Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Morphologie
DOI 10.1051/cmlf08184

Quelques préalables au calcul de la productivité des règles constructionnelles et premiers résultats

Georgette Dal

U Lille 3, et UMR 8163 STL, CNRS & U Lille 3 et Lille 1
georgette.dal@univ-lille3.fr

Bernard Fradin & Clément Plancq

UMR 7110 LLF, CNRS & U Paris 7
{bernard.fradin, clement.plancq}@linguist.jussieu.fr

Natalia Grabar

DIH/HEGP, U Paris 5; U872 INSERM, SPIM
natalia.grabar@club-internet.fr

Stéphanie Lignon & Fiammetta Namer

U Nancy 2, et UMR 7118 ATILF, CNRS & Nancy-Université
{stephanie.lignon, fiammetta.namer}@univ-nancy2.fr

François Yvon

U Paris Sud 11, et UPR 3251 LIMSI, CNRS
francois.yvon@limsi.fr

Pierre Zweigenbaum

UPR 3251 LIMSI, CNRS et EA 2520 ERTIM, INALCO
pz@limsi.fr

Au plan international, la réinscription, à l'agenda des morphologues, de la question de la productivité remonte à une quinzaine d'années. Elle est consécutive au fait que de nouvelles mesures de la productivité, beaucoup plus satisfaisantes que les précédentes, ont été proposées en 1991 par Harald Baayen, aux Pays Bas. Depuis, de nombreux travaux, alliant des compétences disciplinaires multiples, ont vu le jour, faisant émerger toute une série de questions auxquelles seule une collaboration forte entre linguistes, talistes, statisticiens et psycholinguistes permet de répondre. En France, la question de la productivité a jusqu'à présent été peu travaillée, en dehors des recherches menées dans l'action « Productivité morphologique » du GDR 2220 « Description et modélisation en morphologie » qui a pris fin en décembre 2007, et dont la particularité était précisément de réunir des compétences disciplinaires variées[1].

Dans la présente contribution, nous donnerons un aperçu des recherches menées dans ce groupe de travail et des résultats auxquels nous sommes parvenus.

1 Définition préalable

Dans le domaine de la morphologie constructionnelle, la productivité est d'abord une notion qualitative. Dans cette perspective, on considère qu'une règle de construction de lexèmes (désormais RCL), conçue comme constituant une généralisation de rapports pouvant être établis entre données observables, est

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Morphologie
DOI 10.1051/cmlf08184

productive quand de nouveaux lexèmes, supposément jamais produits auparavant ou inconnus du locuteur ou du scripteur qui les produit, peuvent l'instancier[2].

La difficulté, avec une telle définition, est qu'au-delà de l'intuition qui peut être celle du descripteur, elle demande qu'on se donne les moyens d'observer la productivité des RCL et de la mesurer sur des corpus authentiques. A ce jour, plusieurs mesures ont été proposées. Les plus utilisées actuellement sont dues à H. Baayen. Elles sont au nombre de deux, et se calculent toujours relativement à un corpus donné, que nous noterons *C* :

- La première, notée P et appelée 'indice de productivité' (d'une RCL donnée), correspond au quotient du nombre n_1 d'hapax (i.e. formes de mots qui n'apparaissent qu'une fois dans *C*) qui instancient le procédé étudié par le nombre *N* d'occurrences de formes de mots qui instancient le procédé dans *C*. C'est ce que note : $P_C = n_1/N$. Cette mesure correspond à la probabilité que la prochaine occurrence de forme de lexème analysable selon la RCL étudiée soit d'un type nouveau. Elle est particulièrement utile pour comparer la productivité de RCL entre elles, ou pour comparer la productivité d'une même RCL dans des corpus différents.
- La seconde, notée P^*, correspond au quotient du nombre n_1 d'hapax instanciant le procédé étudié par le nombre total N_1 d'hapax dans le corpus *C* : $P^*_C = n_1/N_1$. P^* donne la contribution de la RCL étudiée à la croissance du vocabulaire dans le corpus considéré.

Idéalement, il convient d'observer simultanément ces deux mesures : chacune prise séparément n'offre en effet qu'une vue partielle sur les phénomènes que l'on évalue.

Ces mesures ont déjà suscité un certain nombre de remarques, que nous reprenons rapidement ici (pour un poinf, cf. Dal 2003 ; nous reviendrons sur certaines de ces remarques plus loin) :

– les hapax, sur lesquels elles reposent, peuvent être constitués de mots rares, bien ancrés dans le lexique,

– certains auteurs font un usage répété d'un lexème qui leur est propre, si bien que ces lexèmes constituent, à leur manière, des hapax d'auteur, qui pourraient être comptabilisés en tant que tels,

– la comparaison des indices de productivité ne vaut que si les corpus sont de taille similaire[3].

2 Objectif et questions

L'objectif initial que nous nous étions fixé était de dresser une cartographie des principales RCL du français comme il en existe dans d'autres langues, en utilisant les mesures *P* et *P**.

Nous nous sommes cependant vite heurtés à deux questions, classiques dès qu'il s'agit d'estimer la productivité d'une RCL, au sens où nous l'avons définie :

- quel corpus utiliser pour que les mesures proposées soient représentatives non seulement du corpus sur lequel elles ont été calculées, mais aussi du français en général (pour autant que cela ait un sens) ?
- sur quels critères se baser pour décider si tel ou tel lexème (ou telle ou telle forme de mots) doit être retenu(e) pour calculer *P* et *P** ?

Ces deux questions structureront notre propos :

– nous commencerons par nous poser la question des corpus (sections 3-4),

– nous enchaînerons avec des questions théoriques et exposerons le processus d'analyse que nous avons mis au point pour déterminer quels candidats retenir pour le calcul de la productivité des RCL (sections 5-6).

Pour terminer, nous donnerons un aperçu de nos résultats (section 7).

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Morphologie
DOI 10.1051/cmlf08184

3 La question des corpus

Les premières tentatives d'évaluation de la productivité des procédés morphologiques ont été menées à partir des dictionnaires publiés dans le commerce. Les résultats qu'elles permettent d'obtenir sont fortement sujets à caution pour des raisons identifiées depuis longtemps et rappelées dans Gaeta et Ricca (2003) :

– les dictionnaires ne constituent pas des échantillons représentatifs d'une langue,

– ils introduisent des biais liés à leur finalité qui se marquent notamment dans l'élaboration de leur nomenclature,

– ils ne permettent pas de mesures statistiques relatives à l'usage.

