Différenciation des noms d'action dérivés : le facteur de technicité étudié en corpus

Marine Wauquier[1], **Cécile Fabre**[1] **et Nabil Hathout**[1]
CLLE, CNRS et Université Toulouse Jean Jaurès[1]

Résumé : Plusieurs travaux ont observé que les noms d'action en *-age* et *-ion* du français se distinguent par leur niveau de technicité. Les noms en *-age* relèvent davantage de domaines industriels ou techniques, contrairement aux noms en *-ion*, qui relèvent des sciences ou de la langue générale. Nous proposons dans ce chapitre de contribuer à l'étude comparative de ces nominalisations et de leur distinction à l'aide d'outils distributionnels et statistiques, en nous dotant au préalable d'une définition de la technicité.

En nous appuyant sur les méthodes de la sémantique distributionnelle, nous montrons qu'il existe bien une différence entre les noms d'action en *-age* et *-ion* que nous attribuons sur le plan référentiel à une différence de technicité. Dans un second temps, nous proposons une définition de la technicité dont nous dérivons un ensemble de corollaires linguistiques et de critères estimés sur corpus et à partir de ressources lexicales. Une étude statistique de la technicité des noms d'action basée sur ces critères confirme qu'il existe une différence entre les noms en *-age* et *-ion*. Ces critères permettent de construire un classifieur capable de prédire le suffixe avec une précision de 72%.

Mots clés : noms d'action, dérivation morphologique, technicité, sémantique distributionnelle, modélisation statistique

1. Introduction

Cette contribution aborde la question des analyses sémantiques dans les textes et les discours spécialisés en s'inscrivant dans le champ de la sémantique distributionnelle : elle associe les méthodes et outils de la linguistique de corpus et du traitement automatique des langues pour explorer le niveau de technicité des noms d'action en *-age* et *-ion* du français.

Il est communément admis que la formation d'un nom d'action à partir d'un verbe, par suffixation ou par conversion, modifie le lexème de base sur le plan catégoriel et le plan formel, mais pas sur le plan sémantique (Chomsky, 1970 ; Roché, 2009). Les verbes et leurs nominalisations sont à ce titre supposés être sémantiquement identiques ou plus probablement similaires, indépendamment du procédé morphologique impliqué. D'autres travaux nuancent fortement cette hypothèse générale en mettant au jour des différences sémantiques entre les procédés. C'est en particulier le cas de travaux sur les suffixes *-age*, *-ion* et *-ment*, qui sont les plus utilisés pour la nominalisation en français (Fradin 2014). Les nominalisations en *-age*, *-ion* et *-ment* se distinguent selon l'agentivité du sujet du verbe de base d'après Kelling (2001) et Martin (2010), et selon la concrétude du référent dénoté par son objet pour Fradin (2014). Il a par ailleurs été évoqué une préférence du suffixe *-age* pour les verbes nécessitant une interprétation physique, ou liés à des procédés industriels, alors que le suffixe *-ion* serait davantage présent dans la terminologie scientifique (Martin, 2010 ; Dubois, 1962). Les nominalisations en *-ment* dénoteraient quant à elles des attitudes et des états psychologiques (Dubois 1962) ou seraient ontologiquement non marquées (Martin 2010).

Fleischman (1990), cité également dans Uth (2010), suggère que la prédominance de noms en *-age* dans le domaine technique serait le fruit de l'histoire du suffixe : les nominalisations en *-age* se seraient multipliés au 19e siècle du fait de la révolution industrielle et du besoin grandissant de désigner des techniques et procédés nouveaux, et seraient notamment issues d'emprunts à la terminologie anglaise. Si Uth (2010) ne souscrit pas à l'hypothèse d'un emprunt massif, elle démontre pourtant la recrudescence de nominalisations en *-age* au 19e siècle. Le suffixe *-age* reste à ce jour un des suffixes les plus productifs du français, et l'on peut se demander si les nominalisations en *-age*, récentes ou anciennes, tendent globalement toutes à une plus forte technicité que les nominalisations en *-ion* et *-ment*.

Nous approfondissons au travers de cette étude la question de la spécialisation des noms d'action dérivés en termes de technicité. Les principales contributions de ce travail sont la proposition d'une définition de la technicité des noms d'action et la caractérisation de la différence entre les noms d'action en *-age*, *-ion* et *-ment* à partir de cette définition.

En nous appuyant sur les méthodes de la sémantique distributionnelle, nous confirmons dans un premier temps l'existence d'une différence entre ces noms d'action suffixés relative à leur technicité dans un corpus contemporain. Dans un second temps, nous proposons une définition de la technicité dont nous dérivons un ensemble de corollaires linguistiques et de critères estimés à partir de corpus et de ressources lexicales. Une étude statistique de la technicité des noms d'action basée sur ces critères corrobore l'hypothèse d'une plus grande technicité des noms d'action en *-age* et d'une moindre technicité des noms d'action en *-ion.*

2. Distinction distributionnelle des noms d'action

Avant de nous interroger sur la notion de technicité et ses manifestations linguistiques, vérifions d'abord qu'il existe bien une distinction sémantique entre les noms en *-age*, *-ion* et *-ment*. Nous utilisons un modèle de sémantique distributionnelle pour agréger l'information contextuelle et observer le profil des noms. Si l'hypothèse d'une plus forte technicité des noms d'action en *-age* est toujours d'actualité dans les usages contemporains, elle devrait se manifester dans la distribution des noms porteurs de ce suffixe.

2.1. Lexeur

Nous utilisons Lexeur (Fabre, Floricic & Hathout, 2004), une ressource lexicale validée manuellement regroupant 5 974 familles dérivationnelles partielles. Chacune d'elle est constituée d'un nom d'agent ou d'instrument en *-eur* (*abatteur, camionneur*), de son ou ses équivalents féminins[1] en *-euse* et *-rice* (*abatteuse, camionneuse*), de sa base verbale (*abattre*) ou nominale (*camion*), et des noms d'action dérivés de cette base (*abat, abattis*, *abattage*). Cette dernière catégorie ne se limite pas aux noms d'action mais comporte également des noms résultatifs (*sculpture*), statifs (*abattement*), entre autres. Nous ne considérons dans ce qui suit que les noms d'action appartenant à des familles dont la base est verbale (4 675 familles, 5 687 noms d'action). Le tableau 1 donne leur répartition pour les 4 298 suffixations en *-age*, *-ion* et *-ment*.

