44-64.

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),* pp. 235-244, Nancy.

Plénat M. (1988). Morphologie des adjectifs en *–able*. *Cahiers de grammaire*. 13, pp. 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*, pp. 44-49, Manchester, UK.