Cette dernière exigence demande de recourir à des corpus au sens strict, c'est-à-dire à des rassemblements de textes sélectionnés selon des critères qui en garantissent la représentativité linguistique. Dans la mesure où, dans le cas présent, l'on souhaite que le corpus reflète la compétence d'un locuteur français éduqué (mais non savant), l'idée de le constituer sur la base de journaux s'impose comme un compromis acceptable. Le fait que les journaux s'adressent à de très nombreux locuteurs ayant des intérêts variés les contraint dans une large mesure à utiliser la langue courante, même si des écarts sensibles quant au niveau de langue existent entre différents journaux et, selon les rubriques, à l'intérieur d'un même journal. Le fait qu'ils traitent aussi de sujets variés sans avoir à se focaliser sur aucun garantit par ailleurs une certaine représentativité des contenus.

Un des problèmes soulevés par l'usage de la mesure de productivité morphologique $P = n_1/N$, mis au jour par Gaeta et Ricca dans plusieurs de leurs travaux, tient au fait que le nombre N du dénominateur correspond au nombre total d'occurrences instanciant la RCL étudiée dans le corpus. De ce fait, pour chaque RCL, P décroît avec le nombre d'occurrences de formes qui instancient cette RCL[4]. Dans la mesure où tous les procédés morphologiques n'offrent pas un nombre d'instances identique dans un corpus donné, comparer différentes RCL sur la totalité d'un corpus conduit à les comparer sur des nombres d'occurrences de N différentes (Gaeta & Ricca, 2003 : 67, mentionnent un rapport pouvant dépasser 1 pour 50). Cela amène donc à surestimer dans des proportions importantes la productivité des RCL offrant peu d'occurrences (c'est-à-dire des procédés faiblement productifs). Pour rectifier ce biais, Gaeta et Ricca proposent d'adopter une approche à corpus variable (*variable-corpus approach*) dans laquelle la productivité est mesurée à partir de sous-corpus présentant un nombre d'occurrences N similaire pour chaque procédé. Cette démarche entraîne qu'on doit utiliser des sous-corpus de tailles différentes pour comparer des procédés qui présentent un nombre d'occurrences attestées dissemblables.

4 Corpus utilisés

Notre corpus de travail est composé d'articles du quotidien français *Le Monde*. Le fait d'avoir choisi ce journal tient à des facilités d'accès, de consultation et de maniement. Mais il est bien clair qu'il faudrait élargir la palette des journaux constituant le corpus, notamment pour ce qui concerne les niveaux de langue. Nous faisons l'hypothèse que les corpus journalistiques donnent un reflet non filtré de la langue courante. Comme leur parution s'étend sur plusieurs années, ils permettent d'observer l'évolution de la langue sur la période étudiée. Notre corpus du *Monde* englobe les années 1991, 1995, 1999 et 2003. Notre objectif étant pour l'instant de construire un premier outil permettant de mesurer la productivité, nous nous sommes restreints aux années 1995 et 1999.

Comme on sait, les articles des journaux sont classés en rubriques suivant leur contenu. Nous avons eu l'idée de tirer parti de cette organisation pour étudier s'il existait une corrélation entre la productivité morphologique et la nature de la rubrique dans lequel figurait le texte du corpus. Notre attention particulière s'est portée sur huit des rubriques du *Monde*. La taille de ces données, en fonction des années et des rubriques, est présentée dans le tableau 1. Ce tableau indique, pour chacune des deux années, le nombre d'articles (articles) et le nombre d'occurrences de mots, ou formes, sans la ponctuation

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Morphologie
DOI 10.1051/cmlf08184

(occurrences) dans chaque rubrique du journal. La taille globale de ce corpus est de plus de 23 millions occurrences.

	Rubrique	articles	occurrences	articles	occurrences
		1995		1999	
AGE	Agenda	1 213	490 663	1 776	650 440
ART	Événements culturels	4 242	1 801 044	4 809	2 018 424
FRA	France	6 331	2 704 350	4 264	1 948 253
INT	International	9 276	3 065 884	8 083	3 211 160
LIV	Livres	1 949	1 350 540	2 388	1 280 437
RTV	Programme TV et radio	1 217	718 586	22	5 471
SOC	Société	4 009	1 678 573	2 823	1 260 588
SPO	Événements sportifs	2 362	894 648	2 825	911 162
Total		30 599	12 500 000	26 988	11 000 000

Tableau 1 : Taille des sous-corpus formés par les rubriques du Monde en 1995 et 1999.

5 Tri des candidats : procédure automatique

Les outils du traitement automatique des langues (désormais TAL) que constituent les étiqueteurs morpho-syntaxiques et les lemmatiseurs constituent une aide appréciable dans le calcul de P et P^* :

– étant admis pour simplifier qu'une RCL donnée forme une catégorie lexicale de lexèmes et une seule, les étiqueteurs morpho-syntaxiques permettent d'étiqueter automatiquement les formes candidates au calcul de la productivité d'une RCL donnée et, par la même occasion, d'évincer les formes qui n'auraient pas la bonne étiquette. Par exemple en français pour la suffixation en *-able* formant des adjectifs, cette phase d'étiquetage automatique permet d'évincer *cartable* et *vocable*, du fait de leur étiquette nominale, et de conserver *mangeable* et *mettable,* catégorisés comme adjectifs ;

– la phase de lemmatisation[5] est cruciale pour l'estimation du nombre d'hapax que contient le corpus sur lequel se fait l'étude, aussi bien pour le calcul de n_1 (P) que pour celui de N_1 (P^*). En imaginant que, dans C, *mangeable* et *mangeables* apparaissent chacun une fois, elle évite, en les ramenant à un lemme unique, de les compter chacun comme hapax. Nous avons pris l'option de nous intéresser au décompte des lemmes et non des formes car cela nous semblait fournir des indicateurs plus compréhensibles de la productivité.

En matière de calcul de productivité, la phase de traitement automatique est donc cruciale.

