Une approche statistique de la concurrence entre procédés constructionnels

La dérivation en -age et en -ment en français

Arthur Lapraye

Mémoire pour l'obtention du Master de Sciences du Langage spécialité Linguistique Informatique

Université Paris-Diderot - Paris VII

UFR de Linguistique

Directeur : **Olivier Bonami**

2017

Table des matières

Chapitre 1

Introduction

De multiples stratégies existent en français pour créer des nominalisations verbales, autrement dit des noms dérivés de verbes, dont le sens basique ou originel désigne en général l'action ou le procès décrit par le verbe, mais qui sont sujets à différentes formes de polysémie, systématique ou non-systématique (cf [Huyghe, 2014]).
Les principales stratégies productives sont la dérivation en *-ion* (par ex. *former-formation*), la dérivation en *-age* (par ex. *bavarder-bavardage*), la dérivation en *-ment* (*bombarder-bombardement*), et enfin la conversion (*envoler-envol*).
D'autre procédés existent mais sont plus marginaux comme la dérivation en *-ure* (*souder-soudure*) ou en *-ade* (*promener-promenade*).

Le nombre de noms attestés dans le lexique Démonette [Hathout et Namer, 2014] pour chaque type de suffixation donne une première idée de la productivité de ces stratégies :

Table 1.1 – Nombre d'attestations pour 5 suffixes dans le lexique Démonette

Suffixe	Attestations
-ure	131
-ade	127
-ment	2488
-age	2939
-tion	2731

À ces chiffres, il faut rajouter 1 101 noms de Démonette qui sont dans une relation de conversion avec des verbes, et on peut également citer le nombre de 3 241 paires de conversion nom-verbe relevées par [Tribout, 2010]. Comme il est généralement impossible de déterminer parmi ces conversions lequel des éléments est dérivé de l'autre, la comparaison avec les autres processus de dérivation est difficile, d'autant plus que la nature du processus de conversion rend ardue sa simple quantification.

Pour les autres formes de dérivation, les indices de productivité introduits par [Baayen, 1993] et calculés sur le corpus FRWIKI (comprenant 279 735 153 tokens dont 6 194 hapax) sont présentés dans le tableau 1.2.

TABLE 1.2 – Indices de Baayen calculés sur FRWIKI

	Hapax	P	$P*$
-ment	200	1.5505×10^{-4}	0.03228
-age	300	3.4890×10^{-4}	0.04843
-ion	178	0.3742×10^{-4}	0.02873
-ation	146	0.7477×10^{-4}	0.02357
-ade	12	1.04751×10^{-4}	0.00193

L'indice $P*$ de Baayen, le *hapax-conditioned degree of productivity* représente la proportion de mots appartenant à une catégorie morphologique donnée parmi l'ensemble des hapax d'un corpus, et est interprété comme une une estimation de la probabilité que le prochain hapax appartienne à la catégorie considérée. Nous voyons ici que les dérivations en *-age* et en *-ment* sont les plus productives dans le corpus FRWIKI, parmi les procédés de nominalisation verbale.

Par opposition à la dérivation en *-ion* qui est associée à des bases savantes, les dérivations en *-age* et *-ment* sélectionnent le même type de radical.
Ayant le même sens, leur productivité les met en situation de concurrence morphologique.
Leur situation est d'autant plus remarquable que ces deux procédés donnent lieu à un nombre élevé de doublets, c'est à dire qu'une même base possède souvent un dérivé en *-age* et un en *-ment*. Le lexique Démonette compte ainsi 1000 paires de tels dérivés, alors qu'il n'y a que 31 paires de dérivés en *-age* et en *-ation* issus de la même base, et 52 pour *-ment* vs *-ation*.

Le premier chapitre est consacré à un état de l'art, tout d'abord des différentes théories liées à la notion de concurrence morphologique en général, puis des multiples études consacrées au cas spécifique des suffixes *-age* et *-ment*.
Le deuxième chapitre est une étude quantitative de l'impact des traits des bases verbales sur le choix parmi les deux types de dérivation.
Enfin, le troisième chapitre examine les utilisations concrètes des doublets dans un large corpus et leurs corrélations avec certains traits syntaxiques.

Chapitre 2

État de l'art

2.1 La question de la concurrence en morphologie constructionnelle

La concurrence entre types de dérivation est étudiée depuis [Aronoff, 1976] qui examine les différences entre deux manières de dériver des noms abstraits de propriété à partir d'adjectifs en anglais : la dérivation en *-ness* et la dérivation en *-ity*, en particulier à partir d'adjectifs en *-ous*.
Aronoff note plusieurs différences entre les deux procédés de dérivation.
D'abord, une différence sémantique dans l'interprétation des noms déadjectivaux en *-ness*, qui ont, selon lui, toujours trois sens, liés de façon stable, pour un adjectif donné X : *the fact that Y is X*, *the extent to which Y is X*, *the quality or state of being X* tandis que les déadjectivaux en *-ity* seraient souvent plus polysémiques, voire ne pourraient pas prendre l'un de ces trois sens.
Ensuite une différence morphologique : la base sélectionnée pour la dérivation en -ness est identique à l'adjectif concerné (par ex. *callous- callousness*) tandis que la dérivation en *-ity* sélectionne un allomorphe, qui n'est pas toujours prévisible (par ex. *luminous-luminosity* vs *various-variety*).
Enfin une différence liée à leur présence dans le lexique et à la possibilité de leur génération en ligne, cette dernière n'étant systématiquement possible selon lui que pour le suffixe *-ness*.

La notion de *concurrence* découle du fait que l'existence d'un mot gêne fortement la création et l'entrée dans le lexique de synonymes directs de ce mot. Ce phénomène, le *blocage synonymique* a pour conséquence que si deux suffixes ont le même sens, la combinaison de l'un des deux avec une base constitue un obstacle fort contre l'utilisation de l'autre suffixe avec la même base, car le mot ainsi dérivé serait un synonyme.
Selon Plag dans son article sur les dérivations verbales en *-ize*, *-ify* et *-ate* [Plag, 1999], le blocage synonymique n'intervient pas au niveau des procédés de dérivation mais bien au niveau du lexique, et plusieurs suffixes synonymes peuvent coexister comme procédés de dérivation productifs.

Le mécanisme de ce blocage tel que le décrit [Rainer, 2012] implique que, tandis que des mots possibles non-attestés peuvent avoir une présence latente dans le lexique qui leur permet de devenir la base de dérivés attestés, une telle chose est impossible pour un dérivé potentiel qui subirait un blocage synonymique : la dérivation d'un mot par un suffixe non seulement empêche la dérivation par un suffixe synonyme sur la même base, mais raye cette possibilité de dérivation de la liste des mots possibles.

[Lindsay et Aronoff, 2013] présentent une théorie de la concurrence entre suffixes dérivationnels montrant comment deux procédés de dérivation concurrents peuvent rester chacun productifs.
Un suffixe est généralement d'autant moins productif qu'il est sélectif, puisque la sélection limite le nombre de bases auxquelles il peut s'associer.
Cependant, le plus sélectif de deux suffixes synonymes peut néanmoins rester productif s'il sélectionne des bases ayant des propriétés suffisamment distinctives. C'est d'autant plus vrai si cet ensemble de bases est susceptible de croître par l'utilisation d'autres procédés de dérivation productifs. Les deux suffixes synonymes se retrouvent alors en distribution complémentaire.
Il en va ainsi en anglais de la dérivation adjectivale en *-ic* et en *-ical* :
La dérivation en *-ic* domine globalement sur l'ensemble des noms mais la dérivation en *-ical* reste productive en sélectionnant les bases se finissant en *-ology*. Le fait que ces bases sont elles-mêmes issues d'un processus de dérivation productif servant à former des noms de science, permet à *-ical* de rester productif.
Lindsay & Aronoff s'intéressent également à la concurrence entre les suffixes de dérivation verbale *-ize* et *-ify*. Étudiant la répartition des dérivés selon le nombre de syllabes de la base, ils montrent que plus la base est longue, plus le suffixe *-ize* est préféré, au point que *-ify*, majoritaire parmi les bases monosyllabiques, ne sélectionne jamais de base de 3 syllabes ou plus. Cependant le suffixe *-ize* sélectionne quelques bases monosyllabiques lui aussi, ce qui montre que cette préférence est de nature graduelle, probabiliste.

Enfin, [Arndt-Lappe, 2014] présente une tentative de quantifier la rivalité entre dérivations en *-ity* et dérivation en *-ness* et en particulier de modéliser l'importance de l'analogie dans la productivité de ces deux suffixes : l'idée est, comme chez Lindsay & Aronoff, que le choix de la dérivation est motivé par la forme de la base, en particulier si cette base est elle-même morphologiquement composée, ce qui divise l'ensemble des bases potentielles des deux suffixes en différents groupes selon leurs suffixes : En l'occurrence, les adjectifs en *-able*, *-ic*, *-ed*, *-y*, *-ous*, *-ish*, *-ing*, *-ive*, *-less* et ceux sans suffixe identifiable. Arndt-Lappe montre que la répartition des bases entre les deux suffixes n'aboutit pas à une distribution complémentaire stricte selon ces groupes, mais dégage plutôt des tendances : si certains groupes morphologiques de bases adjectivales sont entièrement accaparés par l'un des deux suffixes (par ex. *-able* pour *-ity*, *-y* pour *-ness*) , pour d'autres, les deux sont présents à l'un ou l'autre degré, la catégorie *-ous* étant celle où les deux suffixes sélectionnent approximativement la moitié des bases.

La concurrence entre *-age* et *-ment* paraît cependant difficilement explicable selon les critères développés par Lindsay & Aronoff ou par Arndt-Lappe, étant donné le très grand nombre de bases pour lesquelles les deux types de dérivés sont attestés : 1000 sur l'ensemble des formes attestées dans Démonette, tandis que *-age* et *-ment* ne partagent chacun avec *-ion* qu'une cinquantaine de bases. Un tel recouvrement montre que les deux suffixes ne sont pas en distribution complémentaire l'un par rapport à l'autre, et qu'il est difficile de délimiter des zones du lexique qui seraient dévolues entièrement à l'un ou à l'autre en utilisant les traits morphologiques employés par Aronoff et Arndt-Lappe.
L'existence de tant de paires de doublets en dépit du blocage synonymique ne semble pouvoir s'expliquer que s'il existe également de vraies différences d'interprétations sémantiques entre les deux types de dérivation.

2.2 Les théories sur la concurrence entre *-age* et *-ment*

La question de ce qui différencie la formation de noms déverbaux en *-age* et *-ment* a été étudiée de longue date et a notamment été abordée sous l'angle de la différence d'interprétation sémantique des dérivés.
[Debaty-Lucas, 1986] formule l'hypothèse que la dérivation en *-age* ou en *-ment* est l'expression d'un unique morphème d'action (« *suffixème* ») sous-jacent, donc qu'il n'existe pas de différence sémantique entre les deux.
Selon [Dubois et Dubois-Charlier, 1999] les dérivés en *-age* proviennent de verbes transitifs et ont un sens d'action (par ex. *affuter-affutage*) ou de résultat de l'action (par ex. *atteler-attelage*), mais peuvent aussi être dérivés de verbes intransitifs ayant un sens de communication ou psychologiques (*vagabondage, bafouillage*), tandis que les dérivés en *-ment* se forment plutôt à partir de verbes intransitifs ou pronominaux et ont un sens d'état plutôt que d'action (*alanguir-alanguissement*, *affaisser-affaissement*), mais peuvent également être dérivés de transitifs factitifs (*décourager-découragement*).
Selon eux, lorsqu'il existe pour une même base un dérivé en *-age* et un en *-ment*, cela s'explique d'abord par la différence syntaxique entre *-age* issu d'un emploi transitif direct ou intransitif mais avec un sens d'action tandis que *-ment* est issu d'un emploi intransitif (en y incluant les réflexifs) ou passif (être + vpp) avec un sens de résultat ou d'objet concret. Par exemple, *appontage* (action) vs *appontement* (objet).
Il peut exister également une différence au niveau du complément de nom en de :
Pour *-age*, c'est le complément d'objet inanimé du verbe transitif tandis que celui de *-ment* est le sujet d'un verbe intransitif ou d'un passif être + vpp (*abattage* vs *abattement*, *attendrissage* vs *attendrissement*).

Cependant il existe un certain nombre de contre-exemples : par exemple dans l'opposition *échafaudage-échafaudement* , *échafaudage* a dans la majorité des cas un sens d'objet concret, résultat de l'action décrite par le verbe-base, tandis qu'*échafaudement* n'a jamais un tel sens mais désigne systématiquement l'action elle-même.