-age	***-ion***	***-ment***
1 697	1 376	1 225

Tableau 1 – Répartition des noms en fonction du procédé morphologique dans Lexeur

1 Les équivalents féminins ont été rajoutés lors de la construction de la base Démonette (Hathout & Namer, 2014).

2.2. Modèle distributionnel Word2Vec

Afin de vérifier en corpus l'existence d'une distinction sémantique entre les noms d'action en *-age*, *-ion* et *-ment*, nous exploitons l'outil de sémantique distributionnelle Word2-Vec (Mikolov, Chen, Corrado & Dean, 2013). La sémantique distributionnelle repose sur l'hypothèse d'une corrélation entre la similarité sémantique des mots et le partage de contextes en corpus (Harris, 1954 ; Firth, 1957). Dans cette approche quantitative du sens, la différence sémantique entre deux mots est évaluée à partir de la différence de leurs contextes, et mise en œuvre dans des modèles vectoriels : le sens des mots est représenté par des vecteurs calculés à partir de l'ensemble de leurs contextes dans un corpus. Les mots qui ont des contextes proches ou similaires sont représentés par des vecteurs proches dans l'espace vectoriel. Les vecteurs étant des objets mathématiques, on peut leur appliquer diverses opérations, telles que le calcul de distance, compris entre 0 (proximité nulle) et 1 (proximité maximale), mais aussi l'addition, la soustraction ou la multiplication, afin de représenter compositionnellement des structures complexes comme des syntagmes ou des phrases (Lenci, 2018).

Dans le cadre de cette étude, nous travaillons avec un modèle calculé à l'aide de Word2Vec sur un corpus constitué à partir de la version française de *Wikipedia* de 2013 et comptant 255 millions mots. Le corpus a été au préalable lemmatisé. Le choix de ce corpus se justifie par sa taille et la diversité des thématiques abordées, deux caractéristiques qui le rendent propice à une observation extensive du vocabulaire présent dans Lexeur.

Nous utilisons les paramètres par défaut de l'outil, qui comprend donc l'architecture CBOW, l'algorithme d'entraînement *Negative Sampling* et un seuil de sous-échantillonnage des mots fréquents de 10^{-3}. Nous fixons le nombre de dimensions des vecteurs à 100. Seuls les mots dont la fréquence est supérieure ou égale à 5 dans le corpus sont représentés. Ces paramètres excluent les mots trop rares (de fréquence inférieure à 5). De ce fait, seule une partie des noms en *-age*, *-ion* et *-ment* est effectivement représentée dans le modèle, comme le montre le tableau 2.

-age	***-ion***	***-ment***
673	1 142	566

Tableau 2 – Répartition des noms en fonction du procédé morphologique dans le modèle vectoriel

2.3. Méthode

Word2Vec permet de représenter dans un espace vectoriel les mots d'un corpus sur la base de leurs contextes, de calculer leur proximité par rapport à l'ensemble des autres mots du corpus et d'obtenir pour chaque mot ses plus proches voisins, c'est-à-dire les mots qui lui sont le plus similaires distributionnellement et, par extension, sémantiquement. Le vecteur du nom d'action *laminage* a par exemple pour plus proches voisins dans l'espace vectoriel résultant les noms *forgeage* (0.81), *extrusion* (0.80), *laminoir* (0.77) et *cintrage* (0.76), désignant, comme *laminage*, des procédés de déformation de la matière, ou l'outil utilisé pour réaliser cette action.

L'espace vectoriel permet aussi de travailler plus globalement sur les classes de noms d'action, en se détachant du niveau des formes particulières. Pour cela, nous construisons une représentation prototypique de l'ensemble des noms d'action en *-age*, *-ion* et *-ment*. Notre hypothèse est que la distinction sémantique entre les noms d'action en *-age*, *-ion* et *-ment* suggérée dans la section 1 est partagée par l'ensemble des noms d'action de chaque suffixation, et qu'elle est donc perceptible sur le plan distributionnel pour l'ensemble de la classe. Sur le plan vectoriel, nous instancions cette représentation prototypique par un vecteur moyen calcu-

lé à partir de l'ensemble des vecteurs d'une classe, en nous inspirant du travail de Kintsch (2001). L'information sémantique associée à ce vecteur moyen peut être examinée en observant ses 50 voisins les plus proches (nombre choisi arbitrairement). Ces voisins peuvent en effet être considérés comme des instances représentatives sur le plan distributionnel de la classe considérée, puisqu'ils correspondent aux mots les plus similaires au prototype des noms d'action agrégés.

Nous reportons en (1), (2) et (3) les 50 premiers voisins des vecteurs moyens des noms d'action en *-age* (VMage), en *-ion* (VMion) et en *-ment* (VMment). Nous signalons par un changement de typographie les voisins des vecteurs moyens qui ne sont pas porteurs du suffixe ciblé.