Cependant, comme l'ont montré avant nous (Evert & Lüdeling, 2001), cette étape automatique ne dispense pas d'une étape manuelle, dans la mesure où tous les candidats retenus à l'issue de la phase de traitement automatique des données n'illustrent pas nécessairement la RCL en jeu. La figure 1 est extraite de leur étude, et porte sur la suffixation en *-sam* en allemand. Elle montre en abscisse le nombre d'occurrences (*tokens*) de formes de mots du corpus, et en ordonnée le nombre de lexèmes différents (*types*) qui instancient le procédé étudié.

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Morphologie
DOI 10.1051/cmlf08184

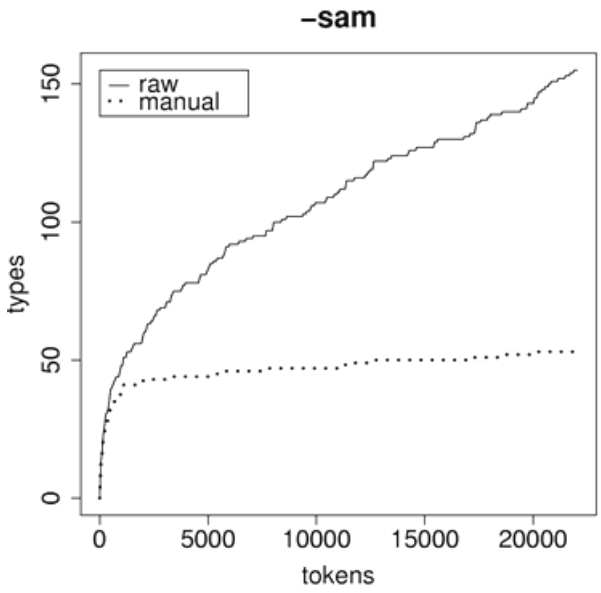
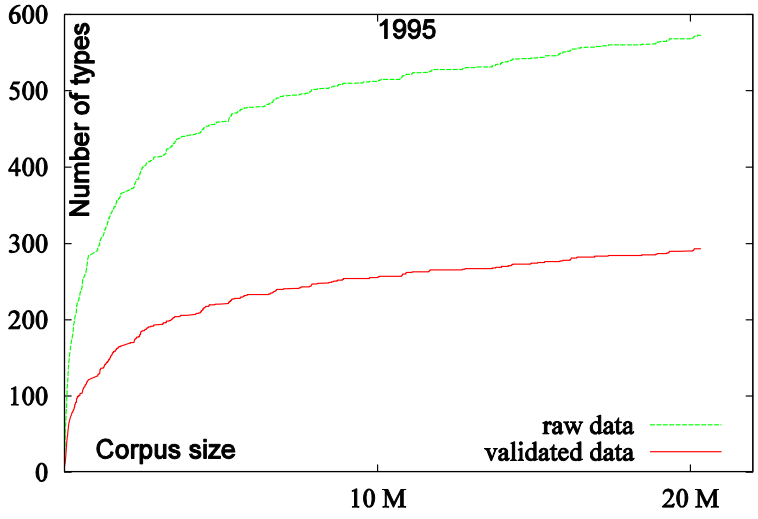


Figure 1 : Courbes de croissance brute et corrigée manuellement pour la suffixation en -sam *en allemand (Lüdeling & Evert, 2001).*

Elle fait apparaître l'écart, dans l'apport des suffixés en *-sam* à la croissance du vocabulaire dans *C*, selon que l'on s'arrête à l'étape automatique (trait continu) ou selon qu'on la complète d'une étape de validation manuelle (trait en pointillés) : sans validation manuelle, la courbe obtenue est typique d'une règle productive ; avec validation manuelle, elle s'aplatit et devient typique d'une règle non-productive.

Le travail que nous avons mené sur le français confirme les observations qui précèdent. Ainsi, sans expliciter pour l'instant ce que suppose la phase de validation manuelle des formes candidates, nous avons observé que le nombre d'adjectifs suffixés retenus au titre de la suffixation en *-able* du français diminue de moitié, selon que l'on s'arrête à la phase automatique du traitement ou qu'on la fait suivre d'une phase de validation manuelle (figure 2).



Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Morphologie
DOI 10.1051/cmlf08184

Figure 2 : Courbes de croissance brute et corrigée manuellement pour la suffixation -able *en français dans* Le Monde *1995 (Grabar* & al., *2006).*

6 Tri des candidats : phase de validation manuelle

Pour un non-morphologue, il peut paraître simple de décider de retenir telle ou telle forme pour le calcul de la productivité de la RCL que l'on étudie. Pour un morphologue, les choses sont autrement compliquées, si bien qu'une bonne partie du travail que nous avons mené a consisté à mettre au point d'une grille d'analyse constructionnelle permettant d'établir si un lexème est construit ou non. Ces problèmes d'analysabilité ne sont pas propres au français et ont été clairement identifiés dans la littérature.

Nous avons ainsi distingué trois catégories principales dans lesquelles nous avons classé manuellement les séquences candidates au calcul de *P* et de *P**, à l'issue de la phase automatique :

1. La première marque les séquences dans lesquelles nous avons reconnu des lexèmes analysables comme construits en français selon la RCL étudiée. Par analysabilité, nous entendons adopter un point de vue strictement synchronique de conformité au patron constructionnel que l'on peut dégager de l'observation des données.

Ainsi, ACCEPTABLE et MANGEABLE sont tous deux conformes au patron correspondant à la suffixation en *-able* du français : ce sont des adjectifs qui peuvent être mis en relation formelle et sémantique avec des verbes, resp. ACCEPTER et MANGER, vis-à-vis desquels ils entretiennent chacun le même rapport sémantique (ils marquent l'expression de la potentialité de réalisation du procès décrit par le verbe-base). Nous les avons donc marqués 1, en distinguant deux cas, selon que l'analysabilité fait ou non intervenir ce que nous avons appelé un accident phonologique. Selon ce critère, nous marquons différemment MANGEABLE, dans lequel la suffixation en *-able* opère directement sur le radical *mang-* de MANGER (cat. 1.a), et CORRUPTIBLE, dont l'analyse sur CORROMPRE met en jeu une allomorphie radicale (*corrupt-* vs *corromp-* ; cat. 1.b).

2. La deuxième catégorie marque les lexèmes que nous avons considérés comme des construits dans d'autres langues, en nous fondant sur les indications lexicographiques du *TLF*, la contrainte étant qu'ils soient analysables dans la langue de départ. Le point de vue est ici celui de la diachronie ou, plus rarement, de la diatopie.