Dubois & Dubois-Charlier mentionnent également qu'il peut exister une différence de domaine d'emploi pragmatique entre les deux dérivés : on parlera de « l'arrosage du jardin » mais de « l'arrosement d'une céramique ».
Enfin, ils mentionnent qu'il peut exister une grande différence sémantique entre les deux dérivés si elle résulte d'une différence de sens entre emploi transitif et intransitif d'un même verbe : *barbotage* serait lié au sens transitif du verbe *barboter*, synonyme de voler, tandis que *barbotement* serait lié au sens intransitif de ce même verbe, alors synonyme de *patauger*.
Cependant, une vérification dans le corpus frWAC montre de nombreuses attestations de *barbotage* dans le sens de l'action de *patauger, barboter (intr)*.

Dans [Kelling, 2001], Carmen Kelling remet en cause plusieurs théories précédentes de la concurrence entre ces deux formes de dérivation.
Outre les théories de [Debaty-Lucas, 1986] et de [Dubois, 1962] [Dubois et Dubois-Charlier, 1999] déjà citées, Kelling critique celle de [Lüdtke, 1978].
Lüdtke considère comme Dubois que la transitivité est le trait déterminant, favorisant la dérivation en *-age*, tandis que *-ment* sélectionnerait les verbes « *intransitifs, réflexifs et passivisés* », et qu'il en résulte une différence d'interprétation des doublets.
Par exemple, l'opposition entre *battage* et *battement*, viendrait de ce que le premier est dérivé du sens transitif de *battre*, tandis que *battement* serait dérivé d'une version intransitive de *battre*, celle qui se retrouve dans les phrases comme « mon coeur bat ».
Mais Kelling présente plusieurs contre-exemples comme *furetage*, dérivé d'un intransitif, et *essouflement*, dérivé d'un transitif, qui contredisent cette hypothèse.

Kelling présente ensuite sa théorie : selon elle, la dérivation en *-age* est d'autant plus favorisée que le verbe possède un argument sujet prototypiquement agentif, et, lorsqu'il existe un doublet, l'interprétation du dérivé en *-age* est systématiquement liée à un sens plus agentif du verbe que celle du dérivé en *-ment*.
L'agentivité d'un participant est définie ici comme une propriété graduelle définie selon cinq critères énoncés dans [Dowty, 1991] :

- L'implication volontaire dans l'événement.
- La conscience de l'événement.
- Le fait de causer événement ou changement d'état chez un autre participant.
- Le mouvement.
- L'existence indépendante du verbe et de l'événement nommé par le verbe.

Plus un argument du verbe correspond à ces critères, plus il est vu comme agentif et par extension plus le verbe lui-même peut être qualifié d'agentif.

Ainsi, la différence d'interprétation entre *étirage* et *étirement* est expliquée par le fait que le premier serait lié au sens transitif d'*étirer*, où ce verbe prend un argument agent et un argument patient, réunissant donc tout les critères de Dowty sauf le quatrième. A contrario, *étirement* serait lié au sens d'*étirer* pris comme verbe réflexif, donc dénué du troisième critère, donc moins agentif.

Kelling pointe, de plus, que l'évolution des deux suffixes est différente.
Selon elle, malgré l'attestation des premiers déverbaux en *-age* au XIII^{e} et XIV^{e} siècle (par exemple *chauffage*), ce n'est qu'au XIX^{e} siècle qu'il devient réellement productif, tandis que *-ment* est productif en tant que suffixe déverbal dès le latin.
Elle mentionne également que des règles de dissimilation phonologiques jouent un rôle, notamment pour empêcher l'apparition de mots ayant deux /ʒ/ à la suite comme dans **saccageage*.
Enfin, elle formule l'hypothèse que certains domaines spécialisés du vocabulaire ont une influence pragmatique qui contrecarre la contrainte d'agentivité, en donnant l'exemple du domaine économique, où la dérivation en *-ment* dominerait au point de susciter l'existence du dérivé *intéressement* plutôt qu'**intéressage*.

Dans ses articles [Fradin, 2014] et [Fradin, 2017], Fradin examine les différences sémantique dans l'interprétation des dérivés en *-age* et en *-ment* issus de la même base.
Il formule l'hypothèse que, par défaut, la dérivation en *-age* est liée à des constructions requérant un contrôle par un agent et donne lieu à des noms ayant un sens événementiel, tandis que les noms en *-ment* seraient corrélés avec des bases ne requérant pas de contrôle d'un agent, et donnent lieu à des interprétations plus variables. Ainsi, si une même base donne lieu à ces deux dérivations, il existera en général un contraste sémantique ou pragmatique dans l'interprétation des dérivés, qui seront de préférence associés à des contexte d'utilisation différents ou à deux sens différents de leur base.
Par exemple « le rasage des aisselles » et « le rasement de la citadelle » correspondent respectivement à un sens littéral et à un sens métaphorique du verbe-base *raser*.
Cependant, à la différence de Dubois, ces différences sont pour Fradin de nature plutôt tendancielle que catégorielle. Il relève comme contre-exemple possible les dérivés *pavage* et *pavement* qui peuvent tout deux désigner aussi bien l'action de paver que son résultat.
Les nuances relevées par Fradin ne semblent pas forcément très nettes en pratique : L'exemple des deux dérivés du verbe rhabiller montrant une différence de contexte entre « le rhabillage des meules » et « le rhabillement des enfants » est nuancé par le fait que « rhabillage des enfants » est attesté, et la difficulté de trouver un exemple pour « rhabillement des meules » est vraisemblablement liée à la rareté du mot *meule* lui-même, comparativement avec le mot *enfant*.
On peut également noter, a contrario, que la paire *pavage-pavement* n'est pas strictement synonyme étant donné que, des deux, seul *pavage* connaît une utilisation abondante dans le domaine mathématique, ce qui semble corroborer en partie les hypothèse de [Kelling, 2001] et [Dubois et Dubois-Charlier, 1999] concernant l'influence pragmatique possible des domaines spécialisés sur le choix entre les deux suffixes.

Ces différents articles montrent que le choix d'une dérivation en *-age* ou en *-ment* ne relève visiblement pas d'un unique critère déterminant, mais plutôt d'un ensemble de critères, dont le plus examiné est un critère sémantique lié à la notion d'agentivité, d'action volontaire ou de contrôle, peut-être lié à un critère syntaxique de transitivité.
Le rôle de la phonotactique est évoqué comme pouvant contrecarrer ces préférences en empêchant les dérivations faisant apparaître deux /ʒ/ à la suite et favorisant donc la dérivation en *-ment* pour les bases se finissant par cette consonne.
Enfin les études sur la concurrence en général montrent que le nombre de syllabes (et plus généralement la prosodie) jouent souvent un rôle dans le choix du suffixe.
Il semble de plus que la préférence entre les deux suffixe ait changé au fil du temps pour favoriser la dérivation en *-age*.

La nature tendancielle des préférences qui existent en matière de création de noms déverbaux en *-age* et en *-ment*, l'existence notamment de nombreux doublets attestés dans les lexiques, en dépit du phénomène de blocage synonymique, rendent nécessaire l'emploi de méthodes quantitatives afin de tester à plus grande échelle les théories sur le choix de ces deux suffixes et de permettre de modéliser plus en détail les modalités de leur concurrence.

Chapitre 3

La concurrence entre *-age* et *-ment* dans le lexique

La question du choix entre deux suffixes de même sens semble avant tout être liée à des traits lexicaux de la base de dérivation, et l'étude de la concurrence entre *-age* et *-ment* doit commencer par l'examen des traits des verbes utilisés pour créer les deux types de dérivés.

3.1 Croisement des données

Afin d'étudier les différents traits susceptibles d'influencer dans le lexique l'existence de noms déverbaux pour l'un ou l'autre suffixe, nous avons constitué une base de verbes associés à une série de traits potentiellement pertinents, en combinant plusieurs ressources pré-existantes.

Démonette [Hathout et Namer, 2014], une base de 31 204 mots du français en relation morphologique, constituée notamment à partir du TLFnome et de Verbaction, et contenant des annotations sémantiques, a servi de base pour constituer une liste de 4258 verbes ayant soit un dérivé en *-age*, soit un en *-ment*, soit les deux à la fois.
Les noms finissant orthographiquement par « age » ou « ment » et paraphrasés comme « action de X » ont été relevés dans Démonette, ainsi que les verbes leur correspondant.
De ces paires verbes-nom ont été retirées les quelques paires où le verbe est dans une relation de conversion avec le nom (par ex. *boniment-bonimenter* ou *partage-partager*). Cette sélection a été faite semi-manuellement en recherchant les finales orthographiques en « ager » et « menter » dans les relevés.

Ensuite, les informations relatives à la phonologie et à la fréquence ont été extraites du Gros Lexique À tout Faire du Français, [Hathout *et al.*, 2014], un lexique de 1 406 857 mots-formes fabriqué à partir du Wiktionnaire, et comportant pour chaque forme une transcription phonémique, des informations de flexion (telles que temps, mode, personne de conjugaison) ainsi que des fréquences calculées sur trois corpus différents dont frWAC [Ferraresi *et al.*, 2010] un corpus de 1 328 628 428 mots constitué à partir d'un grand nombre de sites internet francophones.

On a considéré que le radical de dérivation pour *-age* et aussi pour *-ment* était le radical du participe présent.
C'est donc la transcription phonémique donnée par le GLÀFF pour cette forme qui a servi a extraire les traits de phonèmes finaux. Le trait de nombre de syllabe a également été calculé sur la base de cette transcription et, en cas de possibilités multiples de syllabations, c'est la plus longue qui a été systématiquement retenue.

Les transcriptions du GLÀFF ont également été utilisées pour classer les verbes selon les trois groupes de conjugaison du français :
Les verbes dont l'infinitif a une terminaison orthographique en « er » sont étiquetés comme appartenant au premier groupe, ceux dont le participe présent a la terminaison phonémique /i.sɑ̃/ et dont l'infinitif a la terminaison orthographique « ir » sont étiquetés comme appartenant au deuxième groupe, et tous les autres verbes présents dans le GLÀFF sont par défaut assignés au troisième groupe.

Les informations de fréquences de formes contenues dans le GLÀFF sont incluses dans le jeu de données et, en particulier, les fréquences relatives (par million de mots) de la forme catégorisée dans le corpus frWAC servent à calculer les divers ratio de fréquences de formes finies du verbe.

Les informations de sous-catégorisation ont été extraites du LEFFF [Sagot, 2010], un lexique de 1 197 649 formes fléchies (dont 967 302 formes verbales) contenant notamment des informations de sous-catégorisation pour chaque forme conjuguée de chaque verbe.
Pour chaque fonction syntaxique, on a considéré qu'elle était sélectionnée par un verbe si au moins l'une des formes de ce verbe l'admettait. Ainsi, si un verbe admet un objet indirect en *de* pour au moins un de ses sens et une de ses formes, la variable correspondante dans le jeu de donnée aura la valeur booléenne "vrai".

Le même raisonnement a été employé pour les informations de sélection de rôles thématiques tirées de Verb∋Net [Danlos *et al.*, 2016], lexique de 2 814 verbes du français organisés selon une base sémantique et regroupés selon les rôles thématiques de leurs arguments. Si un verbe est présent dans une *frame* de Verb∋Net associé à un rôle thématique, par exemple un rôle d'agent, le trait correspondant du jeu de données, portant le même nom, aura la valeur booléenne « vrai » et sinon la valeur booléenne « faux ».

Enfin Google Ngrams [Michel *et al.*, 2011], un corpus de plus de 155 milliards de mots, constitué via la numérisation de millions d'ouvrages et recouvrant une très large période temporelle, a servi a calculer les dates de première attestation des noms déverbaux.
Cependant ce corpus est très bruité car, en raison de la façon dont il a été construit, des erreurs peuvent se retrouver à tout les niveaux, aussi bien pour la reconnaissance optique des caractères que pour des erreurs d'annotation de la date de publication, qui n'est pas toujours enregistrée correctement ou correspond parfois à une réédition, et sa taille même empêche la moindre vérification manuelle des données.

Étant donnée la fiabilité inégale de Google Ngrams, pour limiter le bruit, une méthode qui consiste à prendre le milieu de la première fenêtre de N années successives pour lesquelles Ngrams indique au moins X occurrences a été utilisée. Plusieurs estimations de la première année d'attestation ont été calculées selon cette méthode pour différentes valeurs de N et de X.