(1) *démoulage, séchage, usinage, remplissage, perçage, soufflage, démontage, sablage, broyage, coulage, étirage,* chargement*, soudage, dégraissage, vissage, trempage, traçage, sciage, polissage, gonflage, nettoyage, lavage,* pulvérisation*,* roulement*, roulage, remontage, réglage, meulage, assemblage,* recuit*,* soudure*, compostage,* dégagement*, salage, foulage, affûtage,* désinfection*, cuivrage, enrobage,* refroidissement*, clouage, décapage, grattage,* vidange*,* étanchéité*,* rechargement*, stockage, rinçage, brasage, dégivrage*

(2) *activation, réévaluation, dégradation, réduction, détérioration, simplification, transformation, modification, manipulation, assimilation, acceptation, dilution, détermination,* surcharge*, stimulation, dispersion, utilisation, application, dénaturation, dilatation, généralisation, évaluation, différenciation, mutation, altération,* survenue*, réaction, vérification, prolifération, complexification, action, limitation, régénération, homogénéisation,* coupure*, fixation, intervention, compréhension, stérilisation, inhibition, imputation, formulation, perception, aggravation, constatation, dissociation, multiplication, appropriation, révision,* mesure

(3) *enfoncement, déplacement, durcissement,* blocage*, échauffement, dépassement, élargissement, relâchement, abaissement, éparpillement, affaissement, isolement, envahissement, dégagement, écoulement, ajustement, allongement, basculement, dysfonctionnement, effritement, ralentissement,* rejet*, rétrécissement, étirement, endommagement, épuisement,* remplissage*, encombrement,* équilibrage*, emballement, tassement, affaiblissement,* accumulation*,* inconfort*,* absence*, ensablement, accroissement, traitement, décollement,* usure*, engorgement,* utilisation*, gonflement, étalement,* colmatage*, éloignement, relèvement,* redémarrage*,* lessivage*, arrachement*

2.4. Analyse des résultats

Sur le plan morphologique, on constate une forte homogénéité des voisins : 78% des voisins de VMage sont eux-même des noms d'action en *-age* ; 92% des voisins de VMion sont suffixés en *-ion* ; 76 % des voisins de VMment sont dérivés par suffixation en *-ment*. Ces dérivés se mélangent donc peu. À titre de comparaison, nous observons une homogénéité moindre lorsque la même procédure est appliquée aux noms d'agent déverbaux en *-eur*, *-euse* et *-rice*, avec respectivement 44 %, 10 % et 16 % de voisins porteurs du suffixe ciblé (Wauquier, Fabre & Hathout, 2018). Ce constat manifeste donc une distinction claire entre les noms d'action en fonction de leur suffixe. En d'autres termes, ces voisinages occupent des régions différentes de l'espace vectoriel, ce qui découle d'une distinction distributionnelle et donc sémantique entre les classes définies par les dérivés en *-age*, *-ion* et *-ment*.

Sur le plan sémantique et référentiel, on observe tout d'abord que les trois listes de voisins comportent exclusivement des noms d'action. Cependant, leur sémantisme précis diffère. Les voisins de VMion contiennent des noms dénotant des procédés ou phénomènes relatifs aux sciences, comme *dilution* ou *dénaturation*, ce qui va dans le sens de la description proposée par Dubois (1962). On y retrouve aussi des noms relatifs à des processus psycholo-

giques, comme *compréhension*, *détermination,* ou *perception*, et, en plus grande quantité, des noms relativement génériques voire sous-spécifiés comme *modification, action* ou *utilisation*, caractérisés par une importante polysémie. Le voisinage de VMage est très différent de VMion : on y trouve un grand nombre de noms relatifs à des techniques ou procédés industriels, tels que *soudage*, *usinage,* ou *brasage* et peu, voire aucun nom générique ou sous-spécifié : les plus génériques comme *stockage*, *nettoyage* ou *lavage* dénotent des actions qui sont intrinsèquement plus techniques que celles d'*utilisation* ou de *modification* (pour VMion). Le voisinage de VMment semble quant à lui plus hétéroclite : on y retrouve des noms relativement génériques comme *déplacement* ou *traitement*, mais aussi des noms plus spécifiques, pour certains relevant de techniques particulières, comme *relèvement* ou *ensablement*.

Ces observations dessinent deux profils sémantiques distincts pour les noms d'action suffixés en *-age* et *-ion*, les premiers exhibant un plus grand degré de technicité, les autres, par contraste, de généricité. Le profil sémantique ébauché pour la suffixation en *-ment* est quant à lui moins net. Nous allons dans la suite de cette étude étayer ces premières observations en nous dotant de critères et d'outils permettant de caractériser la technicité des noms d'action.

3. La notion de technicité

Les précédentes observations relatives à la technicité des noms d'action, rappelées en introduction, se heurtent à l'absence de définition de cette notion, peu abordée dans la littérature, et à l'absence de critères permettant de la caractériser. C'est à ces deux difficultés que tente de répondre cette section.

3.1. Définition générale

Simondon (1958) propose définition de la technicité qui comporte plusieurs volets. Le premier d'entre eux est l'agentivité. Il envisage en effet la technicité comme étroitement liée à l'homme : « [l'homme] est parmi les machines qui opèrent avec lui » (p. 12). C'est par ailleurs à l'aune des objets qu'il qualifie de techniques que le philosophe définit la technicité : « la technicité se manifestant par l'emploi d'objets techniques » (p. 156). Le degré de technicité d'un objet dépend de son niveau de perfectionnement et de complexité. La complexité est elle-même estimée selon le caractère inné ou acquis de la connaissance nécessaire à son utilisation (Simondon, 1958). Un savoir inné, non réfléchi, lié à un objet d'usage ou de la vie quotidienne, traduira une action moins technique qu'une action qui est le fruit d'une « opération réfléchie, d'une connaissance rationnelle élaborée par les sciences » (*ibid*, p. 85). Plus l'opération dénotée par le nom d'action nécessite une connaissance construite et acquise, plus elle présentera un haut degré de technicité. D'un point de vue ontologique, Simondon indique que « la technique touche au commerce, à l'agriculture, à l'industrie » (p. 97).

3.2. Perspective linguistique

La notion de technicité est mal définie dans la littérature en linguistique, comme le souligne Mudraya (2006). Divers travaux en terminologie se sont intéressés au degré de spécialisation, évalué en fonction de la spécialisation des émetteurs et destinataires des contenus textuels (voir Josselin-Leray, 2005, pour une typologie précise), mais, si l'on se rapporte aux propos de Simondon, la technicité est une notion plus précise que la simple spécialisation.