Ainsi, FRIABLE est-il un emprunt à un adjectif latin construit (< lat. FRIABILIS 'qui se rompt facilement'), de même qu'AMIABLE (< lat. AMICABILIS 'qui témoigne de l'amitié') ou ABOMINABLE (< lat. médiéval ABOMINABILIS 'qui inspire l'aversion'). La séquence /bl/ qu'ils comportent est le résultat de l'évolution phonétique du suffixe latin /bilis/ et correspond à celle qu'on observe en français dans les adjectifs régulièrement analysables comme construits par la RCL en question. Pour autant, la sémantique de ces adjectifs n'est pas compositionnelle en français, et il est même difficile de voir sur quel verbe simple du français construire l'interprétation. Ces adjectifs sont des lexèmes complexes non construits et correspondent à ce que (Corbin, 1987) dénommait des « mots complexes non construits ». Outre l'emprunt, leur origine peut aussi résulter du fait que le lexème-base a disparu de la langue, comme cela s'est passé avec GALET, régulièrement formé sur l'ancien français °GAL 'caillou' aujourd'hui disparu. Comme le notent (Gaeta & Ricca, 2006 : 74) à propos de ce qu'ils appellent les « baseless derivatives », la question de la prise en compte de ces cas pour la mesure de la productivité se pose de façon cruciale.

L'emprunt à une autre langue n'interdit pas l'analysabilité en français (ou le contraire). Ainsi, selon le *TLF*, ACCEPTABLE est-il un emprunt à l'adjectif latin construit ACCEPTABILIS. Il s'ensuit qu'ACCEPTABLE est à la fois marqué 1 (analysable en français), et 2 (c'est un emprunt). En revanche, FRIABLE ne sera marqué que 2, dans la mesure où ce lexème ne se laisse formellement apparier à aucun verbe-base du français.

Les deux points de vue (celui de l'analysabilité en français, celui de l'emprunt à une autre langue) sont orthogonaux l'un à l'autre : le premier met en œuvre la compétence linguistique que les locuteurs du

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Morphologie
DOI 10.1051/cmlf08184

français ont sur leur propre langue ; le second fait appel à des connaissances qui relèvent de la culture personnelle (par exemple, connaissance du latin). On sait de toute façon qu'on ne peut pas toujours déterminer si un lexème analysable, qui possède un étymon construit, n'a pas été re-construit sur la période française.

3. La troisième catégorie réunit les lexèmes dans lesquels l'affixation étudiée est à l'œuvre du point de vue de l'analysabilité, mais ne constitue pas la dernière opération constructionnelle. Il s'agit, si on veut, d'une analysabilité indirecte en français. Ainsi, pour la suffixation en *-able*, nous avons marqué de la sorte tous les lexèmes de forme *inXable*, dans lesquels la dernière opération constructionnelle intervenue est la préfixation en *in-* (et non la suffixation en *-able*) : par exemple, IMMANGEABLE ou IMMETTABLE.

Le tableau 2 reprend ces différents cas de figure :

Type d'analysabilité	Exemples de lexèmes
1. Analysabilité synchronique	ACCEPTABLE (< ACCEPTER) MANGEABLE (< MANGER)
2. Analysabilité diachronique (diatopique)	ACCEPTABLE (< lat. ACCEPTABILIS) FRIABLE (< lat. FRIABILIS))
3. Analysabilité indirecte	AUTOCASSABLE (< CASSABLE < CASSER) IMMANGEABLE (< MANGEABLE < MANGER)

Tableau 2 : Types d'analysabilité.

La phase de validation manuelle a ainsi consisté en ce que nous avons marqué manuellement chaque lexème rencontré en corpus, en nous répartissant la tâche. Nous avons ainsi traité des suffixations en *-able*, *-eux* (ACRIMONIEUX), *-if* (IRRUPTIF), *-fier* (AMPLIFIER), *-ion* (UNION), *-ique* (ALGEBRIQUE), *-iser* (AMERICANISER), *-oir(e)* (ABATTOIR), et partiellement de la préfixation en *in-*, et nous avons consacré plusieurs séances de travail à discuter de nos marquages pour les cas litigieux. Nous voulions en effet nous assurer que les analyses que chacun proposait étaient partagées, et partageables, par tous. Le travail est fastidieux, mais c'est grâce à lui que nous avons pu affiner notre grille au fil des séances.

7 Calculs de productivité

7.1 Principes linguistiques d'utilisation de la grille d'analyse

Par rapport à la problématique de la productivité, du point de vue du morphologue qui travaille en synchronie sur une RCL donnée et qui cherche à en évaluer la productivité, seule la catégorie 1 est en théorie pertinente.

Si la notion de productivité a une quelconque légitimité du point de vue psycholinguistique, il nous a cependant paru que les catégories 2 et 3 pouvaient elles aussi être intéressantes, dans la mesure où il n'est pas déraisonnable de penser qu'à chaque fois qu'un locuteur produit ou est exposé à une séquence comme FRIABLE (cat. 2) ou comme INCOLLABLE (cat. 3), cette dernière incrémente aussi la productivité de la suffixation en *-able.* Nous faisons cette hypothèse à titre d'heuristique sans nous prononcer sur sa légitimité, et laissons le soin aux psycholinguistes de dresser les protocoles expérimentaux qu'il faut. A titre indicatif, nous préciserons seulement que, lorsqu'ils calculent la taille des familles constructionnelles, les psycholinguistes incluent les lexèmes que nous avons marqués 3[6].

Aussi avons-nous décidé d'effectuer des calculs de productivité différenciés qui figurent dans le tableau 3, et qu'illustrent les exemples en *-able* vus jusqu'ici :

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Morphologie
DOI 10.1051/cmlf08184

Catégorie(s)	Lexèmes en *-able* retenus
1	ACCEPTABLE, MANGEABLE, ~~FRIABLE~~, ~~AUTOCASSABLE~~, ~~IMMANGEABLE~~
1 ∪ 2	ACCEPTABLE, MANGEABLE, FRIABLE, ~~AUTOCASSABLE~~, ~~IMMANGEABLE~~
1 ∪ 2 ∪ 3	ACCEPTABLE, MANGEABLE, FRIABLE, AUTOCASSABLE, IMMANGEABLE
1 ∪ 3	ACCEPTABLE, MANGEABLE, ~~FRIABLE~~, AUTOCASSABLE, IMMANGEABLE
1 ¬ 2	~~ACCEPTABLE~~, MANGEABLE, ~~FRIABLE~~, ~~AUTOCASSABLE~~, ~~IMMANGEABLE~~

Tableau 3 : Calculs de productivité différenciés.