Ces différentes ressources ne se recouvrant que partiellement, beaucoup de verbes présentent des valeurs manquantes pour au moins une partie des traits utilisés. Par exemple, sur les 4258 verbes relevés dans Démonette, seuls 1603 sont également présents dans Verb∋Net.
Les autres verbes ont donc des valeurs manquantes pour les traits booléens définis à partir des rôles thématiques définis dans Verb∋Net.
Seuls 1023 verbes ne présentent aucune valeur manquante. A contrario, 5 des verbes relevés dans Démonette ne sont présents dans aucune autre ressource (*abatre, bêtir, entr'ouvrir, préapprendre* et *blétir*), et ont donc des valeurs manquantes pour chacun des traits.
La figure 3.1 (dont le tableau 3.1 donne la légende) montre le recouvrement de chacun des trait en abscisse, c'est à dire le nombre de verbe pour lesquelles ce trait a une valeur définie, et présente les traits définis par le jeu de données. Les traits issus du GLÀFF sont ceux qui ont la couverture la plus large, puis les traits de valence issus du LEFFF. Enfin, les traits de rôles thématiques tirés de Verb∋Net sont ceux qui ont la couverture la plus étroite.

FIGURE 3.1 – Nombre de valeurs présentes par traits

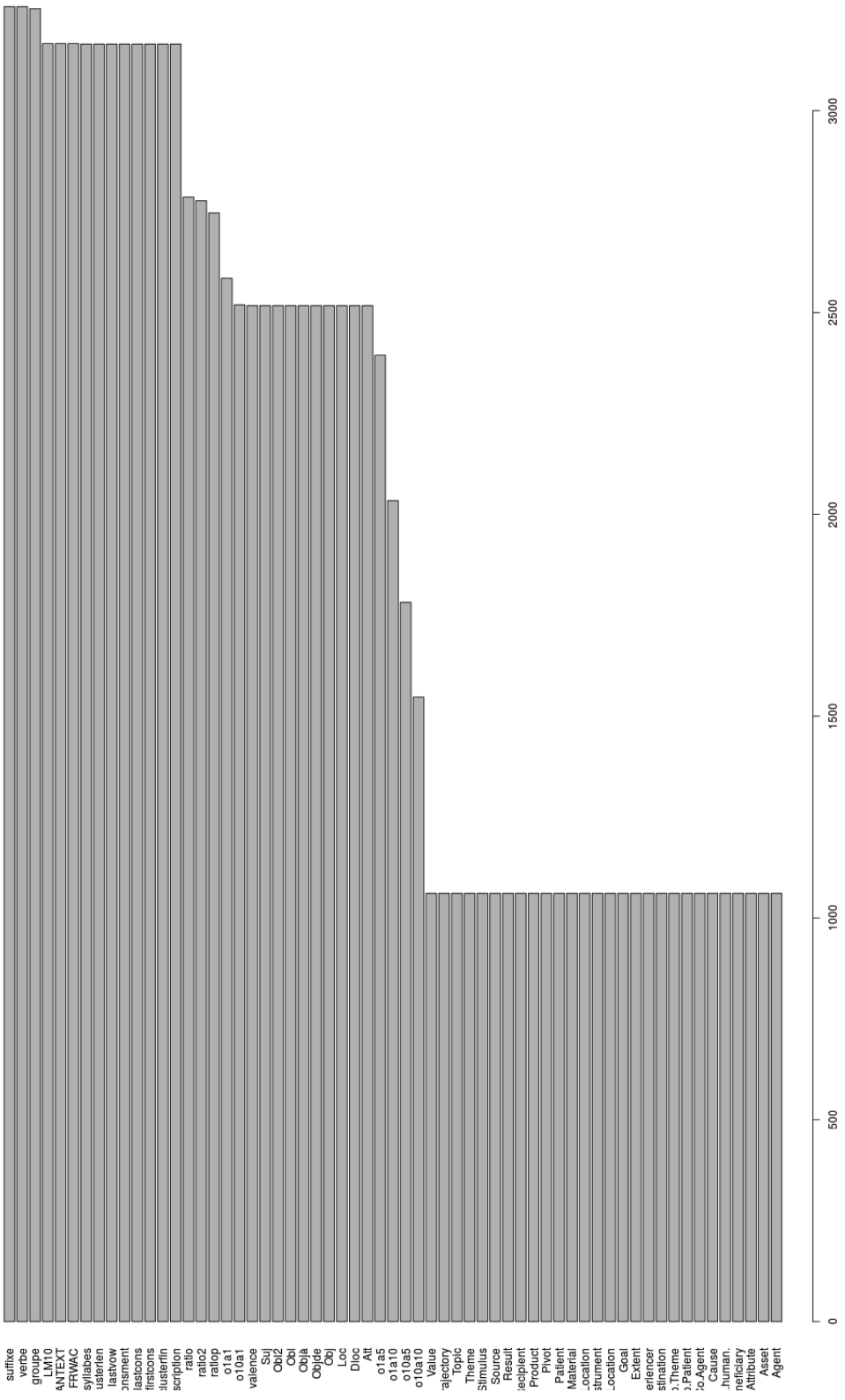


TABLE 3.1 – Légende des traits utilisés

verbe	Le verbe tiré des relevés dans Démonette
suffixe	Le suffixe associé au dérivé du verbe
transcription	La transcription du verbe au participe présent
groupe	Le groupe du verbe considéré
clusterfin	Le groupe de dernières consonnes du participe présent (attaque de la dernière syllabe d'après la syllabation du GLÀFF)
firstcons	La première consonne de clusterfin
lastcons	Dernière consonne de clusterfin
consment	la consonne appartient au groupe de trois consonnes [ʒ ɲs]
lastvow	Dernière voyelle de la base
clusterlen	Longueur de clusterfin
nbsyllabes	Nombre de syllabes au participe présent
frWAC	Fréquence du lemme dans le corpus frWAC
ratio ratio2 ratiop	Proportion de fréquence des formes des deux premières personnes parmi les formes conjuguées
FRANTEXT	Fréquence du lemme verbal dans FRANTEXT
LM10	Fréquence du lemme verbal dans LM10
valence	Nombre total d'arguments rattaché au verbe (incluant le sujet)
Att Dloc Loc Obj Objde Objà Obl Obl2 Suj	Traits de sous-catégorisation booléens tirés du lefff.
Agent Asset Cause Experiencer Goal Instrument Patient Stimulus Theme Topic	Traits de rôles sémantique tirés de Verb∋Net.
o1a1 o10a1 o1a5 o1a10 o10a5 o10a10	Traits de première attestation, où oNaM renvoie le milieu des premières M années telles que pour chaque année il y a au moins N occurrences du dérivé dans Google Ngrams.

3.2 Statistiques descriptives

La section qui suit présente un ensemble de statistiques descriptives, calculées sur l'ensemble des verbes des données pour différents traits susceptibles d'avoir une influence sur le choix d'une stratégie de dérivation, et les répartitions de ces verbes suivant qu'ils sont associés dans Démonette à un dérivé en *-age*, un dérivé en *-ment*, ou bien les deux à la fois.

Étant donné que chaque variable n'est pas toujours définie pour l'ensemble des verbes, le nombre total de verbes sur lesquelles ces statistiques sont calculées varie d'un tableau à l'autre et est indiqué à chaque fois.

3.2.1 Répartition par dernière consonne de la base

Le premier trait que nous examinons est un trait phonotactique, la dernière consonne de la base. Nous montrons ici la répartition des verbes de nos données pour chaque consonne finale et chaque type de dérivation.

TABLE 3.2 – Répartition des dérivés selon la consonne finale du radical du participe présent

	age		ment		ment&age		Total
_	8	31%	8	31%	10	38%	26
ɲ	12	24%	27	55%	10	20%	49
ɥ	9	53%	7	41%	1	6%	17
ʃ	68	42%	39	24%	56	34%	163
ʁ	175	55%	74	23%	69	22%	318
b	17	46%	8	22%	12	32%	37
d	100	56%	38	21%	42	23%	180
f	13	43%	5	17%	12	40%	30
g	22	81%	0	0%	5	19%	27
j	150	38%	140	35%	110	28%	400
k	68	62%	22	20%	19	17%	109
l	182	42%	122	28%	132	30%	436
m	64	73%	18	20%	6	7%	88
n	263	44%	207	35%	129	22%	599
p	65	57%	24	21%	26	23%	115
s	120	20%	312	53%	158	27%	590
t	343	56%	155	25%	113	18%	611
v	31	43%	14	19%	27	38%	72
w	6	29%	5	24%	10	48%	21
z	66	44%	49	32%	36	24%	151
ʒ	20	16%	89	73%	13	11%	122
Total	1802	43%	1363	33%	997	24%	4162

On constate qu'en général, les dérivés en *-age* sont plus nombreux que ceux en *-ment*, à l'exception de trois consonnes finales : /ɲ/ /ʒ/ et /s/.
Pour les radicaux verbaux se finissant par ces trois consonnes, les dérivations en *-ment* sont bien plus fréquentes.

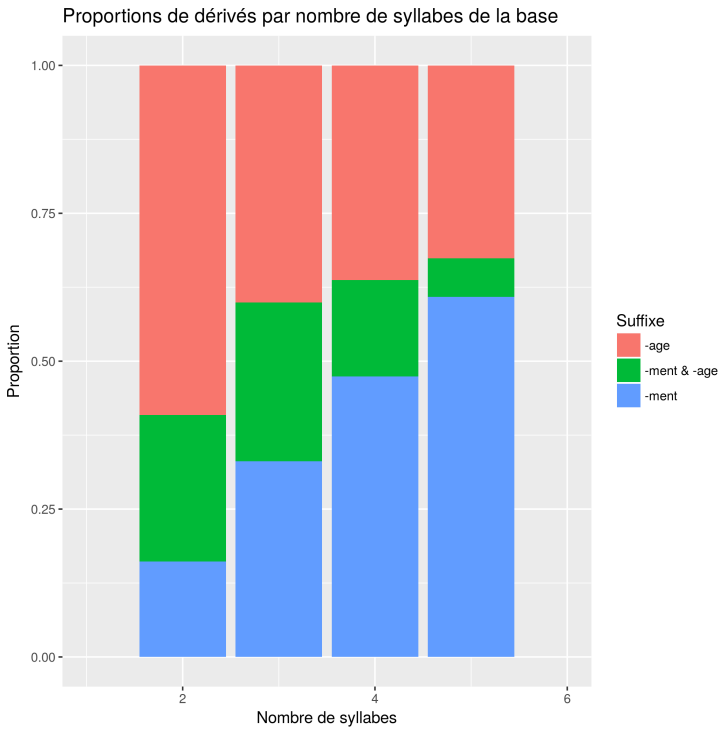
Le fait que ces phonèmes ne forment pas une classe naturelle conduit à l'hypothèse que c'est pour des raisons indépendantes qu'ils sont plus associés à la dérivation en *-ment*.
Spécifiquement, on peut faire l'hypothèse qu'une grande partie des radicaux en /s/ sont ceux des verbes du deuxième groupe, dont on verra plus bas qu'ils sont plus associés à la dérivation en *-ment*.
A contrario, le faible nombre de dérivations en *-age* pour /ʒ/ s'expliquerait plutôt par le fait d'une réticence phonotactique au fait d'avoir deux /ʒ/ à la suite, celui de la base et celui du suffixe, comme le supposait Kelling [Kelling, 2001], et on constate un effet symétrique pour le /m/ final, où la proportion de dérivations en *-ment* est plus particulièrement basse que pour la plupart des autres consonnes.
Quant à /ɲ/, la raison pour laquelle il est associé à un plus grand nombre de dérivations en *-ment* qu'en *-age* n'est pas claire, et n'est peut-être qu'un effet du hasard, compte tenu du petit nombre de verbe qu'elle concerne. Pour chaque consonne finale, le nombre de verbes ayant simultanément un dérivé en *-age* et un dérivé en *-ment* est généralement plus proche du plus petit des nombres de verbes ayant uniquement l'un des deux. Par exemple pour /l/ final, il existe 132 verbes ayant les deux dérivés, contre 122 ayant uniquement un dérivé en *-ment* et 182 ayant un dérivé en *-age*, tandis que pour /s/ final, on relève 158 verbes possédant les deux dérivés contre 120 ayant uniquement un dérivé en *-age* et 312 ayant uniquement un dérivé en *-ment*.
Néanmoins les verbes ayant un /v/ ou un /b/ final font exception à cette observation puisque pour les verbes en /v/, 27 verbes ont les deux dérivés, ce qui est plus proche du nombre de 31 verbes ayant un dérivé en *-age* que du nombre de verbes ayant un dérivé en *-ment*, qui est de 14.

3.2.2 Nombre de syllabes

Un autre trait susceptible d'influencer dans le choix d'une dérivation est un trait de nature prosodique, le nombre de syllabes. Ce trait peut jouer un rôle déterminant dans le choix d'un suffixe plutôt qu'un autre, ainsi que le montrent [Lindsay et Aronoff, 2013] pour *-ize* et *-ify* en anglais.

Nous montrons ici la répartition des dérivations selon le nombre de syllabes des verbes-bases, mesuré sur le radical du participe présent, qui est le radical utilisé pour la dérivation en *-age* et en général aussi pour celle en *-ment*.

