Description et analyse des verbes désadjectivaux et dénominaux en *-ifier* et *-iser*

Benoît Sagot [1], Karën Fort [2]
INRIA Paris-Rocquencourt / Paris 7, INIST

Résumé

L'objectif de cet article est d'étudier le processus productif de dérivation morphologique qui construit des verbes désadjectivaux et dénominaux à l'aide des suffixes *-iser* et *-ifier*. Ce processus est étudié aux niveaux morphologique, syntaxique et sémantique, à la fois sous l'angle lexicographique (extension du lexique morphologique et syntaxique Le*fff*) et sous l'angle dynamique (détection et interprétation automatique de dérivés néologiques).

Abstract

This work aims at studying the productive process of morphological derivation which allows the creation of deadjectival and denominal verbs using *-iser* and *-ifier* suffixes in French. This process is analysed at the morphological, syntactic and semantic levels, both from the lexicographic point of view (extension of the morphological and syntactic lexicon Le*fff*) and from the dynamic point of view (automatic detection and interpretation of neologisms created by derivation).

Mots-clés : morphologie dérivationnelle, lexique syntaxique, sémantique lexicale, détection de néologismes.

1. Introduction

La création de ressources linguistiques de qualité pour le Traitement Automatique du Langage (TAL) est à la fois indispensable et très coûteuse. On peut cependant réduire ce coût en utilisant des mécanismes permettant de factoriser le travail manuel, à l'image de la morphologie dérivationnelle, qui identifie des liens réguliers entre entrées lexicales. Toutefois, ces liens ne sont pas toujours faciles à mettre au jour. Ainsi, les adverbes en *-ment* (Sagot & Fort 2007) ne peuvent pas tous être interprétés sémantiquement selon le patron *d'une manière Adj*. Nous nous sommes penchés, dans cette étude, sur les verbes dérivés en *-iser* et *-ifier*, qui nous paraissaient particulièrement intéressants de par leur productivité et leur utilisation dans les processus de néologisation.

Cet article est organisé comme suit. La section 2 fait le point sur les principaux travaux antérieurs portant sur les verbes en *-iser* et *-ifier*. La section 3 décrit les trois ressources à la base de ce travail et leur conversion dans un format unique. La section 4 détaille nos observations concernant les verbes en *-iser* et *-ifier* identifiés précédemment. Enfin, la section 5 en décrit la mise en œuvre sur corpus (pondération des entrées et extraction de nouvelles entrées).

[1] ALPAGE, INRIA Paris-Rocquencourt / Paris 7, benoit.sagot@inria.fr
[2] INIST, karen.fort@inist.fr

2. Les verbes dérivés en *-iser* et *-ifier* : travaux antérieurs

La dérivation de verbes à partir de bases adjectivales et nominales a été étudiée par de nombreux auteurs, notamment Willems (1979), qui a travaillé sur environ 3 000 verbes.

Willems a recensé 238 verbes désadjectivaux (dérivés d'adjectifs), dont 62 verbes en *-iser* et 18 en *-ifier*. Selon elle, ces verbes fonctionnent syntaxiquement comme suit :

```
SN1 + Vnd (rendre)+ SN2 + ADJx -> SN1 + Vd (ADJx) + SN2
```

Autrement dit, les constructions résultantes sont très courtes, intransitives ou transitives directes à un seul complément. Willems en déduit que pour les verbes dérivés, la syntaxe intra-proposition est remplacée par une syntaxe intra-mot ou morphosyntaxe.