Différents travaux évoquent la notion de corpus technique, mais sans définir au préalable la notion de technicité. Par exemple, Siddiqui, Ren, Parameswaran et Han (2016) considèrent comme techniques des documents aussi divers que des brevets, des documents juridiques, des accords immobiliers, des archives historiques et des articles scientifiques (entre autres). D'autres travaux fondent leur analyse du domaine technique sur des corpus relevant du domaine nautique (Baroni et Bisi, 2004), du pétrole et des sciences et technologies (Mu-

draya, 2006), des technologies nucléaires et de la médecine (Habert, Naulleau & Nazarenko, 1996), des télécommunications (Drouin, 2003), ou de l'informatique (Wang, Lo, Jiang, Zhang & Mei, 2009). D'autres études encore font le choix de travailler sur l'encyclopédie en ligne *Wikipedia* (Nazar, Vivaldi & Wanner, 2012). La diversité des corpus dits techniques illustre le manque de contours nets de la notion de technicité.

Dans ces études, néanmoins, la langue technique est généralement définie par opposition à la langue générale : cette opposition se fait notamment au niveau du lexique (Lerat, 1997). C'est en effet la présence ou l'absence de termes techniques qui permet de statuer sur le caractère technique ou général de la langue (Koch, 2016 ; Forner & Thörle, 2016). Pourtant, Mudraya (2006) signale ainsi que « the division between technical and non technical vocabulary is far from distinct » (p. 238), lorsque l'idée même d'une technicité intrinsèque des unités lexicales n'est pas remise en question (Cabré, 2016). Ainsi, Wang, Lo, Jiang, Zhang et Mei (2009) évoquent des « technical terms ». Chez Drouin (2003) et Fuentes (2001), cette opposition se traduit notamment par la caractérisation des corpus utilisés, où l'on retrouve les expressions « non technical corpora » et « academic and technical corpora ». L'adjectif *technique* y est à chaque fois utilisé sans qu'une réelle définition ne soit proposée. Notons qu'une seconde opposition, tout aussi peu définie, émerge entre langue technique et langue scientifique, telles que mises en regard dans (Nazar, Vivaldi & Wanner, 2012) avec l'expression « scientific and technical corpora » ou « technical or specialized meanings » dans (Fuentes, 2001, p. 111). Lerat (2016) ébauche une distinction sur la base d'une opposition entre connaissance (relevant du domaine de la science) et savoir-faire (relevant du domaine technique).

3.3 Critères d'approximation de la technicité

En l'absence de définition de la technicité, nous proposons de la construire en nous appuyant sur les éléments apportés par les travaux de Simondon. Nous traduisons ensuite cette définition en différentes caractéristiques linguistiques, dont nous tirons des critères opératoires pour une expérience combinant l'utilisation de corpus et de ressources lexicales.

3.3.1. Définition de la technicité des noms d'action

Dans la lignée de Simondon (1958), nous considérons qu'un nom d'action est technique lorsqu'il dénote une action relevant d'un domaine technique. Nous entendons par domaine technique l'industrie et l'agriculture, tels que proposés par Simondon (1958) et Dubois (1962), domaines auxquels nous ajoutons l'artisanat : si Simondon l'excluait du fait d'une moindre technicité, il nous semble qu'il s'y rattache néanmoins en raison notamment des connaissances et des outils nécessaires pour la réalisation des actions liées au domaine. Plus que l'appartenance à ces domaines précis, nous retiendrons qu'un nom technique est spécifique à un domaine particulier, et qu'il ne sera donc pas utilisé dans d'autres domaines ou dans un contexte plus général (Koroucek 1982). Du fait de sa spécificité, le nom technique est peu transparent pour un public non initié, et nécessite des connaissances particulières pour comprendre l'action dénotée par le nom (Koroucek 1982). Ces différents éléments nous permettent de proposer une définition des noms d'action techniques :

> Nom peu transparent pour un public non initié, dénotant une action précise complexe, dont la réalisation et la connaissance nécessitent un savoir acquis et qui est spécifique à un domaine particulier. Les noms d'action techniques appartiennent aux domaines de l'industrie, de l'agriculture et de l'artisanat.

3.3.2. Caractéristiques linguistiques de la technicité

La définition ci-dessus permet de déduire 3 propriétés linguistiques (ci-après désignées *T1*, *T2* et *T3*) de la technicité des noms d'action. Nous proposons par la suite pour chacune d'elles une mise en œuvre basée sur des critères calculables à partir de corpus et de ressources.

La première propriété que nous dégageons est la spécialisation (*T1*) : du fait de la spécificité de l'action dénotée et de son lien à un domaine particulier et de spécialité, nous faisons l'hypothèse qu'un nom technique sera davantage utilisé dans un contexte spécialisé que dans un contexte général. Nous tirons par ailleurs de cette spécificité dénotationnelle une deuxième propriété, l'opacité du nom (*T2*) : ce dernier nécessite une explicitation de l'action dénotée à l'intention du public non spécialiste ne disposant pas des connaissances pour comprendre le nom. Cette explicitation passera par exemple par la définition du nom et de son action dans une ressource de type encyclopédique. La troisième et dernière propriété que nous dégageons est l'univocité du nom d'action technique (*T3*), par opposition à l'équivocité du nom d'action générique. Ce dernier est en effet défini réciproquement, à l'aune de notre définition, comme un nom transparent pour un public de non spécialistes, et dont l'action dénotée n'est ni spécifique, ni particulière à un domaine de spécialité. Nous approchons de fait l'univocité par le biais de l'équivocité, plus facilement quantifiable, qui pourra se traduire en termes d'appartenance à plusieurs domaines, à l'image du lexique scientifique transdisciplinaire (Tutin 2007), et en termes de polysémie voire de sous-spécification.