FIGURE 3.2 – Proportions par longueur du verbe



Cette figure montre que le nombre de syllabes influe directement sur la préférence pour un type de dérivation : plus le verbe est long, plus la dérivation en *-ment* est préférée et moins la dérivation en *-age* est tolérée. Le nombre des dérivés en *-ment* est nettement inférieur à ceux des dérivés en *-age* en dessous de 4 syllabes, et à partir de 4 ils deviennent plus nombreux que les dérivés en *-age*. On note également que les bases permettant la dérivation en *-age* et en *-ment* sont les plus nombreuses au milieu de l'intervalle, comme on s'y attendrait.

Le détail de la répartition des dérivés selon le nombre de syllabes est donné dans le tableau 3.3, incluant les verbes ayant un participe présent comptant moins de 2 syllabes ou plus de 5, ces deux groupes étant trop faibles numériquement pour être inclus dans la figure précédente.

TABLE 3.3 – Nombre de dérivés par longueur en syllabe des verbes au participe présent

	age	ment&age	ment	Total
1	7 58,33%	3 25,00%	2 16,67%	12 100,00%
2	520 59,09%	218 24,77%	142 16,14%	880 100,00%
3	959 40,09%	642 26,84%	791 33,07%	2392 100,00%
4	283 36,28%	127 16,28%	370 47,44%	780 100,00%
5	30 32,61%	6 6,52%	56 60,87%	92 100,00%
6	3 50,00%	1 16,67%	2 33,33%	6 100,00%
Total	1802	997	1363	4162

3.2.3 Répartition par groupe morphologique

Les verbes du français sont répartis traditionnellement en deux groupes de conjugaison, formant des ensembles cohérents sur le plan de la morphologie flexionnelle, auxquels s'adjoint un troisième groupe contenant l'ensemble des verbes dont la conjugaison ne rentre pas dans les deux premières catégories. Nous examinons ici la répartition des types de dérivations en fonction du groupe de conjugaison auquel appartient le verbe-base.

TABLE 3.4 – Nombre de dérivés par groupe verbal du radical

	G1	G2	G3	Total
age	1793 46%	30 11%	40 43%	1863 44%
ment	1155 30%	194 71%	41 44%	1390 33%
-ment & *-age*	937 24%	50 18%	13 14%	1000 24%
Total	3885	274	94	4253

Les verbes du deuxième groupe sont bien plus fortement associés à la dérivation en *-ment* qu'à la dérivation en *-age* puisque plus de 70% d'entre eux ont uniquement un dérivé en *-ment*. Pour le premier groupe, les proportions sont similaires à la répartition globale des dérivés dans le lexique, ce qui suggère qu'il n'influe pas dans le choix de dérivation dans un sens ou dans l'autre. Enfin, il ne peut pas être tiré directement de conclusions à partir de la répartition des dérivations dans le troisième groupe en raison de sa nature hétéroclite.

3.2.4 Admission d'un rôle d'agent

Le rôle d'agent est le rôle thématique d'un actant verbal qui a prototypiquement la propriété d'être volontairement à l'origine de l'événement décrit par le verbe, d'être l'initiateur d'un changement d'état chez un autre actant. La base lexicale Verb∋Net propose pour chaque verbe, une annotation des rôles thématiques qui peuvent être assignés par celui-ci. Le tableau qui suit montre la répartition des dérivations selon le trait de sélection d'un rôle thématique d'Agent tel qu'encodé par Verb∋Net.

TABLE 3.5 – Répartition des dérivations selon la sélection d'un Agent

	N	Y	Total
age	32	441	473
ment	143	445	588
ment&age	49	493	542
Total	224	1379	1603

Au vu des chiffres montrés dans ce tableau, il semble que ce n'est pas tant l'agentivité d'un verbe qui favorise la dérivation en *-age*, que son absence qui l'inhibe par rapport à la dérivation en *-ment* : le nombre de verbes sélectionnant un agent, et possédant les deux dérivés attestés, est proche de 500, tandis qu'il y en a 445 qui n'ont qu'un dérivé en *-ment* d'attesté, ce qui est plus élevé que le nombre de verbes ayant un dérivé en *-age* et non en *-ment*. A contrario, le nombre de verbes n'admettant pas d'agent qui ont une dérivation en *-age* est bien moins élevé que ceux qui ont un dérivé en *-ment*.

3.2.5 Rôle de stimulus et expérienceur

Les rôles sémantiques de stimulus et d'expérienceur sont souvent associés : le rôle d'expérienceur est celui d'une entité consciente du procès (donc prototypiquement un humain), qui n'a pas de contrôle dessus et chez qui se produit un changement d'état psychologique. Le rôle de stimulus est, a contrario, celui d'une cause du procès, sans impliquer nécessairement de conscience ou de volition. Ces deux rôles thématiques remplissent donc chacun en partie les critères d'agentivité de Dowty repris par Kelling et peuvent potentiellement avoir une influence dans le choix de la dérivation en *-age* ou en en *-ment*. Contrairement au rôle d'agent, qui est systématiquement associé au sujet syntaxique dans les phrases actives, les rôles de stimulus et d'expérienceur peuvent être associés à différentes fonctions syntaxiques en fonction du verbe, voire en fonction de différents emplois du verbe : Par exemple dans « Jean regarde la fille », le sujet syntaxique est l'expérienceur, « Jean », tandis que « Cette histoire ne regarde pas Jean », le sujet syntaxique est le stimulus « Cette histoire ».

Pour cette raison nous présentons séparément la répartition des verbes admettant ces rôles, suivant qu'ils sont sujets ou non.

Le tableau 3.6 présente la répartition des verbes de Verb∋Net qui admettent un expérienceur sujet et le tableau 3.7, celle des verbes qui admettent un stimulus qui n'est pas un sujet syntaxique. La répartition montrée dans ces deux tableau est assez similaire, comme on peut s'y attendre car la corrélation entre ces deux variables est très forte : leur corrélation de Pearson est de 0.80.

Table 3.6 – Admission d'un expérienceur sujet

	age	ment&age	ment	Total
Non	460	518	555	1533
Oui	13	24	33	70
Total	473	542	588	1603

Table 3.7 – Admission d'un stimulus non-sujet

	age	ment&age	ment	Total
Non	461	527	557	1545
Oui	12	15	31	58
Total	473	542	588	1603

Ces deux tableaux montrent une légère préférence pour la dérivation en *-ment* parmi les verbes ayant cette configuration syntaxique.

Le tableau 3.8 présente la répartition des verbes de Verb∋Net admettant un stimulus comme sujet syntaxique, et le tableau 3.9 celle des verbes admettant un rôle d'expérienceur dans une fonction autre que sujet. À nouveau, la répartition est similaire car la corrélation entre ces deux variables est forte, le test de Pearson donnant aussi une valeur de 0.94.

Table 3.8 – Admission d'un stimulus sujet

	age	ment&age	ment	Total
Non	449	471	466	1386
Oui	24	71	122	217
Total	473	542	588	1603

Table 3.9 – Admission d'un rôle d'expérienceur non-sujet

	age	ment&age	ment	Total
Non	450	476	465	1391
Oui	23	66	123	212
Total	473	542	588	1603

Ici nous voyons une préférence plus nette en faveur de la dérivation en *-ment* parmi ces verbes, puisqu'il y en a 5 fois plus que parmi les verbes ayant uniquement un dérivé en *-age*. Les verbes psychologiques sélectionnant expérienceur et stimulus favorisent donc plutôt la dérivation en *-ment* et plus nettement si le stimulus est le sujet syntaxique.

3.2.6 Sous-catégorisation

La sous-catégorisation d'un verbe est constituée par le nombre et la réalisation syntaxique des arguments nécessaires pour que le syntagme verbal soit bien formé. Ce trait varie suivant les formes du verbe et aussi selon les sens qui peuvent être associés à une même forme verbale. Nous avons considéré qu'un verbe-base admettait un argument syntaxique si au moins un sens d'une forme du paradigme de conjugaison tel que présenté dans le LEFFF admettait cet argument.

Les tableaux suivants montrent la répartition des dérivations selon l'admission d'un objet direct par les verbes-bases, puis selon l'admission d'un objet indirect en *de*, puis selon l'admission d'un objet indirect en *à*.

TABLE 3.10 – Admission d'un objet direct

	N	Y	Total
age	72	1282	1354
ment	179	984	1163
ment&age	53	892	945
Total	304	3158	3462

Le fait d'admettre un objet direct est nettement plus répandu parmi les verbes ayant un dérivé en *-age*, à contrario les verbes ayant un dérivé en *-ment* sont nettement majoritaires parmi les verbes n'admettant pas d'objet direct.

TABLE 3.11 – Admission d'un objet indirect en *de*

	N	Y	Total
age	1291	63	1354
ment	991	172	1163
ment&age	858	87	945
Total	3140	322	3462

À l'inverse, l'admission d'un objet indirect en *de* est plus répandue chez les verbes possédant un dérivé en *-ment* que parmi ceux n'ayant qu'un dérivé en *-age*.

TABLE 3.12 – Admission d'un objet indirect en *à*

	N	Y	Total
age	1270	84	1354
ment	1050	113	1163
ment&age	854	91	945
Total	3174	288	3462

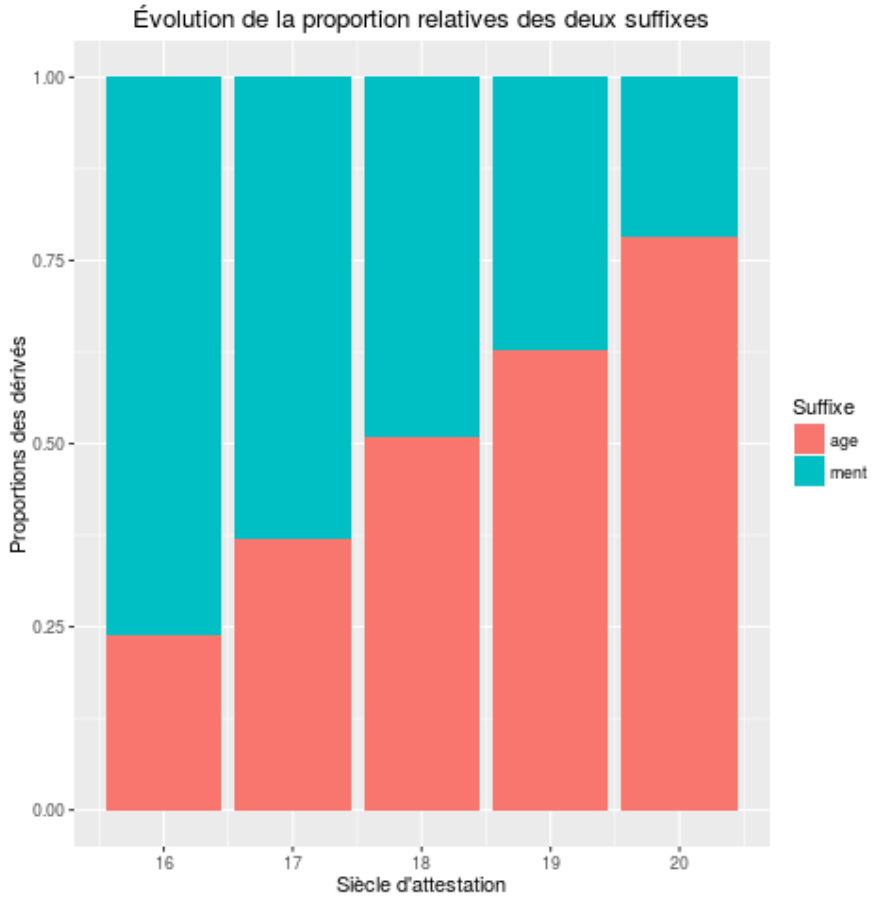
De même, la dérivation en *-ment* est plus fréquente chez les verbes admettant un objet indirect en *à*.

3.2.7 Répartition par date d'attestation

Le fait que la préférence pour un type de dérivation par rapport à une autre évolue au fil du temps a été noté dans plusieurs cas de concurrence, par exemple, en anglais, pour les dérivations en *-ment*, *-ity* et *-ation* par [Lindsay et Aronoff, 2013], et dans le cas spécifique de *-age* et *-ment* par [Kelling, 2001].

La figure suivante montre l'évolution de la proportion des deux types de dérivation parmi les nouvelles attestations pour chaque siècle.