– le premier calcul ne prend en considération que les lexèmes analysables en français, qu'ils soient ou non par ailleurs des emprunts (le point de vue est synchronique). Si l'on reprend l'ensemble des exemples en *-able* cités jusqu'ici, c'est, parmi eux, le cas d'ACCEPTABLE et de MANGEABLE,

– le deuxième réunit les lexèmes des catégories 1 et 2, conjoignant ainsi synchronie et diachronie (les lexèmes doublement marqués 1 et 2 ne sont comptés qu'une fois) : parmi nos cinq lexèmes en *-able* sont concernés ACCEPTABLE, MANGEABLE et FRIABLE,

– le troisième correspond à l'extension maximale (synchronie, diachronie et synchronie étendue) et prend en considération les catégories 1, 2 et 3 : les cinq adjectifs en *-able* cités sont ainsi retenus,

– nous avons également calculé la productivité de la suffixation en *-able* en ne prenant que les lexèmes marqués 1 et 3. Le point de vue est alors celui de l'analysabilité, directe ou non, en français : ACCEPTABLE, MANGEABLE, AUTOCASSABLE et IMMANGEABLE sont dans ce cas,

– enfin, il nous a paru intéressant d'effectuer également des calculs de productivité, en ne prenant en compte que les lexèmes marqués 1 qui ne soient pas en même temps des emprunts. On est alors dans le cas de la synchronie pure, et selon les procédés étudiés, ce calcul permet d'évaluer l'influence du lexique emprunté sur la vitalité du procédé. Ce calcul permet également en partie d'éviter l'écueil de compter comme hapax des séquences éventuellement entrées dans le lexique de longue date, et peu employées parce qu'archaïsantes. Parmi nos cinq adjectifs, seul est alors retenu MANGEABLE.

7.2 Quelques exemples

La figure 3 donne en ordonnée l'indice de productivité *P* de huit RCL du français, calculé dans la rubrique 'France' de l'année 1995 du quotidien *Le Monde*. En abscisse figure le nombre de types illustrant chaque RCL dans le corpus considéré.

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Morphologie
DOI 10.1051/cmlf08184

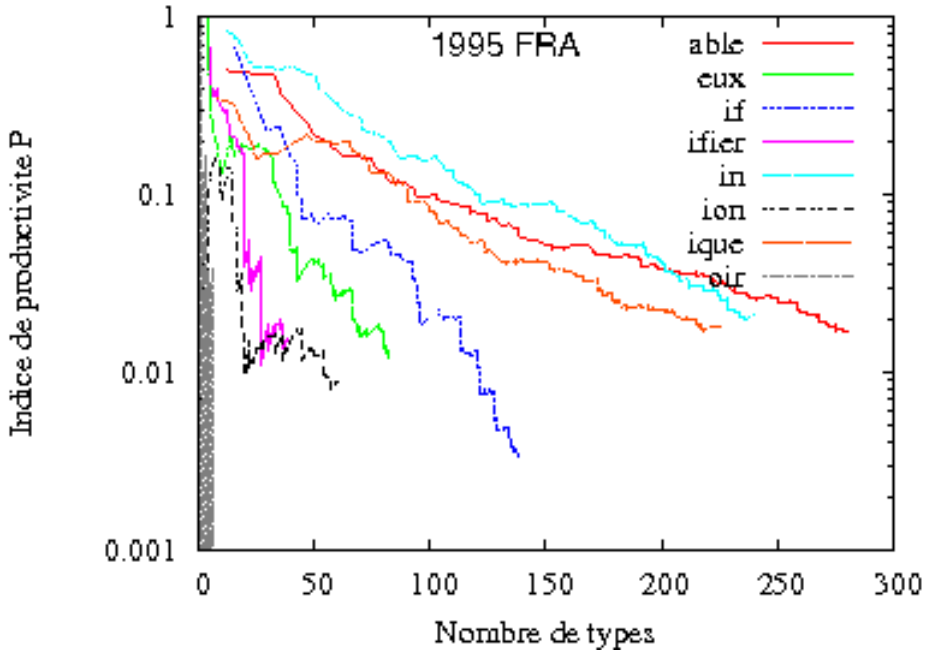


Figure 3 : Comparaison de la productivité des procédés étudiés.

La lecture du graphique suit donc l'évolution de ces deux valeurs : avec l'accroissement des valeurs de l'axe des abscisses, le vocabulaire d'un procédé augmente dans la rubrique considérée, tandis que les valeurs plus grandes sur l'axe des ordonnées marquent une potentialité plus forte dans la création de nouveaux lexèmes. L'évolution de ces valeurs est conforme à ce que l'on attend : avec le parcours de corpus, le vocabulaire d'un procédé augmente alors que son indice de productivité diminue. La capacité à garder un indice *P* aussi élevé que possible, tout en produisant de nouveaux lexèmes, marque une productivité élevée. C'est, par exemple, le cas de la suffixation en *-able* sur cette figure.

Ainsi, la projection des huit procédés dans cet espace permet de distinguer trois groupes :

– les suffixations en *-able* et *-ique*, la préfixation en *in-* ont les indices de productivité *P* les plus élevés, tout en étant présentes dans un nombre de types relativement important,

– les suffixations en *-if, -eux, -ion* et *-ifier* ont des indices de productivité moyens, et sont instanciées dans un nombre de type lui aussi moyen,

– la suffixation en *-oir(e)* est présente dans quelques types, mais n'est guère productive.

Les observations que nous faisons ici sur la rubrique 'France' de l'année 1995 *du Monde* se retrouvent dans d'autres rubriques de la même année ainsi que dans d'autres années du quotidien. Nous retiendrons donc un procédé de chacun de ces groupes (les suffixations en *-able,* en *-ion* et en *-oir(e)*) pour la présentation des résultats qui vont suivre.