3.3.3. Ressources

Nous présentons dans le tableau 3 les ressources et corpus que nous utilisons pour opérationnaliser les propriétés linguistiques de la technicité que nous avons précédemment présentées. Nous utilisons pour la suite de ce travail une version plus récente du corpus *Wikipedia* pour garantir une plus large couverture des noms à annoter

Nom	**Aperçu quantitatif**	**Description**
Wikipedia2018	600 millions de mots	Corpus encyclopédique constitué de la version française de *Wikipedia* de 2018
LM10	200 millions de mots	Corpus journalistique constitué d'articles du journal *Le Monde* publié entre 1991 et 2000
DES	83 395 entrées	Dictionnaire électronique de synonymes (Manguin et coll, 2004)
TLFi	54 280 entrées	Version électronique du dictionnaire *Trésor de la Langue Française* (Dendien & Pierrel, 2003)
GLAWI	1 481 346 entrées	Dictionnaire électronique construit à partir du *Wiktionnaire* français (Hathout & Sajous, 2016)
LexiTrans	1 611 entrées	Lexique scientifique transdisciplinaire (Drouin, 2010)
LexNSS	305 entrées	Liste de noms sous-spécifiés (extraits de Legallois & Gréa, 2006)

Tableau 3 – Présentation des ressources et corpus utilisés

3.3.4. Critères

Le tableau 4 présente les critères associés aux propriétés décrites dans la section 3.3.2. Nous apportons ci-après quelques précisions quant à l'implémentation de certains critères.

Nom	Propriété	Description
RATIO_FREQ	*T1*	Ratio des fréquences relatives (par million de mots) dans Wikipedia2018 et dans LM10
PAGE_W18	*T2*	Présence ou absence d'une page dans Wikipedia2018
NB_CAT_W18	*T3*	Nombre de catégories associées à la page du nom dans Wikipedia2018. Égal à 0 lorsqu'il n'y a pas de page
NB_SYN	*T3*	Nombre de synonymes dans DES
NB_DEF_G	*T3*	Nombre de définitions dans GLAWI
NB_DEF_T	*T3*	Nombre de définition dans TLFi
NB_DOM_G	*T3*	Nombre de marqueurs lexicographiques de domaines dans GLAWI
NB_DOM_T	*T3*	Nombre de marqueurs lexicographiques de domaines dans TLFi
LST	*T3*	Présence ou absence dans LexiTrans
NSS	*T3*	Présence ou absence dans LexNSS

Tableau 4 – Présentation des critères de technicité

La présence en grand nombre de critères sanctionnant la propriété *T3* (8 sur 10) s'explique par la nature composite de l'équivocité, comme évoqué dans la section 3.3.2. Les différents critères permettent d'approcher différentes facettes de l'équivocité, comme la polysémie (NB_DEF), la sous-spécification (NSS) et l'appartenance à plusieurs domaines (NB_DOM, NB_CAT_W18, LST). Le critère RATIO_FREQ se base sur la mesure présentée dans (Hatier, 2016), consistant pour un nom donné à diviser sa fréquence relative dans un corpus d'analyse par sa fréquence relative dans un corpus de référence. Notre corpus de référence est le corpus LM10, considéré comme un exemplaire parmi d'autres de discours non technique. Notre corpus d'analyse est le corpus Wikipedia2018, dont nous faisons l'hypothèse que son contenu est plus spécialisé et technique, du fait de son caractère encyclopédique. Nous faisons ce choix malgré la diversité des domaines couverts par Wikipedia2018 car il n'existe pas encore à notre connaissance de corpus technique à la couverture suffisamment large. Nous testons également dans le corpus Wikipedia2018 la présence ou non d'une page décrivant l'action dénotée par le nom (PAGE_W18). Seules les pages dont le titre correspond exactement au nom d'action sont prises en compte. On considère par exemple que le nom *serrage* ne fait pas l'objet d'une page puisqu'on trouve des pages comme « Collier de serrage » ou « Noix de serrage », mais pas de page « Serrage ».

4. Modélisation statistique de la technicité

Les critères que nous venons de présenter permettent de réaliser une étude empirique testant l'hypothèse d'une plus grande technicité des noms d'action en *-age* et d'une moindre technicité des noms d'action en *-ion* et *-ment*.

4.1. Annotation des traits de technicité

Pour l'ensemble des 5 687 noms d'action de Lexeur, nous procédons à l'annotation automatique des critères de la section 3.3.4. Au regard de la définition de la technicité de la

section 3.3.1, nous nous attendons à ce qu'un nom d'action technique ait des valeurs plus élevées que les noms d'action non techniques pour les critères relevant de *T1* et *T2*, et des valeurs plus faibles pour ceux relevant de *T3*.

Si les noms en *-age* sont plus techniques que les noms d'action en *-ion* et *-ment*, ils devraient présenter une valeur de RATIO_FREQ plus élevée que les seconds. Nous nous attendons par ailleurs à ce que les noms en *-ion* et *-ment* fassent moins souvent l'objet d'une page dans Wikipedia2018 que les noms en *-age*. Enfin, les noms en *-ion* et *-ment* devraient avoir un nombre de synonymes (NB_SYN), de définitions (NB_DEF), et de marqueurs de domaines (NB_DOM) plus élevé que les noms en *-age*, et être davantage représentés dans le lexique scientifique transdisciplinaire (LST) et parmi les noms sous-spécifiés (NSS). Nous traduisons ces hypothèses par les signes (+) et (-) dans le tableau 5.

Le tableau 5 illustre le résultat de l'annotation pour 4 noms, choisis de manière à opposer 2 noms techniques (*arcure* et *drave*) et 2 noms plus génériques (*baisse* et *démarche*). Ces 4 exemples sont relativement bien décrits par les critères que nous avons implémentés : les noms techniques *arcure* et *drave*, correspondant respectivement à une technique liée à l'agriculture et à un ensemble d'activités relatives au métier du bois, présentent des valeurs proches de celles attendues. De la même façon, la généricité des noms semble globalement captée, comme illustré par les noms *démarche* et *baisse*.