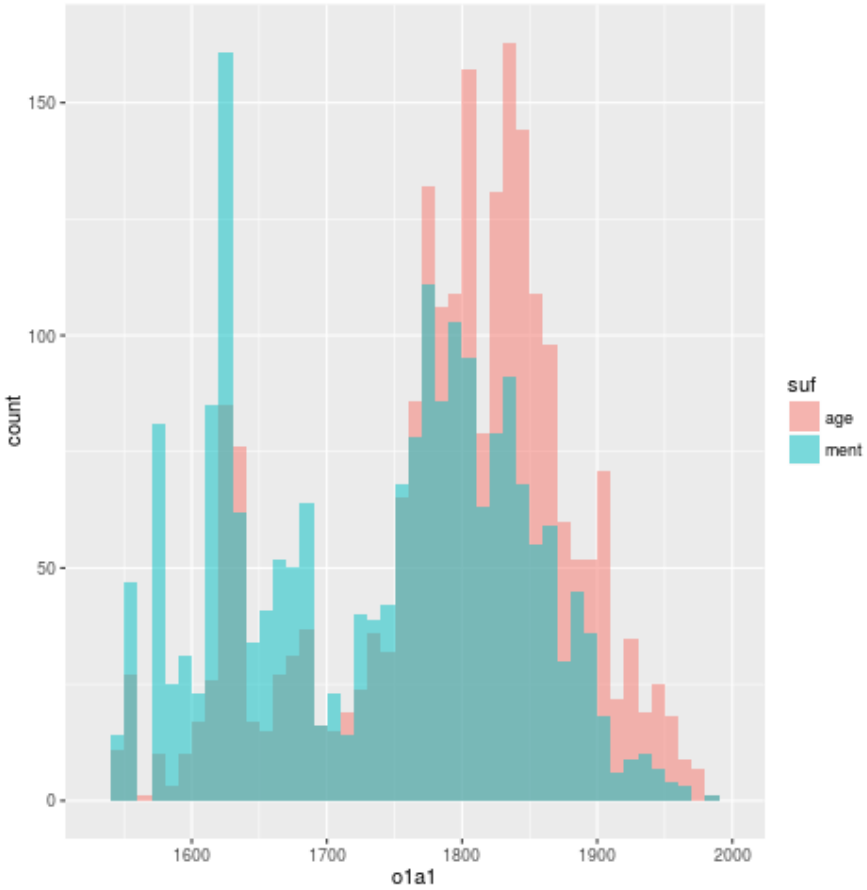
FIGURE 3.3 – Influence de la diachronie



On constate que parmi les nouveaux dérivés en *-age* et en *-ment*, la proportion de dérivés en *-age* s'accroît sans cesse au fil du temps.

La figure 3.4 montre l'évolution de la quantité brute de nouveaux dérivés pour les deux suffixes :

FIGURE 3.4 – Nouveaux dérivés au cours du temps



Les deux suffixes sont bien productifs depuis le moyen-âge, et la grande période de production de nouveaux dérivés en *-age* semblent commencer plutôt au XVIII^{e} siècle qu'au XIX^{e} comme le dit Kelling. La différence est due vraisemblablement à la source utilisée pour relever les dates, ici, Google Ngrams, et probablement un dictionnaire pour Kelling.

3.3 Modèles de régression logistique sur le lexique

Nous allons présenter une tentative de quantifier l'effet de différents traits lexicaux du verbe sur le choix entre les deux types de dérivations au moyen de la régression logistique, une méthode statistique modélisant une variable binaire sous la forme d'une estimation de probabilité. Le système employé pour effectuer la modélisation est le logiciel R [R Core Team, 2017].
Le principe de la régression logistique est de rechercher les paramètres d'un polynôme tel que l'erreur, c'est à dire la différence entre la composition de ce polynôme par la fonction logistique et les points de donnée réels, soit minimisée [Cox, 1958].

La probabilité qu'il est le plus tentant de vouloir modéliser de prime abord serait la probabilité, pour un verbe donné, de la possibilité d'une dérivation selon l'une ou l'autre stratégie.
Cependant des difficultés de plusieurs ordres empêchent d'aborder le problème de cette façon :

Il est impossible en pratique de prouver à coup sûr qu'un mot n'existe pas car aucun corpus (ni a fortiori aucune ressource) ne recouvre parfaitement toute la langue dans son usage, et en conséquence il serait dangereux de se baser en quelque façon sur la fréquence faible ou l'absence d'un mot dans un corpus pour en déduire que ce mot n'existe pas.
De plus, à supposer qu'on puisse prouver de façon convaincante qu'un mot n'existe pas dans la langue, il resterait à donner une interprétation théorique de cette inexistence, compte tenu du fait que l'ensemble des mots qui existent n'est qu'un sous-ensemble des mots possibles, et que l'inexistence d'un mot ne permet pas vraiment de conclure qu'il est impossible (ou moins possible) qu'il puisse exister. [Rainer, 2012] montre au contraire qu'un mot possible mais non-attesté peut par exemple avoir des dérivés attestés.

Pour cette raison, il est impossible de modéliser simplement la probabilité de l'utilisation d'un procédé de dérivation dans l'absolu.

D'autre part, étant donné qu'on ne s'attend pas à ce qu'il y ait des traits d'un verbe qui favorisent spécifiquement l'existence des deux dérivés à la fois par opposition à favoriser simplement l'un des deux, on ne tiendra pas compte dans ces régressions des verbes ayant à la fois un dérivé en *-age* et un en *-ment*.

Se limiter aux verbes n'ayant qu'un seul dérivé attesté permettra d'obtenir un modèle plus facile d'interprétation et dont les résultats seront plus tranchés quant aux préférences existantes dans la langue.

Les modèles sont donc ici restreints aux verbes des données ayant l'une ou l'autre dérivation mais pas les deux, la probabilité modélisée par ces modèles correspondant donc à la probabilité d'obtenir un dérivé en *-ment*, sachant que le verbe a un dérivé en *-age* ou en *-ment* mais pas les deux.
Les modèles peuvent ainsi porter sur un total de 3258 verbes. En pratique, cependant, le nombre de verbes considérés dépend des traits utilisés : seuls les verbes n'ayant aucune valeurs manquante parmi les traits utilisés sont pris en considération parmi les modèles de régression.

Nous allons d'abord présenter les résultats de modèles de régression univariés pour chaque trait pris séparément, afin d'examiner la contribution de chaque trait pour le plus grand nombre de verbes possibles, puis des modèles de régression multivariés afin de comparer entre elles les contributions des différents traits.

3.3.1 Trait de dernière consonne

Les régressions, portant sur un total de 3165 verbes, sur le trait de dernière consonne montrent que ce dernier n'est généralement pas significatif, l'exception principale étant la finale en ʒ qui favorise la dérivation en *-ment*, ($p < 0.01$), ce qui est probablement le résultat d'un effet de dissimilation, d'une préférence pour éviter de mettre deux ʒ à la suite.

3.3.2 Trait de nombre de syllabes

La longueur en nombre de syllabes d'un mot est corrélée positivement avec la probabilité d'une dérivation en *-ment*, de façon hautement significative ($p < 10^{-15}$), sur la base de régressions effectuées sur un total de 3165 verbes.

3.3.3 Trait d'agentivité

Les régressions logistiques effectuées sur 1061 verbes et portant sur le trait d'agentivité issu de l'annotation de Verb∋Net montrent qu'il favorise la dérivation en *-age* de façon très fortement significative,($p < 10^{-12}$), confirmant l'importance du rôle de l'agentivité dans la formations de noms déverbaux en *-age*.

3.3.4 Trait d'expérienceur et de stimulus

Les régressions logistiques univariées effectuées séparément sur les traits d'expérienceur non-sujet et de stimulus sujet parmi les 1061 verbes de Verb∋Net montrent que ces deux variables ont chacune le même effet fortement significatif ($p < 10^{-12}$) favorisant la dérivation en *-ment*.

Les régressions logistiques univariées effectuées sur les traits d'expérienceur sujet et de stimulus non-sujet pour les mêmes verbes montrent que ces deux variables ont, quant à elles, un effet faiblement significatif ($p < 0.05$) favorisant également la dérivation en *-ment*

3.3.5 Ratio de fréquences des formes conjuguées

Nous avons inclus parmi les traits la proportion de fréquences d'utilisations du verbe à la première et à la deuxième personne du pluriel parmi les premières, deuxièmes et troisièmes personnes.

Nous faisons l'hypothèse qu'une plus grande proportion de premières et de deuxième personnes parmi les formes du verbe permet de mesurer indirectement la préférence de ces verbes pour un sujet humain.
Ainsi que l'hypothèse que, pour la plupart des verbes admettant un sujet humain, ce sujet est un agent prototypique, ce qui permet de supposer que ce ratio peut être considéré comme un proxy pour mesurer l'agentivité d'un verbe, d'une façon différente de Verb∋Net, plus graduelle, sans les biais possibles liés au choix d'inclusion des verbes dans cette base.

Néanmoins ce ratio reste d'interprétation difficile car sans doute lié à plusieurs traits sémantiques différents pour chaque verbe, en particulier du fait qu'un sujet humain n'est pas nécessairement un agent, mais peut aussi être notamment un expérienceur.

Le calcul a été effectué en faisant la somme des fréquences relatives des formes dans frWAC pour chaque personne de chaque verbe.
Seules les fréquences de formes de conjugaisons au pluriel ont été retenues en raison du nombre important de syncrétismes existant pour les formes du singulier, notamment pour le premier groupe : « saute », peut être aussi bien une première qu'une troisième personne du singulier, à l'indicatif et au subjonctif, ou bien une deuxième personne de l'impératif.
De même « sautais » peut être une forme de première ou de deuxième personne de l'imparfait. Au deuxième groupe, il y a de même des syncrétismes systématiques, par exemple « finis » est à la fois une forme de première et de deuxième personne du singulier.
L'existence de ces syncrétismes bruite fortement la variable et ce dans des directions imprévisibles puisque ce ne sont pas les mêmes en fonction du groupe verbal considéré.

Aussi, seules les formes de pluriel ont été retenues, car elles ne sont pas sujettes à ces syncrétismes.

Les modèles de régression effectués sur les 2747 verbes pour lesquels le trait est défini révèlent son caractère hautement significatif, favorisant la dérivation en *-age* ($p < 10^{-11}$).

De plus, des régressions logistiques ont été effectuées sur un jeu de données constitué à partir de l'ensemble des verbes de Verb∋Net ayant un ratio défini dans le GLÀFF, et selon les mêmes principes que le jeu de données constitué pour les verbes ayant un dérivé en *-age* ou en *-ment*.
Ces régressions menées sur un ensemble de 2521 verbes ont permis de conclure que ce ratio était un prédicteur hautement significatif de l'agentivité,($p < 10^{-15}$).

3.3.6 Trait de classe morphologique

L'appartenance d'un verbe au deuxième groupe ressort comme un prédicteur fortement significatif ($p < 10^{-07}$) d'une préférence pour la dérivation en *-ment* pour les régressions effectuées uniquement sur la base de ce trait, pour lequel 3253 verbes sont définis. L'appartenance au troisième groupe favorise également la dérivation en *-ment*, mais est faiblement significative ($p < 0.05$), ce qui résulte probablement de la nature hétéroclite de ce groupe, qui rassemble des verbes de classes flexionnelles très variées et aux comportements très diversifiés.

3.3.7 Sous-catégorisation

La significativité des traits de sous-catégorisation comme prédicteur de la dérivation est testée par des régressions sur les 2517 verbes pour lesquels ces prédicteurs sont définis.

Le fait pour un verbe de sélectionner un objet direct est un trait fortement significatif ($p < 10^{-04}$) favorisant la dérivation en *-age*, ce qui pourrait tendre à corroborer l'hypothèse d'un rôle de la sous-catégorisation syntaxique elle-même, mais pourrait aussi s'expliquer par une corrélation forte entre sélection d'un objet direct et agentivité du verbe.

La sélection d'un objet indirect en *de* est un trait hautement significatif ($p < 10^{-04}$) qui favorise la dérivation en *-ment*. La sélection d'un objet en *à* est moyennement significative ($p < 0.01$) et favorise de même une dérivation en *-ment*.

3.3.8 Date de première attestation

L'année de première attestation est un prédicteur hautement significatif ($p < 10^{-15}$) dans tout les cas, quelle que soit la modélisation retenue parmi les six possibles.

TABLE 3.13 – Nombre de verbes sur laquelle la régression est menée pour chaque trait

Trait utilisé	o1a1	o10a1	o1a5	o1a10	o10a5	o10a10
Nombre de verbes	2585	2519	2394	2034	1782	1547

Pour chaque trait $oNaM$, N est le nombre d'occurrences minimales par année de la fenêtre temporelle et M sa durée en années.

Aussi bien le nombre d'occurrence minimales par années de la fenêtre temporelle utilisée, que la largeur en années de cette fenêtre n'ont aucun impact sur le degré de significativité de ce trait d'années d'attestation.