Dal et Namer (2000) ont par ailleurs montré que les adjectifs en *-able* « sont des bases très improbables pour des verbes désadjectivaux de changement d'état, comme l'illustrent les exemples agrammaticaux suivants : **lavabiliser*, **imperméabl(ir|er)*, **aportabl(ir|er)*, **croyabilifier*. »

Willems ne distingue pas de sémantique différente selon qu'il s'agit d'un verbe en *-iser* ou en *-ifier*. Namer est quant à elle plus précise à ce sujet: « L'ensemble des opérations de construction de verbes désadjectivaux (affixation et conversion) produisent par essence des prédicats de changement d'état, l'état final étant décrit par l'adjectif en position de base, qui de ce fait exprime une propriété que l'on peut qualifier d'extrinsèque. Quand *a-* fabrique un verbe à partir d'une base adjectivale (*aplatir*, *allonger*), celui-ci est causatif transitif et le référent de son objet direct se retrouve dans l'état décrit par l'adjectif à la fin du déroulement du procès. A quelques exceptions près (*brutaliser*, *bêtifier*), il en est de même pour *-is(er)* et *-ifi(er)* » (Namer 2002).

Ainsi, l'outil DériF (Namer 2003) code automatiquement les verbes en *-is(er)* ou *-ifi(er)* sur base adjectivale comme des causatifs transitifs et leurs arguments reçoivent les cas *cause* (pour le sujet) et *thème* (pour l'objet direct).

En ce qui concerne les verbes dérivés de noms, ou dénominaux, le classement de Willems est moins explicite quant aux verbes en *-iser* et *-ifier*. Elle distingue en effet 6 classes de verbes dérivés dénominaux intransitifs dont aucune ne contient explicitement de verbes en *-iser/-ifier*, si ce n'est *bêtifier* et *ironiser*, que l'on trouve dans la classe des verbes se paraphrasant en *dire*. Quant aux transitifs, elle les classe également en 6 classes, dont une seule contient explicitement des verbes en *-iser* (18) et en *-ifier* (3). Cette classe, dont les verbes se paraphrasent en *pourvoir/rendre*, recueille entre autres des verbes comme *alcooliser*, *évangéliser* ou encore *valoriser*. On trouve cependant des dénominaux en *-iser/-ifier* dans trois autres de ces classes, sans qu'ils soient classés en tant que tels. Ainsi, on trouve dans la classe *soumettre/exposer* des verbes comme *expertiser* ou *martyriser*, dans *faire/transformer* les verbes *capitaliser* et *caraméliser* et enfin, dans la classe *avec*, des cas particuliers tels que *galvaniser*.

3. Construction d'un lexique morphologique, syntaxique et sémantique des verbes dérivés en *-iser* et *-ifier*

3.1. Ressources de départ

Nous nous sommes appuyés sur des données lexicales issues de trois ressources électroniques librement disponibles:

LVF (Lexique des Verbes Français) : ce dictionnaire des verbes a été développé par Dubois et Dubois-Charlier (Dubois & Dubois-Charlier 1997) sous la forme d'un thésaurus de classes syntactico-sémantiques, c'est-à-dire de classes sémantiques définies par la syntaxe. C'est une ressource verbale à large couverture, qui regroupe des informations syntaxiques mais également sémantiques. Les différentes classes, qui regroupent 25 610 entrées, sont définies par un agrégat de propriétés syntaxiques et lexicales.

DICOVALENCE : le dictionnaire de valence DICOVALENCE (van den Eynde & Mertens 2006) est une ressource informatique qui répertorie les cadres de valence de plus de 3 700 verbes simples du français, soit plus de 8 000 entrées. Le dictionnaire explicite en outre certaines restrictions sélectionnelles, certaines formes de réalisation (pronominales, phrastiques) des termes, la possibilité d'employer le cadre valenciel dans différents types de passif, etc. La particularité essentielle du dictionnaire réside dans le fait que les informations valencielles sont représentées selon les principes de « l'Approche Pronominale » en syntaxe (Blanche-Benveniste *et al.* 1984). Pour chaque place de valence (appelée *paradigme*) le dictionnaire précise le paradigme de pronoms qui y est associé et qui couvre *en intention* les lexicalisations possibles. Il précise aussi les *reformulations* possibles, comme le passif.