	Technicité	***drave***	***arcure***	***démarche***	***baisse***
RATIO_FREQ	+	26.81	9.39	0.26	0.09
PAGE_W18	+	Oui	Oui	Non	Non
NB_CAT_W18	-	0	1	0	0
NB_SYN	-	0	3	41	34
NB_DEF_G	-	8	5	9	8
NB_DEF_T	-	0	4	4	11
NB_DOM_G	-	1	4	1	0
NB_DOM_T	-	0	3	0	0
LST	-	Non	Non	Oui	Oui
NSS	-	Non	Non	Oui	Non

Tableau 5 – Valeurs de chaque critère pour les noms drave*,* arcure*,* démarche *et* baisse

Notons cependant que certains critères pris individuellement ne valident pas nécessairement nos hypothèses. On remarque ainsi dans le tableau 5 que l'absence d'un nom du lexique NSS ne garantit pas sa technicité (*baisse*), ou que certains noms techniques ont autant de définitions que des noms peu techniques (respectivement 8 pour *drave* dans contre 8 et 9 pour *démarche* et *baisse* dans GLAWI). D'autres critères valident quant à eux nos hypothèses, à l'image du nombre de synonymes ou le ratio des fréquences.

Nous reportons dans le tableau 6 les valeurs moyennes des critères de technicité pour les 4 298 noms d'action en *-age*, *-ion* et *-ment*. L'appartenance au LST et aux NSS est traduite par le pourcentage de noms appartenant à ces lexiques. La présence d'une page correspondant au nom dans le corpus Wikipedia2018 est elle aussi rendue sous la forme d'un pourcentage.

	-age	***-ion***	***-ment***
RATIO_FREQ	1.4	1.9	0.8

PAGE_W18 (%)	25.7	60.1	19.2
NB_CAT_W18	0.39	0.82	0.24
NB_SYN	1.8	11.4	6.2
NB_DEF_G	1.3	2.2	1.4
NB_DEF_T	1.5	4.8	2.4
NB_DOM_G	0.4	0.8	0.3
NB_DOM_T	0.6	2.2	0.7
LST (%)	0.3	6.5	1.3
NSS (%)	0.05	1.4	0.5

Tableau 6 – Valeurs moyennes de chaque critère en fonction du procédé morphologique

Les résultats du tableau 6 corroborent l'hypothèse d'un plus grand degré de technicité des noms d'action en *-age*, et d'un plus grand degré de généricité pour les noms d'action en *-ion*. En effet, on constate que les noms d'action en *-age* ont en moyenne un nombre de synonymes, de définitions, de catégories et de domaines plus faible que les noms d'action en *-ion*. Ils sont par ailleurs proportionnellement moins nombreux à appartenir au LST ou au lexique des noms sous-spécifiés que les noms d'action en *-ion*. Notons que les différences de valeur obtenues à partir du TLFi sont plus marquées que celles obtenues avec GLAWI, en nombre de définitions comme en nombre de domaines. On retrouve ainsi parmi les noms présentant des degrés de technicité élevés au regard de nos critères *alésage*, *cardage*, *forgeage* et *filetage*, et parmi les noms présentant des degrés de technicité faibles *réduction*, *interrogation*, *exclusion* et *déculpabilisation.*

Dans le tableau 6, deux traits ne vont pas dans le sens de nos hypothèses, à savoir RATIO_FREQ et PAGE_W18. Le ratio des fréquences relatives en corpus est ainsi un peu plus faible pour les noms en *-age* que pour les noms en *-ion* (1.4 vs 1.9), ce qui signifie que les noms d'action en *-ion* sont plus fréquents que les noms d'action en *-age* dans le corpus technique que dans le corpus journalistique. Le pourcentage de noms d'action en *-age* faisant l'objet d'une page dans le corpus Wikipedia2018 est quant à lui nettement plus faible que celui pour les noms d'action en *-ion* (25.7 vs 60.1). Les critères issus des ressources sont donc ici plus fiables que ceux issus des corpus. Nous expliquons cela par le choix des corpus utilisés. Nous avons en effet fait l'hypothèse que le corpus Wikipedia2018 constituait un corpus technique, par contraste avec le corpus LM10. Or, Wikipeda2018 contient aussi bien des pages relatives à l'industrie que des pages relevant des sciences ou de la philosophie. Ce corpus revêt donc un aspect transdisciplinaire davantage caractéristique des noms en *-ion* que des noms en *-age*. Wikipedia2018 ne représente donc pas un corpus d'analyse technique satisfaisant, tout comme le corpus LM10 n'est pas nécessairement un corpus de référence satisfaisant. Un travail conséquent de préparation de corpus serait nécessaire pour permettre d'étudier plus finement la technicité en corpus.

Enfin, le tableau 6 montre que les noms d'action en *-ment* occupent une position intermédiaire : on note en effet que les valeurs moyennes du suffixe *-ment* pour les critères issus de ressources sont situées entre celles de *-age* et celles de *-ion*, à l'exception du nombre de domaines dans GLAWI, qui est légèrement inférieur à celui du suffixe *-age*. Les valeurs des critères obtenus à partir des corpus se révèlent elles aussi plus faibles que celles du suffixe *-age,* ce qui semble rapprocher le suffixe *-ment* du suffixe *-age* sur le plan de la technicité des actions dénotées.

4.2. Pouvoir prédictif des traits

Il est possible d'évaluer de façon systématique dans quelle mesure nos critères de technicité permettent de prédire le suffixe d'un nom d'action. Nous faisons le choix de nous concentrer sur les suffixes *-age* et *-ion* car ils exhibent les comportements les plus distincts.

Pour estimer le pouvoir discriminant de nos critères, nous réalisons une régression logistique à partir des critères précédemment décrits et qui permet de créer un classifieur visant à séparer les noms d'action en deux classes définies par les suffixes *-age* et *-ion,* dont nous postulons qu'elles incarnent respectivement la technicité et la généricité.

La précision de ce classifieur est de 0.719, ce qui signifie que près de 72% des noms sont bien catégorisés. Le tableau 7 reporte ainsi le nombre de mots bien et mal catégorisés par notre modèle.

		Observés	
		-age	***-ion***
Prédits	***-age***	1 421	591
	-ion	276	785

Tableau 7 - Matrice de confusion de la catégorisation

On retrouve parmi les noms bien catégorisés plusieurs cas de figures : des noms en *-age* techniques (*écrémage*, *écharnage*, *ramardage*, *essimplage*), des noms en *-age* non techniques (*cafetage, enjambage*, *éclairage*, *bafouage*), des noms en *-ion* non techniques (*imposition*, *versification*, *classification*, *intoxication*) et des noms en *-ion* techniques (*irrigation*, *galvanisation*). Ce dernier cas de figure est peu représenté.