3.3.9 Modèle de régression multivarié

Sur les divers modèles de régression multivariés testés sur notre jeu de données, les traits hautement significatifs séparément tendent à le rester, tandis que les traits faiblement significatifs ne le sont plus du tout du fait de la diminution des points de données entrainée par le croisement de ces différentes variables.

C'est le cas notamment sur ce modèle prenant en compte 948 verbes, pour le trait de consonne final en /ʒ/, qui bien que significatif dans les régressions effectuées sur l'ensemble des 3165 verbes où le trait de dernière consonne est défini, ne ressort plus ici comme significatif.

TABLE 3.14 – Modèle multivarié

formule	ratiop + Agent + groupe + lastcons + nbsyllabes + o1a1 + Obj + Objà + Objde + Experiencer + Stimulus				
	Coefficients :				
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	9.882e+00	2.121e+00	4.660	3.16e-06	***
ratiop	-4.125e+00	8.023e-01	-5.142	2.72e-07	***
AgentTRUE	-9.362e-01	3.176e-01	-2.948	0.00320	**
groupeG2	1.811e+00	6.225e-01	2.909	0.00362	**
groupeG3	7.821e-01	4.248e-01	1.841	0.06561	.
lastconsɲ	1.823e-02	1.513e+00	0.012	0.99039	
lastconsɥ	6.742e-01	1.745e+00	0.386	0.69931	
lastconsʃ	-1.701e+00	1.461e+00	-1.164	0.24424	
lastconsʁ	-1.770e+00	1.410e+00	-1.255	0.20947	
lastconsb	-1.941e-01	1.570e+00	-0.124	0.90163	
lastconsd	-2.037e+00	1.442e+00	-1.413	0.15764	
lastconsf	-1.176e+00	1.815e+00	-0.648	0.51684	
lastconsg	-1.639e+01	5.474e+02	-0.030	0.97612	
lastconsj	-1.030e+00	1.401e+00	-0.735	0.46214	
lastconsk	-1.831e+00	1.470e+00	-1.245	0.21313	
lastconsl	-7.539e-01	1.406e+00	-0.536	0.59194	
lastconsm	-1.779e+00	1.453e+00	-1.224	0.22084	
lastconsn	-1.184e+00	1.400e+00	-0.845	0.39788	
lastconsp	-1.292e+00	1.468e+00	-0.880	0.37869	
lastconss	-5.520e-01	1.414e+00	-0.390	0.69630	
lastconst	-1.116e+00	1.392e+00	-0.802	0.42278	
lastconsv	-2.514e+00	1.608e+00	-1.564	0.11792	
lastconsw	-1.393e+00	1.882e+00	-0.740	0.45912	
lastconsz	-1.372e+00	1.458e+00	-0.941	0.34681	
lastconsʒ	5.807e-01	1.454e+00	0.400	0.68951	
nbsyllabes	1.056e+00	1.500e-01	7.045	1.86e-12	***
o1a1	-5.425e-03	8.789e-04	-6.173	6.70e-10	***
ObjTRUE	-1.528e+00	3.104e-01	-4.924	8.48e-07	***
ObjàTRUE	3.199e-01	2.477e-01	1.291	0.19658	
ObjdeTRUE	6.145e-01	2.649e-01	2.320	0.02035	*
ExperiencersujTRUE	6.004e-01	4.859e-01	1.236	0.21660	
StimulussujTRUE	1.212e+00	3.248e-01	3.732	0.00019	***

Parmi les traits significatifs, l'effet le plus fort est celui du ratio de formes conjuguées, de -4.24, suivi de l'effet de l'appartenance au deuxième groupe, qui est de 1.88. Les traits suivants par ordre décroissant de la taille de l'effet sont le trait d'agentivité de Verb∋Net qui est de -1.52, puis celui du trait de sélection d'un objet direct, qui est de 1.41. On peut aussi noter l'importance du fait d'avoir un stimulus sujet, avec une taille d'effet de 1.21 et une significativité persistante malgré la petite taille du modèle.
Le plus petit effet parmi les traits significatif est l'effet lié à l'année d'attestation : même en tenant compte du fait qu'il est associé à des valeurs numériques de l'ordre de 10^3, le multiplier par 100 (ce qui équivaut à donner la taille de l'effet liée au siècle d'attestation) ne permet pas de trouver une taille d'effet supérieure à celle de la présent d'un objet indirect en de.

On observe que les traits de sous-catégorisation, en particulier celui d'objet direct, le trait booléen d'Agentivité de Verb∋Net et le ratio de fréquences de personnes (ratiop) ressortent tous comme significatifs y compris quand ils sont tous inclus en même temps dans le même modèle.

Cela laisse penser qu'ils représentent chacun des facteurs d'influence sur le choix du suffixe.
Ainsi, on peut supposer que les propriétés syntaxiques du verbe, et les propriétés de sélection sémantique, jouent des rôles qui ne sont pas totalement réductibles les uns aux autres dans ce choix, mais au contraire plutôt indépendants.
La mesure du facteur d'inflation de la variance généralisé [Fox et Monette, 1992] pour les variables utilisées dans ce modèle montre que ces variables ne sont pas fortement corrélées entre elles, puisque pour chacune des variables, il est bien en dessous de la valeur conventionnelle de 4, au delà de laquelle la corrélation est considérée comme compromettant l'estimation de l'effet des variable et de leur significativité.

Table 3.15 – Facteur d'inflation de la variance généralisé

	$GVIF$	Df	$GVIF^{(1/(2\times Df))}$
ratiop	1.104535	1	1.050968
Agent	1.234188	1	1.110940
groupe	1.630864	2	1.130068
lastcons	2.233377	20	1.020291
nbsyllabes	1.264227	1	1.124378
o1a1	1.167718	1	1.080610
Obj	1.074921	1	1.036784
Objà	1.108245	1	1.052732
Objde	1.094322	1	1.046098
Experiencer	1.209306	1	1.099684
Stimulus	1.083862	1	1.041087

Chapitre 4

La concurrence entre *-age* et *-ment* en corpus

Nous avons jusqu'ici modélisé la concurrence entre les deux suffixes selon les termes classiques de l'étude de la concurrence en morphologie dérivationnelle, comme un phénomène affectant plutôt le lexique, et considéré comme un jeu à somme nulle, où une base sélectionnée par un des deux suffixes est perdue pour l'autre.
Cependant, le nombre important de bases possédant des dérivés attestés pour les deux suffixes est le trait peut-être le plus remarquable, le plus spécifique de la concurrence entre *-age* et *-ment*. Nous nous proposons d'étudier à présent les modalités de la concurrence entre les éléments de ces paires de doublets dérivationnels dans le contexte de leurs occurrences en corpus.

4.1 Constitution des données

Nous avons extrait d'un corpus des occurrences d'utilisations de noms déverbaux en *-age* ou en *-ment* répertoriés par Démonette afin d'étudier les prédicteurs, dans le contexte syntaxique de ces noms, du choix d'un dérivé en *-age* ou en *-ment*.
Le corpus utilisé est un corpus FRWIKI de 279 735 153 tokens, fourni par Franck Sajous [1], constitué à partir de la version francophone de l'encyclopédie Wikipédia et étiqueté en parties du discours et en dépendance syntaxiques avec le logiciel Talismane [Urieli, 2013].
Pour chaque occurrence d'un dérivé en *-age* ou en *-ment* relevé dans Démonette et associé à sa base, les traits suivants ont été relevés dans le corpus :
le genre et le nombre de l'occurrence, le lemme, la catégorie, le genre et le nombre de la tête associée à l'occurrence, ainsi que le temps et la personne pour les cas où la tête est un verbe, la nature de la dépendance syntaxique entre cette tête et l'occurrence, puis les traits associés aux mots dépendant de l'occurrence :
la nature de la détermination, le nombre de syntagmes prépositionnels en *de* et en *par* associés à l'occurrence et le nombre total de dépendants.

1. Université Toulouse 2 Jean Jaurès/CNRS - Laboratoire CLLE- équipe ERSS

4.2 Fréquence relative des occurrences de doublets

Sur un ensemble de paires de dérivés en *-age* et en *-ment* issus de la même base relevés dans Démonette, nous avons relevé, pour chacun, leur nombre d'occurrences dans frWAC. Puis nous avons calculé la proportion d'occurrences de chaque dérivés par base et lequel des deux avait la fréquence relative la plus élevée.
Enfin, la moyenne de ces ratios a été calculée pour chacun des deux suffixes.

L'hypothèse est que s'il existe une préférence globale pour l'un des deux types de dérivation, alors cela devrait se refléter dans les fréquences relatives des doublets de deux façons :
D'une part, le dérivé ayant le suffixe préféré devrait être plus souvent majoritairement utilisé par rapport à l'autre, et, d'autre part, dans les cas où c'est l'autre dérivé qui l'emporte, la différence de fréquence entre les deux devrait alors être plus faible.

Sur les 682 paires de dérivés en *-age* et en *-ment* dont au moins l'un des deux possède une occurrence dans frWAC, dans 349 cas, c'est le dérivé en *-ment* qui possède la fréquence relative la plus élevée, tandis que dans 333 cas, c'est le dérivé en *-age*.

Le ratio de fréquence moyen sur l'ensemble des données est de 48% pour les dérivés en *-age* et de 52% pour les dérivés en *-ment*.

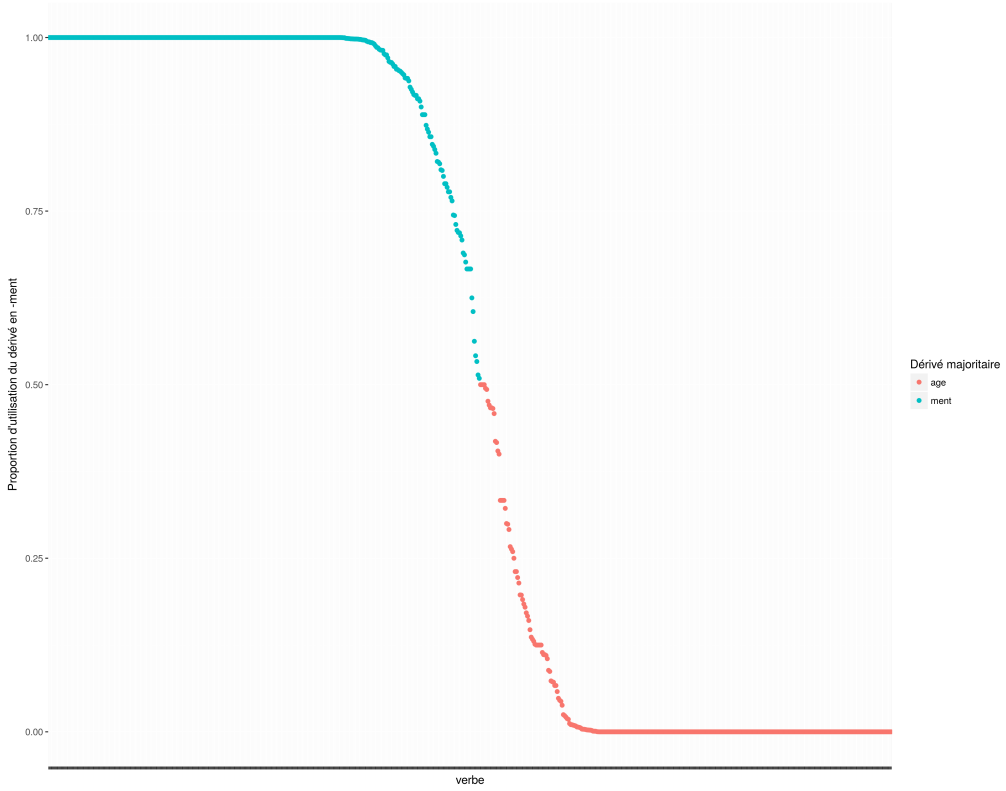
En se limitant aux doublets où le dérivé en *-age* est le plus fréquent, la proportion moyenne d'occurrence des dérivés en *-age* est de 95.1% En se limitant aux doublets où le dérivé en *-ment* est le plus fréquent, la proportion moyenne d'occurrence des dérivés en *-ment* est de : 96%

En se limitant aux 209 paires où les deux dérivés ont chacun au moins une occurrence dans frWAC : Dans 96 cas, le dérivé en *-age* est le plus fréquent, et le dérivé en *-ment* l'est dans 113 cas. Sur l'ensemble de ces 209 paires, la proportion moyenne des dérivés en *-ment* est de 55%. En se limitant aux cas où le dérivé en *-ment* est majoritaire, cette proportion moyenne est de 87%.

Des proportions aussi écrasantes sont indicatives de la force du blocage synonymique dans les deux cas. Le fait que les deux suffixes "gagnent" chacun la moitié du temps où deux dérivés sont en concurrence directe montre qu'il n'existe pas de tendance globale claire à préférer l'un plutôt que l'autre, ce qui ne préjuge pas de préférences plus localisées.