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Morphologie
DOI 10.1051/cmlf08184

7.3 Productivité de la suffixation en *-able*

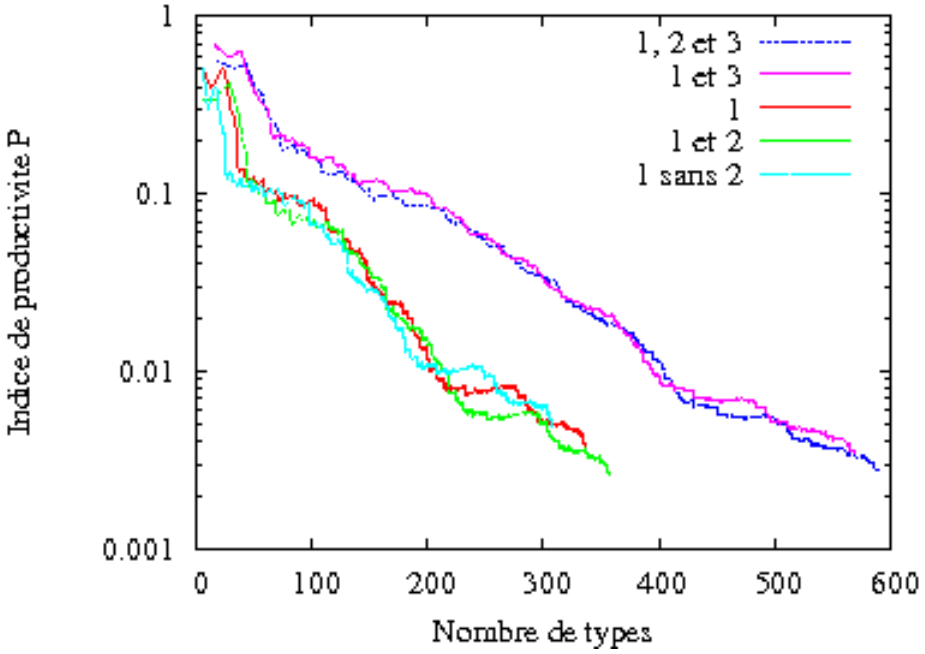


Figure 4 : Productivité différenciée de la suffixation en -able *dans les huit rubriques du Monde en 1995 et 1999.*

La figure 4 illustre la productivité de la suffixation en *-able* dans l'ensemble des rubriques étudiées, dans deux années *du Monde* (1995 et 1999). On y voit :

– que, quelles que soient les combinaisons retenues à partir des catégories 1, 2 et 3 de notre grille d'analyse, *P* décroît à mesure que l'on avance dans le corpus. Ce résultat est attendu : plus on avance dans le corpus, moins on a de chances de rencontrer d'hapax d'un type constructionnel donné,

– qu'en pourcentage, une part non négligeable de suffixés en *-able* relève en fait d'une analysabilité diachronique (comparaison traits turquoise et vert),

– que, pour la suffixation en *-able*, l'analysabilité indirecte (catégorie 3) joue un rôle prépondérant (traits turquoise, vert et rouge d'une part, rose et marine de l'autre). Ceci est dû essentiellement aux préfixés par *in-* de suffixés en *-able*. Les prendre en considération dans le calcul de *P* pour la suffixation en *-able* est d'autant plus légitime que beaucoup n'ont pas de positif en *-able* (sur ce point, cf. Dal *& al.*, 2007). On peut donc considérer qu'ils contribuent indirectement à la productivité de cette règle.

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Morphologie
DOI 10.1051/cmlf08184

7.4 Productivité de la suffixation en *-ion*

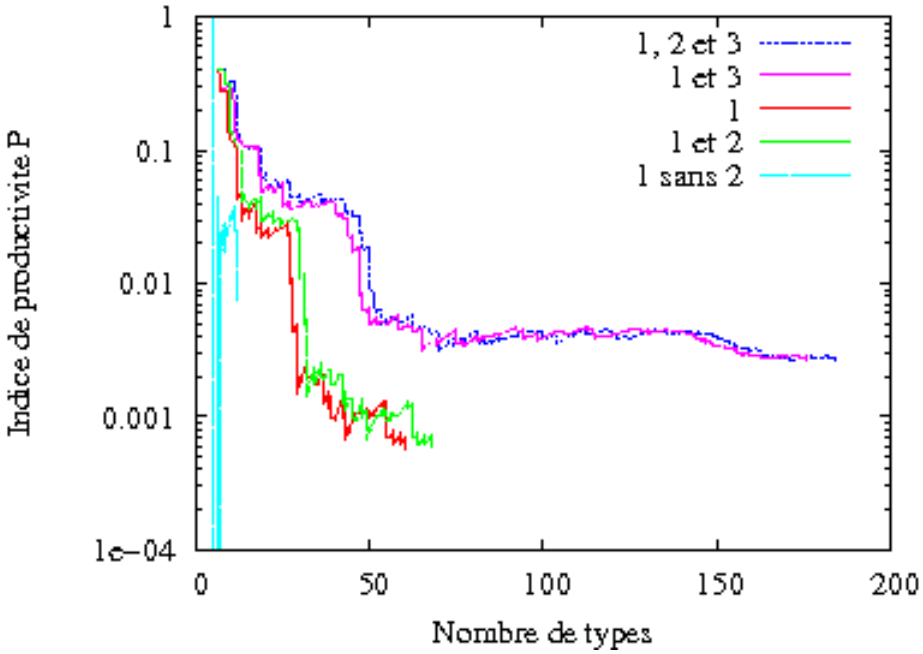


Figure 5 : Productivité différenciée de la suffixation en -ion *dans les huit rubriques du Monde en 1995 et 1999.*

De manière similaire à celle de la figure 4, la figure 5 illustre la productivité de la suffixation en *-ion* en fonction des catégories de la grille. Notons tout de suite que la valeur la plus basse de l'axe des ordonnées est 10 fois moins élevée que pour la suffixation en *-able* : en d'autres termes, la suffixation en *-ion* montre moins de potentialité à former de nouveaux lexèmes. On remarque aussi que, dans le même corpus, le vocabulaire en *-ion* est près de trois fois moins important que le vocabulaire en *-able*.