Les catégories représentées par les dérivés en *-age* et *-ion* ne sont cependant pas prédites avec la même précision. Le tableau 6 montre que les noms en *-age* sont mieux catégorisés que les noms en *-ion* (83.7% de bonne catégorisation pour les premiers contre 57% pour les seconds). Pour comprendre ce décalage, nous avons observé plus en détail les noms en *-ion* qui ont été mal catégorisés.

Parmi les noms en *-ion* mal catégorisés, nous retrouvons des noms relatifs aux sciences (*dessication, photocoagulation, sulfitation, transfusion*) et des noms relevant de concepts et de phénomènes sociaux (*judaïsation*, *désassimilation*) par exemple. On retrouve enfin des noms techniques comme *désamiantation* ou *cimentation*. Sur le plan morphologique, nous constatons que 37.6% des noms en *-ion* mal catégorisés sont suffixés en *-isation* (*fiscalisation*, *porphyrisation*) et *-ification* (*démystification*, *acétification*). Mieux encore, plus de 68% des noms en *-isation* sont catégorisés comme des noms en *-age* (149 sur 219) et 100% des noms en *-ification* sont catégorisés en *-age* (73 sur 73). Cela est dû au fait que ces deux suffixes permettent de dériver des noms d'action à partir de verbes eux-mêmes suffixés en *-ifier* et *-iser*, porteurs d'un degré de spécialisation élevé.

5. Conclusion

L'objectif de cette étude était d'approfondir la question de la distinction sémantique des noms d'action en *-age* et en *-ion* en français sur la base de la technicité des actions qu'ils dénotent. Nous avons à ce titre proposé une définition des noms d'action techniques comme étant des noms peu transparents pour un public non initié, dénotant une action spécifique complexe, qui nécessite des connaissances acquises et qui relève des domaines de l'industrie, de l'agriculture et de l'artisanat. Sur la base de cette définition, nous avons montré que des traits linguistiques pouvaient être calculés à partir de corpus et de ressources pour caractériser

la technicité des noms d'action, et qu'ils permettaient effectivement de discriminer les noms d'action en *-age*, plus techniques, des noms d'action en *-ion*, moins techniques.

Nous avons constaté que les noms d'action en *-ion* semblaient plus hétérogènes que prévu en termes de technicité, avec la présence de noms en *-isation* et *-ification* dont on peut soupçonner que la technicité intrinsèque du verbe de base bruite la prédiction. Un découpage plus fin des noms d'action déverbaux sur la base de leur procédé morphologique permettrait certainement d'obtenir des classes plus homogènes pour permettre une analyse plus fine.

Nous envisageons dans la suite de ce travail d'étendre l'analyse de la technicité des noms d'action à d'autres nominalisations, comme les conversions. Nous souhaitons par ailleurs confronter ces critères à une annotation manuelle de la technicité des noms d'action. On peut en effet se demander dans quelle mesure les traits utilisés dans la présente étude sont corrélés à la perception de la technicité effective des noms d'action par les locuteurs.

Remerciements

Le modèle vectoriel a été généré sur la plateforme OSIRIM, administrée par l'IRIT et soutenue par le CNRS, la région Occitanie, le gouvernement français et le FEDER[2]. Nous remercions l'ATILF (UMR 7118) de nous avoir fourni le TLFi.

Références bibliographiques

BARONI Marco & BISI, Sabrina, 2004, « Using Cooccurrence Statistics and the Web to Discover Synonyms in a Technical Language ». *Proceedings of the 4th Language Resources and Evaluation Conference (LREC)*, Lisbonne, p. 1725-1728.

CABRÉ Teresa, 2016, « La terminologie », dans W. Forner et B. Thörle (eds), *Manuel des langues de spécialité*, Berlin/Boston, De Gruyter, p. 68-81.

CHOMSKY, Noam, 1970, « Remarks on Nominalization », dans R. Jakobs et P. Rosenbaum (eds.), *Readings in English Transformational Grammar*, Waltham Mass.; Toronto; London, Ginn and Company, p. 184-221.

DENDIEN Jacques & PIERREL Jean-Marie, 2003, « Le Trésor de la Langue Française Informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence », *Traitement Automatique des Langues*, vol. 44, n°2, p. 11-37.

DUBOIS Jean, 1962, *Étude sur la dérivation suffixale en français moderne et contemporain : essais d'interprétation des mouvements observés dans le domaine de la morphologie des mots construits*, Paris, Larousse.

DROUIN Patrick, 2003, « Term Extraction Using Non-Technical Corpora as a Point of Leverage », *Terminology*, vol. 9, n° 1, p. 99-115.

DROUIN Patrick, 2010, « Extracting a Bilingual Transdisciplinary Scientific Lexicon », *eLexicography in the 21st century: new challenges, new applications.* Louvain-la-Neuve, Presses Universitaires de Louvain/Cahiers du CENTAL, p. 43-53.

FABRE Cécile, FLORICIC Franck & HATHOUT Nabil, 2004, « Collecte outillée pour l'analyse des emplois discordants des déverbaux en *–eur* », communication présentée aux journées d'étude « La place des méthodes quantitatives dans le travail du linguiste », Toulouse, France.

FIRTH John R., 1957, « A Synopsis of Linguistic Theory 1930-1955 », dans J.R. Firth (eds), *Studies in linguistic analysis*, Oxford, Basil Blackwell, p. 1-32.

FLEISCHMAN Suzanne, 1990, *The French Suffix -age: its Genesis, Internal Growth, and Diffusion*, Ann Arbor, Univ. Microfilm Intern.

FORNER Werber & THÖRLE Britta, 2016, « Introduction », dans W. Forner et B. Thörle (eds), *Manuel des langues de spécialité*, Berlin/Boston, De Gruyter, p. 1-50.