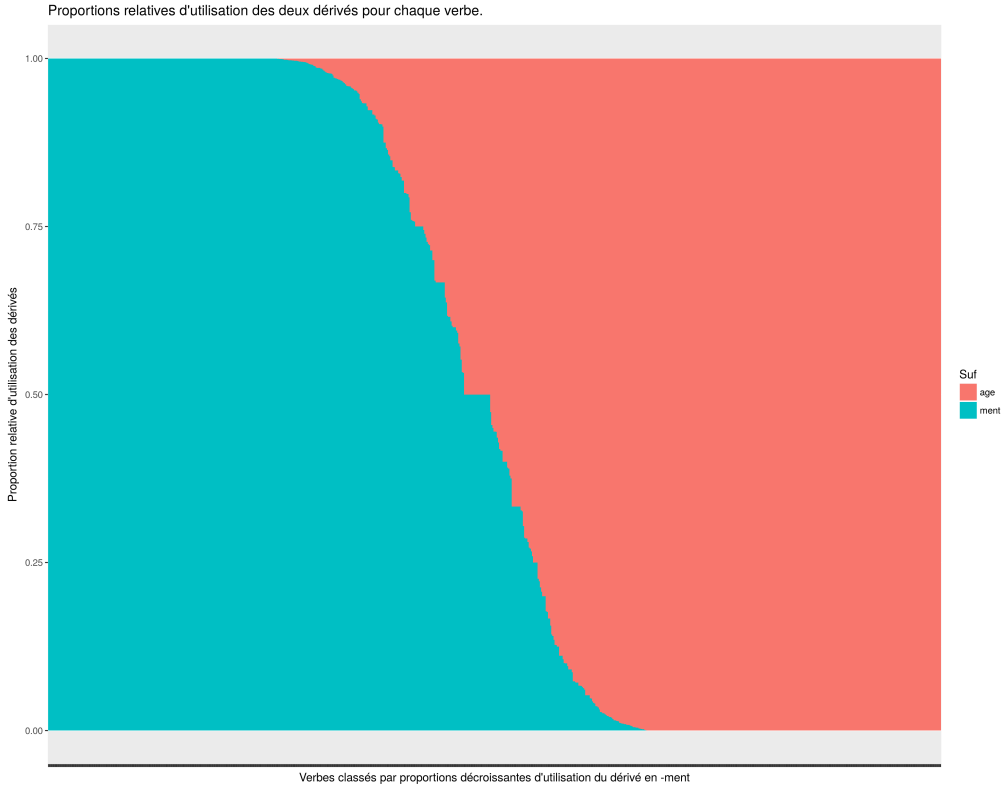
La figure 4.1 donne pour chaque verbe ayant des dérivés en *-age* et en *-ment* dans Démonette, la proportion d'utilisation du dérivé en *-ment* parmi les utilisations des deux dérivés dans frWAC. Les verbes sont ordonnés selon cette proportion en ordre décroissant.

FIGURE 4.1 – Proportion d'utilisations du dérivé en *-ment* dans frWAC



Cette puissance du blocage synonymique se vérifie dans le corpus FRWIKI : La figure suivante montre, pour 788 verbes ayant les deux types de dérivés attestés dans Démonette et dont au moins l'un des deux dérivés est utilisé dans le corpus FRWIKI, la proportion d'occurrences de ces dérivés parmi l'ensemble des occurrences des deux dérivés du verbes. Les verbes sont ordonnés selon la proportion d'utilisation du dérivé en *-ment* en ordre décroissant de gauche à droite.

FIGURE 4.2 – Proportion d'utilisations du dérivé en *-ment* dans FRWIKI



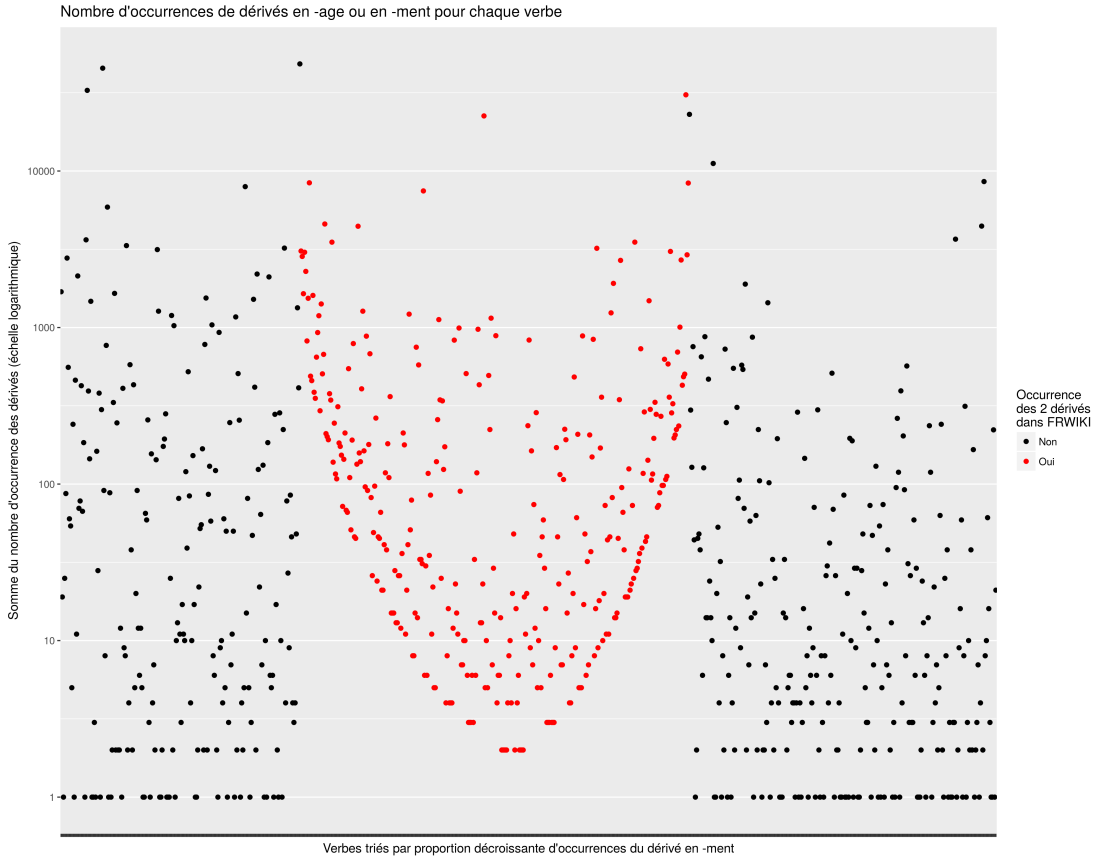
Une telle répartition est le signe que, même lorsque deux dérivés d'un même verbe sont attestés dans le lexique, le blocage synonymique empêche dans la majorité des cas que les deux soient effectivement utilisés.

Néanmoins, nous constatons aussi que la quasi-exclusivité suscitée par le blocage synonymique connaît un certain nombre d'exceptions, matérialisées sur ces figures par les verbes situés au milieu de la courbe, dont les deux dérivés sont non seulement attestés dans le corpus FRWIKI, mais dans des proportions relatives telles que l'on peut penser qu'il existe une concurrence réelle dans l'utilisation de ces deux dérivés.

La première hypothèse que nous pouvons formuler concernant les verbes situés dans cette zone intermédiaire est que leurs dérivés sont globalement d'utilisation moins fréquente : on peut en effet supposer que le blocage synonymique suscité par un mot est d'autant plus fort que le mot est fréquemment employé, et, a contrario, que si le dérivé le plus employé est peu fréquent, alors son concurrent a d'autant plus de chances d'être utilisé. Il est probable que pour les moins fréquents d'entre ces noms, leurs utilisations soient des recréations en ligne du dérivé.

La figure 4.3 montre, pour le même ensemble de verbes que la figure précédente, ordonné de la même façon, la somme du nombre d'occurrences de leurs deux dérivés dans FRWIKI, présentée sur une échelle logarithmique afin de faciliter les comparaisons :

FIGURE 4.3 – Fréquences absolues des dérivés dans FRWIKI par bases



Cette figure montre clairement que les bases peuvent se répartir en trois groupes : à gauche et à droite, les verbes dont seul l'un des deux dérivé est présent dans le corpus FRWIKI ne montrent pas d'organisation particulière. En revanche, nous pouvons voir que, pour les verbes dont les deux dérivés sont présents dans le corpus, la somme du nombre d'occurrences des deux dérivés décroît quand on approche du milieu de la distribution, puis croît quand on s'en éloigne, ce qui montre qu'il y a bien un lien entre la rareté générale de l'utilisation des dérivés d'un verbe et le fait qu'il existe une incertitude dans le choix de l'un ou de l'autre.
Nous voyons cependant que la corrélation est loin d'être parfaite : de nombreux verbes situés au milieu de la figure ne semblent pas avoir des dérivés utilisés moins fréquemments que ceux situés à gauche ou à droite.
Par conséquent, la rareté d'utilisation des dérivés d'un verbe ne permet pas d'expliquer entièrement la concurrence entre *-age* et *-ment*.

4.3 Données quantitatives sur les occurrences de doublets

Si les suffixes *-age* et *-ment* présentent une réelle différence sémantique, alors, cette différence doit se refléter par des environnements syntaxiques différents : Ainsi, [Grimshaw, 1990], citée par [Kerleroux, 2012], formule une liste de critères associés spécifiquement aux noms d'action ou de résultat, par opposition aux noms déverbaux ayant d'autres interprétations comme noms d'objet ou d'entité. Outre la réalisation obligatoire d'un complément prépositionnels, les autres critères sont l'utilisation exclusive d'un déterminant défini et du singulier.

Le choix en ligne entre le nom en *-age* et celui en *-ment* dérivé de la même base devrait donc être corrélé avec certains types de dépendances syntaxiques de ce nom.

Le premier trait que nous examinons est la présence d'un complément de nom introduit par la préposition *par*.
L'expression des arguments d'un verbe peut se faire de plusieurs façons dans les dépendants d'une nominalisation de ce verbe.
La présence d'un dépendant syntaxique d'une nominalisation introduit par la préposition *par* est généralement le signe que le dépendant en question correspond à l'agent de l'action décrite par la nominalisation.
Par exemple, dans « le ramassage des ordures par les éboueurs », où « les éboueurs » est l'agent.
Le rôle de l'agentivité dans le choix du type de dérivé paraît établi, et justifie l'hypothèse que la présence d'un tel complément est potentiellement corrélée à l'utilisation de l'un ou l'autre type de dérivé.

TABLE 4.1 – Répartition selon la présence d'au moins un complément de nom en *par*

	-age		-ment		Total
Absent	159256	36%	277689	64%	436945
Présent	1236	24%	3932	76%	5168
Total	160492	36%	281621	64%	442113

On constate que parmi les occurrences de dérivés ayant au moins un complément en *par*, les dérivés en *-ment* sont légèrement plus nombreux en proportion que leur proportion globale d'occurrences dans le corpus, ce qui semble indiquer qu'ils sont plus fréquemment accompagnés d'un agent exprimé que les dérivés en *-age*.

Cependant, le complément de nom en *par* n'est pas le seul moyen d'exprimer un agent : Une autre façon d'exprimer un agent parmi les dépendants syntaxiques d'une nominalisation verbale est l'utilisation d'un déterminant possessif. Par exemple dans « son raisonnement n'est pas brillant », « son » renvoie à l'agent associé à « raisonnement ».

Le tableau 4.2 donne la répartition par type de détermination classés selon leur possessivité :

TABLE 4.2 – Répartition selon le type de déterminant (1)

	-age		-ment		Total
Amalgame	50559	37,17%	85444	62,83%	136003
Autre	62309	36,67%	107603	63,33%	169912
Numéral	18547	39,83%	28013	60,17%	46560
Possessif de 1e et 2e personnes	66	44,59%	82	55,41%	148
Possessif de 3e personne	7219	47,45%	7994	52,55%	15213
Inconnu	21792	29,34%	52485	70,66%	74277
Total	160492	36,30%	281621	63,70%	442113

Comparée à leur répartition globale, la répartition des dérivés pour la détermination possessive semble plus équilibrée, puisque les dérivés en *-ment* ne représentent que 55% des dérivés ayant un déterminant possessif de première ou deuxième personne, et 52,5% de ceux en ayant un de troisième personne. Il semble donc y avoir une légère corrélation entre détermination possessive et dérivés en *-age*.

Cependant, il est difficile d'en tirer des conclusions, tant l'effectif de ces catégories est inférieur à celui des autres formes de détermination, ce qui s'explique par le fait que le corpus utilisé est un corpus encyclopédique, où l'utilisation de la première ou de la deuxième personne est sans doute inhabituellement basse.