La figure 5 fait également apparaître que la part de l'analysabilité diachronique (catégorie 2) est plus prépondérante pour la suffixation en *-ion* qu'elle ne l'est pour la suffixation en *-able* (comparaison traits turquoise et vert). Ainsi, lorsque l'on adopte le point de vue de la synchronie pure (catégorie 1 sans 2), la courbe montre une chute vertigineuse d'abord, remonte ensuite, et s'arrête très vite : de ce point de vue, le procédé n'est que difficilement productif.. En revanche, une vision englobante, avec la prise en compte des catégories 1 à 3 ou 1 et 3, produit une vision assez productive de cette suffixation.

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Morphologie
DOI 10.1051/cmlf08184

7.5 Productivité de la suffixation en *-oir(e)*

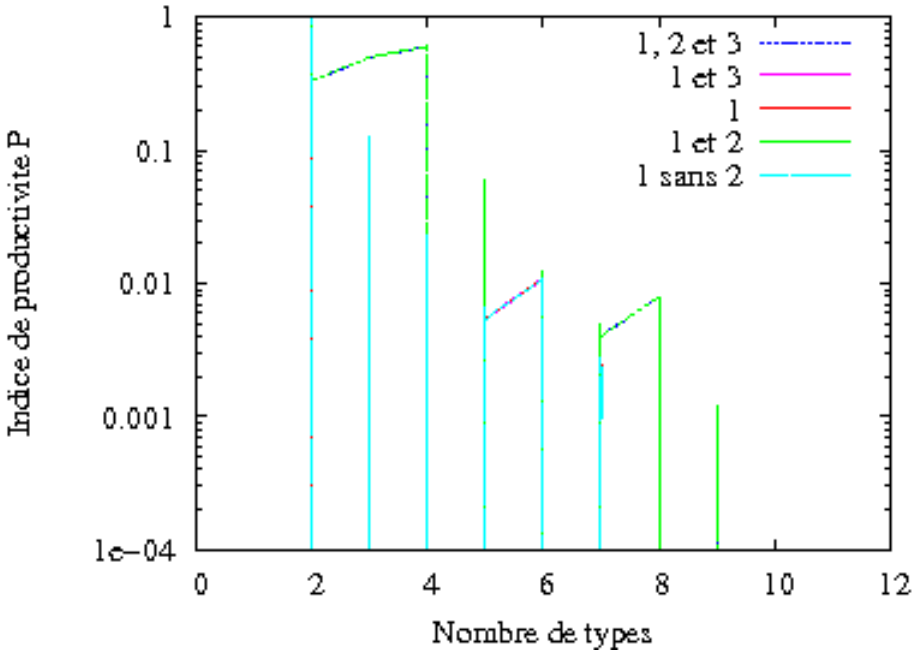


Figure 6 : Productivité différenciée de la suffixation en -oir(e) *dans les huit rubriques du Monde en 1995 et 1999.*

La figure 6 illustre la productivité de la suffixation en *-oir(e)* dans le corpus étudié. L'axe des ordonnées présente les mêmes valeurs que sur la figure 5. En revanche, la taille du vocabulaire est extrêmement faible, puisque, sur les deux années *du Monde* considérées, il ne compte que neuf lexèmes. La vue générale des courbes, quel que soit le point de vue adopté,, est caractéristique d'un procédé non productif. Nous pouvons voir en particulier leur forme « en escalier » qui témoigne de l'apparition non régulière de nouveaux lexèmes dans un corpus de plus de 23 millions d'occurrences.

Ce résultat est conforme aux analyses énoncées dans les grammaires et travaux de recherche (par ex. Meyer-Lübke, 1894, Nyrop, 1899-1930, Dubois, 1962, Corbin, 1987), selon lesquels on ne forme plus guère de noms en *-oir*. Il est corroboré par les données du *TLF*, qui atteste peu de noms en *–oir(e)* postérieurs au milieu du 20ème siècle.

8 Conclusion

Dans le travail présenté ici, nous étudions la productivité de plusieurs procédés constructionnels du français. Nous proposons en plus une réflexion poussée sur la notion d'analysabilité des lexèmes et établissons une grille d'analyse dédiée à cette tâche. L'originalité de ce travail consiste certainement à étudier l'impact de la prise en compte de différents degrés d'analysabilité sur le calcul de la productivité.

Nous avons ainsi observé la productivité de trois procédés qui montrent des comportements différents et semblent être représentatifs de trois classes : suffixations en *-able, -ion* et *-oir(e)*. Dans le corpus étudié, la suffixation en *-able* montre une productivité des plus importantes (taille du vocabulaire et indice de productivité *P* élevés), avec une différence notable entre les analysabilités diachronique et synchronique. Pour les suffixés en *-ion*, cette différence est encore plus grande. Finalement, la suffixation en *-oir(e)* ne compte que quelques occurrences sporadiques dans le corpus étudié et n'est pas productive. Sachant que

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Morphologie
DOI 10.1051/cmlf08184

le *TLF* recense bien plus de suffixés en *-oir(e)* (il en compte 283), nous supposons que *Le Monde* n'est sans doute pas le corpus idéal pour l'étude de ce procédé.

Cette dernière observation pose de façon aiguë la question de la représentativité du corpus. Si le quotidien *Le Monde* propose des volumes respectables de données et constitue un échantillon assez fidèle de la langue générale, nous devons cependant prendre en compte d'autres sources de corpus afin d'avoir une image plus complète des procédés du français. Par exemple, nous pouvons les étudier sur la base des articles du *Chasseur français*, certainement riche en lexèmes en *-oir(e)*.

Se pose aussi l'impact que peuvent avoir d'autres paramètres comme le registre discursif ou l'intention du locuteur. Toujours à propos de la suffixation en *-oir(e)*, (Namer & Villoing, ce volume) ont mis sur pied une enquête menée sur la Toile, visant à en extraire l'ensemble des noms en *-oir(e)* absents du *TLF*, ainsi que leurs contextes d'utilisation. Une fois éliminés les résultats parasites, elles ont ainsi collecté 304 nouveaux noms, et, conformément à Sablayrolles (2000) et Roché (à paraître), ont dégagé deux facteurs favorisant ces créations lexicales :

– les besoins énonciatifs. Une partie des créations a la même capacité référentielle que des dénominations existantes, et en constituent des doublons, volontaires ou accidentels. C'est par exemple le cas dans les noms renvoyant à des forums de discussion et à l'activité qui s'y pratique (par exemple : *baratinoir*, *critiquoir*, *déclamoir*). Les créations peuvent avoir une visée humoristique (par exemple, *appeloir* pour 'téléphone', *enfournoir* pour 'bouche'), ou correspondre à des variations diatopiques (par exemple, *flashoir* pour nommer des radars routiers en Belgique, *shootoir* pour nommer un local aménagé pour les drogués dans certaines municipalités suisses) ;

– les besoins morphologiques. Quelques créations en *-oir(e)* sont des doublons de formes existantes, sans visée particulière sauf celle d'être conformes à la spécificité sémantique de la suffixation en *-oir*, qui forment des noms de lieux ou d'instruments : *pissoir* (pour *pissotière*) en est un exemple.