FRADIN Bernard, 2014, « La variante et le double », dans F. Villoing, S. David et S. Leroy (dir), *Foisonnements morphologiques. Études en hommage à Françoise Kerleroux*, Paris, Presses universitaires de Paris Nanterre, p. 109-147.

FUENTES Alejandro Curado, 2001, « Lexical Behaviour in Academic and Technical Corpora: Implications for ESP Development », *Language Learning and Technology*, vol. 5, n° 3, p. 106-129.

HABERT Benoît, NAULLEAU Elie & NAZARENKO Adeline, 1996, « Symbolic Word Clustering for Medium-Size Corpora », *Proceedings of the 16th conference on Computational linguistics (COLING)*, Copenhagen, p. 490-495.

2 Voir http://osirim.irit.fr/site/fr (consulté le 04/12/2019)

HARRIS Zellig S., 1954, « Distributional Structure », *Word*, vol. 10, n° 2-3, p. 146-162.
HATHOUT Nabil & NAMER Fiammetta, 2014, « Démonette, a French Derivational Morpho-semantic Network », *Linguistic Issues in Language Technology*, vol. 11, n° 5, p. 125-168.
HATHOUT Nabil & SAJOUS Franck, 2016, « Wiktionnaire's Wikicode GLAWIfied: a Workable French Machine-Readable Dictionary », *Proceedings of the tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož , p. 1369-1376.
HATIER Sylvain, 2016, *Identification et analyse linguistique du lexique scientifique transdisciplinaire. Approche outillée sur un corpus d'articles de recherche en SHS*, thèse de doctorat, université Grenoble Alpes, Grenoble, disponible en ligne sur https://tel.archives-ouvertes.fr/ [consulté le 14/05/2018].
JOSSELIN-LERAY Amélie, 2005, *Place et rôle des terminologies dans les dictionnaires généraux unilingues et bilingues: étude d'un domaine de spécialité: volcanologie*, thèse de doctorat, université Lyon 2, Lyon.
KELLING Carmen, 2001, « Agentivity and Suffix Selection », In *Proceedings of the Lexical-Functional Grammar '01 Conference (LFG'01)*, Hong Kong, p. 147–162.
KINTSCH Walter, 2001, « Predication », *Cognitive science*, vol. 25, p. 173-202.
KOCK Christian, 2016, « Textes et discours en musicologie », dans W. Forner et B. Thörle (eds), *Manuel des langues de spécialité*, Berlin/Boston, De Gruyter, p. 169-184.
KOCOUREK Rostislav, 1982, *La langue française de la technique et de la science*, Amsterdam, John Benjamins Publishing Company.
LEGALLOIS Dominique & GRÉA Philippe, 2006, « L'objectif de cet article est de… Construction spécificationnelle et grammaire phraséologique », *Cahiers de praxématique*, vol. 46, p. 161-186.
LENCI Alessandro, 2018, « Distributional Models of Word Meaning », *Annual review of Linguistics*, vol. 4, p. 151–171.
LERAT Pierre, 1997, « Approches linguistiques des langues spécialisées », *ASp*, n° 15-18, p. 1-10, disponible en ligne sur http://journals.openedition.org/asp/2926 [consulté le 29/08/2019].
LERAT Pierre, 2016, *Langue et technique*, Paris, Hermann.
MANGUIN Jean-Luc, FRANÇOIS Jacques, EUFE Rembert, FESENMEIER Ludwig, OZOUF Corinne & SÉNÉCHAL, Morgane, 2004, « Le dictionnaire électronique des synonymes du CRISCO : un mode d'emploi à trois niveaux », *Cahiers du CRISCO*, n° 7.
MARTIN Fabienne, 2010, « The Semantics of Eventive Suffixes in French », dans A. Alexiadou, et M. Rathert (eds), *The Semantics of Nominalizations across Languages and Frameworks*, Berlin/New York, Mouton de Gruyter, p. 109-141
MIKOLOV Tomas, CHEN Kai, CORRADO Greg, & DEAN Jeffrey, 2013, « Efficient Estimation of Word Representations in Vector Space », *Proceedings of International Conference on Learning Representations (ICLR)*, Scottsdale.
MUDRAYA Olga, 2006, « Engineering English: A Lexical Frequency Instructional Model », *English for Specific Purposes*, vol. 25, n° 2, p. 235-256.
NAZAR Rogelio, VIVALDI Jorge & WANNER Leo, 2012, « Co-occurrence Graphs Applied to Taxonomy Extraction in Scientific and Technical Corpora », *Procesamiento del Lenguaje Natural*, vol. 49, p. 67-74.
ROCHÉ Michel, 2009, « Pour une morphologie lexicale », *Mémoires de la Société de Linguistique de Paris*, n° 17, p. 65-87.
SIDDIQUI Tarique, REN Xiang, PARAMESWARAN Aditya & HAN Jiawei, 2016, « Facetgist: Collective extraction of document facets in large technical corpora », *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, Indianapolis, p. 871-880.
SIMONDON Gilbert, 1958, *Du mode d'existence des objets techniques*, Paris, Éditions Aubier-Montaigne.
TUTIN Agnès, 2007, « Traitement sémantique par analyse distributionnelle des noms transdisciplinaires des écrits scientifiques », *Actes de Traitement Automatique des Langues Naturelles (TALN)*, Toulouse, p. 283-292.
UTH Melanie, 2010, « The rivalry of French -ment and -age from a diachronic perspective », dans A. Alexiadou, et M. Rathert (eds), *The Semantics of Nominalizations across Languages and Frameworks*, Berlin/New York, Mouton de Gruyter, p. 215-244.
WANG Xiaoyin, LO David, JIANG Jing, ZHANG Lu & MEI Hong, 2009, « Extracting Paraphrases of Technical Terms from Noisy Parallel Software Corpora », *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP) Conference Short Papers*, Singapore, p. 197-200.
WAUQUIER Marine, FABRE Cécile & HATHOUT Nabil, 2018, « Différenciation sémantique de dérivés morphologiques à l'aide de critères distributionnels », *SHS Web of Conference,* vol. 46.