Le complément de nom en *de* semble le plus souvent introduire un complément de nom ayant un rôle de patient dans le procès auquel réfère la nominalisation verbale. Cependant, en raison de la grande diversité parmi les rôles possibles de ce type de complément, l'interprétation de la répartition des dérivés parmi les occurrences ayant un complément de nom en *de* n'est pas aisée.

TABLE 4.3 – Répartition selon la présence d'au moins un complément de nom en *de*

	-age		-ment		Total
Absent	111546	39%	174073	61%	285619
Présent	48946	31%	107548	69%	156494
Total	160492	36%	281621	64%	442113

Ce tableau montre que, s'il semble y avoir une légère préférence pour les dérivés en *-ment* parmi les occurrences ayant au moins un complément de nom en *de*, la différence par rapport aux proportions globales des deux types de dérivés semble trop faible pour qu'on puisse en tirer des conclusions.

Nous examinons enfin la répartition des occurrences selon deux des critères énoncés par Grimshaw, le trait de définitude du déterminant, et le trait de nombre de l'occurrence :

Table 4.4 – Répartition selon le type de déterminant (2)

	-age		-ment		Total
Indéfini	21693	39,15%	33710	60,85%	55403
Défini	114427	37,45%	191120	62,55%	305547
Total	136120	37,71%	224830	62,29%	360950

Si le tableau 4.5 indique que les noms en *-ment* sont proportionnellement plus utilisés au singulier, le tableau 4.4 montre qu'ils sont plutôt moins fréquents parmi les occurrences utilisées avec des déterminants indéfinis, ce qui revient à dire que les critères énoncés par Grimshaw ne semblent pas concorder ici.

Table 4.5 – Répartition selon le nombre

	-age		-ment		Total
Pluriel	23731	28,82%	58624	71,18%	82355
Singulier	136761	38,01%	222997	61,99%	359758
Total	160492	36,30%	281621	63,70%	442113

4.4 Régression logistique à effet mixte

Les différences de répartition des types de dérivés selon les traits de leurs occurrences en corpus étant relativement faibles et d'interprétation difficile, nous allons tenter de modéliser l'occurrence d'un type de dérivation en fonction de ces traits tirés du contexte syntaxique, de manière à tester lesquels sont véritablement significativement corrélés au choix en ligne entre les deux types de dérivés.

La diversité des bases des dérivés pose ici un problème, car leur influence est évidemment très importante dans le choix du dérivé. Utiliser un modèle de régression classique reviendrait à modéliser avant tout l'importance des traits de la base, ce qui était déjà le but des modélisations menées sur le lexique.
Nous cherchons ici, au contraire, à modéliser des corrélations avec des traits liés à des préférences indépendantes de la base de dérivation utilisée.
Étant donnée la répartition des occurrences des dérivés pour chaque base, le poids d'une telle variable risquerait en outre d'écraser toutes les autres. Par ailleurs, la diversité des bases pose un problème pratique : il est simplement impossible de faire un modèle de régression logistique classique incluant une variable catégorielle comprenant plus de 700 niveaux.

Cependant, étant donnée l'influence prépondérante de la base, il n'est pas non plus possible de se contenter d'ignorer son influence sur le choix de l'un ou l'autre suffixe. Par conséquent, nous faisons le choix de contrôler cet effet en incluant la base de dérivation comme effet aléatoire dans nos modèles de régression qui seront donc des modèles mixtes, générés en utilisant le paquet Lme4 [Bates *et al.*, 2015] dans R.

Le premier modèle est généré à partir d'un ensemble de 360 950 observations pour un ensemble de 768 bases :

Table 4.6 – Modèle de régression mixte

formule	Suf (1|Verbe) + par + de + detdef + Number				
	Coefficients :				
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	0.73929	0.23136	3.20	0.0014	**
parTRUE	0.17651	0.14672	1.20	0.2290	
deTRUE	0.35997	0.02318	15.53	<2e-16	***
detdefY	-0.37803	0.02935	-12.88	<2e-16	***
Numbers	-0.93264	0.02560	-36.43	<2e-16	***

L'effet de la présence de compléments en *par* est non-significatif, ce qui est assez surprenant, étant donné le sens généralement agentif de ces compléments de nom. On constate que la présence de compléments en *de* représente un prédicteur hautement significatif d'un dérivé en *-ment*, tandis que l'utilisation d'un déterminant défini et d'un dérivé au singulier sont des prédicteurs très fortement significatifs de l'utilisation d'un dérivé en *-age*, ce qui tend à montrer que, selon les critères de [Grimshaw, 1990], les dérivés en *-age* sont plus fréquemment des noms désignant une action ou un résultat d'action de façon stable, ce qui corrobore les théories de [Dubois et Dubois-Charlier, 1999] sur le contraste d'interprétation entre les deux catégories de suffixes.

Le modèle suivant est généré à partir d'un ensemble de 41 570 observations pour un ensemble de 163 verbes-bases, un sous-ensemble du modèle précédent restreint aux bases telles que les proportions d'utilisations des deux dérivés de la base dans le corpus sont comprises entre 90% et 10%, donc aux bases pour lesquelles les deux dérivés sont véritablement en concurrence dans notre corpus.

Table 4.7 – Modèle de régression mixte №2

formule	Suf (1|Verbe) + par + de + detdef + Number				
	Coefficients pour les effets fixes :				
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.14702	0.11303	10.15	<2e-16	***
parTRUE	0.01998	0.18090	0.11	0.912	
deTRUE	0.31623	0.02485	12.72	<2e-16	***
detdefY	-0.39341	0.03205	-12.27	<2e-16	***
Numbers	-1.01376	0.02715	-37.34	<2e-16	***

L'absence de différence de résultat entre ce modèle et le précédent semblent montrer que les cas où l'un des deux dérivés ne domine pas exclusivement sont bien, dans l'ensemble, de simples cas intermédiaires entre les cas de dominations exclusive du dérivé en *-ment* et ceux de *-age*, sans propriété particulière qui pourrait expliquer une apparente singularité.

Chapitre 5

Conclusion

L'étude de la concurrence entre nominalisations verbales en *-age* et en *-ment* montre qu'un ensemble de facteurs complexes est à l'œuvre dans le choix parmi les deux suffixes, chacun jouant un rôle relativement indépendant des autres.
La théorie présentant le degré d'agentivité du verbe-base comme le critère central pour le choix parmi ces deux suffixes paraît corroborée en grande partie, mais avec de nombreuses nuances, telles que, par exemple, le rôle du critère purement morphologique que constitue l'appartenance de la base au deuxième groupe de conjugaison.
Le poids relatif de ces traits fortement diversifiés, ainsi que l'existence de nombreux doublets, montrent que l'opposition entre les deux types de dérivation est une opposition tout à fait graduelle.

Cependant, cette vision est elle-même nuancée par l'examen des fréquences relatives de ces doublets, qui montre que le blocage synonymique est un phénomène qui agit à de multiples niveaux : empêchant la création de doublets, mais également leur usage, ou alors dans un sens différencié, témoin l'une des paires de doublets les plus utilisées, la paire *équipage-équipement*, dont aucun des deux ne peut être considéré comme étant avant tout un nom d'action.

Le fait que les modèles de régression sur les occurrences de dérivés et leurs propriétés syntaxiques fassent ressortir des traits correspondant aux critères de Grimshaw et les associe à la dérivation en *-age* peut permettre de conjecturer que, comme le supposaient Dubois & Dubois-Charlier, les dérivés en *-age* sont malgré tout des noms d'action plus prototypiques que les dérivés en *-ment*.

Une différence sémantique sous-jacente entre *-age* et *-ment*, liée à la notion de contrôle suggérée par Fradin, et reflétée dans la différence, décrite par Dubois & Dubois-Charlier, du sens du complément de nom en *de* pour les deux types de dérivés, expliquerait la corrélation observée entre emploi d'un type de dérivé et emploi de certains types de complément.
Une modélisation plus sophistiquée de la notion d'agentivité pourrait sans doute permettre de mettre en évidence la nuance sémantique entre les deux suffixes comme reflétant la gradualité de l'agentivité elle-même, comme Kelling le supposait.

Bibliographie

[Arndt-Lappe, 2014] Arndt-Lappe, S. (2014). Analogy in suffix rivalry : the case of english-ity and-ness. *English Language and Linguistics*, 18(03):497–548.

[Aronoff, 1976] Aronoff, M. (1976). Word formation in generative grammar. *Linguistic Inquiry Monographs Cambridge, Mass.*, (1):1–134.

[Baayen, 1993] Baayen, H. (1993). *On frequency, transparency and productivity*, pages 181–208. Springer Netherlands, Dordrecht.

[Bates *et al.*, 2015] Bates, D., Mächler, M., Bolker, B. et Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

[Cox, 1958] Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2):215–242.

[Danlos *et al.*, 2016] Danlos, L., Pradet, Q., Barque, L., Nakamura, T. et Constant, M. (2016). Un Verbenet du français. *Traitement Automatique des Langues*, 57(1):25.

[Debaty-Lucas, 1986] Debaty-Lucas, T. (1986). *Théorie fonctionnelle de la suffixation : appliquée principalement au français et au wallon du Centre* . Éd. Les Belles Lettres.

[Dowty, 1991] Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, 67(3):547–619.

[Dubois, 1962] Dubois, J. (1962). *Étude sur la dérivation suffixale en français moderne et contemporain : essai d'interprétation des mouvements observés dans le domaine de la morphologie des mots construits.* Librairie Larousse.

[Dubois et Dubois-Charlier, 1999] Dubois, J. et Dubois-Charlier, F. (1999). *La dérivation suffixale en franc ais* . Nathan, Paris.

[Ferraresi *et al.*, 2010] Ferraresi, A., Bernardini, S., Picci, G. P. et Baroni, M. (2010). *Web corpora for bilingual lexicography. A pilot study of English/French collocation extraction and translation* , pages 337–359. Cambridge Scholars Publishing.

[Fox et Monette, 1992] Fox, J. et Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association*, 87(417):178–183.

[Fradin, 2014] Fradin, B. (2014). *La variante et le double*, pages 111–148. Presses Universitaires de Paris Ouest, Nanterre.

[Fradin, 2017] Fradin, B. (2017). Competition in derivation : What can we learn from duplicates ?

[Grimshaw, 1990] Grimshaw, J. (1990). *Argument structure*. the MIT Press.

[Hathout et Namer, 2014] Hathout, N. et Namer, F. (2014). La base lexicale démonette : entre sémantique constructionnelle et morphologie dérivationnelle. *In Actes de la 21e conférence sur le Traitement Automatique des Langues Naturelles*, pages 208–219, Marseille, France. Association pour le Traitement Automatique des Langues.

[Hathout *et al.*, 2014] Hathout, N., Sajous, F. et Calderone, B. (2014). GLÀFF, a Large Versatile French Lexicon. *In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

[Huyghe, 2014] Huyghe, R. (2014). La sémantique des noms d'action : quelques repères. *Cahiers de Lexicologie*.

[Kelling, 2001] Kelling, C. (2001). Agentivity and suffix selection. *In* Butt, M. et King, T. H., éditeurs : *The Proceedings of the LFG '01 Conference*. CSLI Publications.

[Kerleroux, 2012] Kerleroux, F. (2012). Il y a nominalisations et nominalisations. *LEXIQUE*, (20):157– 172.

[Lindsay et Aronoff, 2013] Lindsay, M. et Aronoff, M. (2013). Natural selection in self-organizing morphological systems. *Morphology in Toulouse*, pages 133–153.

[Lüdtke, 1978] Lüdtke, J. (1978). *Prädikative Nominalisierungen mit Suffixen im Französischen, Katalanischen und Spanischen*. Beihefte zur Zeitschrift für romanische Philologie. Niemeyer.

[Michel *et al.*, 2011] Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A. et Aiden, E. L. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014):176–182.

[Plag, 1999] Plag, I. (1999). *On the mechanisms of morphological rivalry : A new look at competing verb-deriving affixes in English*, pages 63–76. Trier : Wissenschaftlicher Verlag Trier.

[R Core Team, 2017] R Core Team (2017). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[Rainer, 2012] Rainer, F. (2012). Morphological metaphysics : virtual, potential, and actual words. *Word Structure*, 5(2):165–182.

[Sagot, 2010] Sagot, B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. *In 7th international conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta.

[Tribout, 2010] Tribout, D. (2010). *Les conversions de nom à verbes et de verbes à nom en français*. Thèse de doctorat, Paris 7. Thèse de doctorat dirigée par Fradin, Bernard.

[Urieli, 2013] Urieli, A. (2013). *Robust French syntax analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Thèse de doctorat, Université de Toulouse II le Mirail.