Ce survol rapide fait apparaître la nécessité de diversifier les types de ressources textuelles pour calculer la productivité, si l'on veut que les résultats obtenus aient une quelconque représentativité de la langue en général.

Références bibliographiques

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

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

Corbin, D. (1987). *Morphologie dérivationnelle et structuration du lexique*. 2 vols, Lille : Presses Universitaires du Septentrion. Edition originale, Tübingen : Niemeyer.

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

Dal, G., Grabar, N., Lignon S., Tribout, D. & Yvon, F. (2007). Les adjectifs en *inXable* du français. In Floricic, F. (éd.), *La négation dans les langues* romanes, Amsterdam/Philadelphia : John Benjamins, coll. « Investigationes Supplementa », 215-234.

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

Evert, S. & Lüdeling, A. (2001). Mesuring morphological productivity: Is automatic preprocessing sufficient?. In Rayson, P., Wilson, A., Mc Enery, T., Hardie, A. & Khoja, S. (éd.), *Proceedings of the Corpus Linguistics 2001 Conference*, 167-175.

Gaeta, L. & Ricca, D. (2003). Frequency and productivity in Italian derivation: A comparison between corpus-based and lexico-graphical data. *Italian Journal of Linguistics, 15-1*, 63-98.

Gaeta, L. & Ricca, D. (2006). Productivity in Italian word formation: a variable-corpus approcah. *Linguistics, 44-1*, 57-89.

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Morphologie
DOI 10.1051/cmlf08184

Grabar, N., Dal, G., Fradin, B., Hathout, N., Lignon, S., Namer, F., Plancq, C., Tribout, D., Yvon, F. & Zweigenbaum, P. (2006). Productivité quantitative de la suffixation par *-able* dans un corpus journalistique du français. In Viprey, J.-M. (éd.), *Peer-reviewed proceedings, Lexicometra : 8e Journées internationales d'Analyse statistique des Données Textuelles*, Besançon, 19-21 Avril 2006, 473-486.

Habert, B. (2000). Des corpus représentatifs: de quoi, pour quoi, comment ? In Bilger, M. (éd.), *Linguistique sur corpus. Etudes et réflexions*, Perpignan : Presses Universitaires de Perpignan, 11-58.

Habert, B., Nazarenko, A. & Salem, A. (1997). *Les linguistiques de corpus*. Paris : Armand Colin / Masson.

Meyer-Lübke, W. (1894). *Grammatik der romanischen Sprachen*. Leipzig : Reisland.

Namer, F. & Villoing, F. (ce volume). Interpréter les noms déverbaux : quelle relation avec la structure argumentale du verbe de base? Le cas des noms en *-OIR*. In Durand, J., Habert, B. & Laks, B. (éds), *Actes du premier Congrès mondial de linguistique française (CMLF-08), Paris, 9-12 juillet 2008*.

Nyrop, K. (1899-1930). *Grammaire Historique de la langue française vol. 3*. Copenhague / Paris : Gyldendal / Nordisk Forlag.

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

Roché, M. (à paraître). Propositions en morphologie lexicale. In *Carnets de grammaire (Rapports internes CLLE-ERSS)*. Toulouse: CLLE-ERSS.

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

Sablayrolles, J.-F. (2000). *La néologie en Français contemporain ; examen du concept et analyse de productions néologiques récentes.* Lexica. Paris : Champion.

Schultink, H. (1961). Produktiviteit als Morfologisch Fenomeen. *Forum der Letteren, 2*, 110-125.

---

[1] Le GDR, qui a débuté le 1e janvier 2000, était dirigé par Bernard Fradin. L'opération « Productivité morphologique » était coordonnée par Georgette Dal. Tous les auteurs du présent travail étaient membres de l'opération. Nous profitons de l'occasion qui nous est offerte pour remercier les collègues qui ont participé, pendant une période donnée, à l'opération. Nous leur sommes en partie redevables des résultats auxquels nous sommes parvenus.

[2] La définition qui précède est librement inspirée de Schultink (1961) : "Onder produktiviteit als morfologish fenomeen verstaan we dan de voor taalgebruikers bestaande mogelijkeid door middel van het morfologisch procédé dat aan de vorm-betekeniscorrespondentie van sommige hun bekende woorden ten grondslag, onopzettelijk een in principle niet telbaar aantal nieuwe formatives te vormen" ['Nous considérons que la productivité est un phénomène morphologique qui permet aux usages d'une langue de créer de façon non intensionnelle un nombre en principe infini de nouvelles formations, au moyen de procédures morphologiques qui se cachent derrière la correspondance entre la forme et le sens de mots connus' (notre traduction)]

[3] Pour plus de détail sur ce point, nous renvoyons entre autres à Gaeta & Ricca (2003) et (2006) et à Evert & Lüdeling (2001).

[4] On considère ici que le cas où n_1 et N croissent à la même vitesse est exceptionnel (il faudrait que chaque type qui incrémente le dénominateur soit lui-même un hapax).

[5] Le terme de *lemme* employé ici appartient au métalangage informatique. Il désigne la forme qui subsume toutes les variations de formes des mots du point de vue du traitement informatique. Les formes lemmatisées correspondent aux formes citationnelles qui se trouvent dans les dictionnaires. Par exemple, toutes les formes d'un verbe ont pour lemme son infinitif, ou encore CE sert de lemme à *ce*, *ces*, *cet*, *cette*. La lemmatisation est le processus permettant de ramener à un même lemme les variations qu'il peut connaître en corpus.

[6] La précision nous a été donnée par Séverine Casalis, que nous remercions.