Le*fff* (Lexique des formes fléchies du français) : ce lexique est une ressource morphologique et syntaxique à large couverture du français, qui couvre l'ensemble des catégories (Sagot *et al.* 2006). Le Le*fff*, développé dans le formalisme lexical Alexina, est orienté vers une utilisation dans des outils de traitement automatique, mais cherche à préserver une pertinence linguistique. Il est ainsi utilisé par exemple dans des analyseurs à grande échelle pour différents formalismes (LFG, LTAG). Des travaux récents en ont amélioré la qualité et la couverture pour certaines classes d'entrées (constructions impersonnelles, constructions pronominales, adverbes en *-ment*), notamment par comparaison et fusion avec d'autres ressources lexicales comme DICOVALENCE et les Tables du lexique-grammaire (Danlos & Sagot 2007; Sagot & Fort 2007; Danlos & Sagot 2008).

3.2. Conversion au format du Le*fff* et fusion des ressources obtenues

Pour construire une base de données lexicales aussi couvrante et complète que possible, nous avons mis en œuvre une version simplifiée de la méthodologie décrite dans (Sagot & Danlos 2008): nous avons extrait des ressources de départ les entrées en *-iser* et *-ifier*[3] et nous

[3] Nous avons éliminé manuellement les entrées verbales correspondant à des lemmes se terminant par *-iser* ou *-ifier* mais ne relevant pas du processus de dérivation étudié ici (exemples: *croiser*, *griser*...).

les avons converties en un format unique, le format Alexina (celui du Le*fff*). La conversion du DICOVALENCE avait déjà été réalisée dans le cadre des travaux décrits dans (Danlos & Sagot 2008). En revanche, la conversion des « constructions » de LVF en cadres de sous-catégorisation et en propriétés syntaxiques complémentaires au sens du Le*fff* a été réalisée assez directement, en interprétant les codes représentant ces constructions, qui sont décrits dans (Dubois & Dubois-Charlier 1997). Seuls les codes représentant les constructions pronominales ont parfois posé problème: en effet, aucune information ne permet de distinguer les différents types de constructions non essentiellement pronominales (constructions réfléchies, réciproques, moyennes, neutres) ; de plus, dans les entrées concernées, les compléments obliques renseignés dans les codes pour la constructions transitive sont souvent ignorés dans le code pour la construction pronominale. Nous avons donc appliqué un certain nombre d'heuristiques, dont le résultat nécessiterait une validation manuelle.

La fusion des trois ressources ainsi obtenues au format Alexina a été réalisée de la même façon que décrite dans (Danlos & Sagot 2008), en préservant toutes les informations issues de chacune des ressources (exemples, identifiants de classes LVF, définitions formalisées LVF, etc.). Le résultat est un lexique morphologique, syntaxique et sémantique des verbes en *-iser* et *-ifier* composé de 2 246 entrées couvrant 1 701 lemmes verbaux distincts. Parmi ces entrées, on compte 1 862 entrées pour des verbes en *-iser* couvrant 1 457 lemmes distincts et 384 entrées pour des verbes en *-ifier* couvrant 244 lemmes distincts.

4. Description et analyse des propriétés des verbes dérivés en *-iser* et *-ifier*

Pour étudier plus précisément les mécanismes de dérivation morphologique à l'œuvre dans les verbes désadjectivaux et dénominaux en *-iser*/*-ifier*, nous avons développé un outil qui extrait automatiquement de notre lexique la base adjectivale ou nominale la plus plausible. Cet outil repose sur une liste de bases candidates provenant de deux sources d'informations: tout d'abord la définition sémantique semi-formalisée de LVF, pour les entrées qui correspondent à une entrée de LVF, et ensuite une liste de transformations possibles que nous avons répertoriées sous forme de patrons et ordonnées manuellement (par exemple, *-ais* > *-iser*, ou encore *-le* > *-iliser*). Les candidats sont alors recherchés dans le Le*fff*, et le premier trouvé est considéré comme la base la plus plausible.

Une fois les bases extraites, nous avons identifié, pour chaque entrée ayant reçu une base, le mécanisme dérivationnel à l'œuvre, sous la forme de patrons de transformation (cf ci-dessus). Les bases trouvées dans LVF ont parfois induit des patrons que nous n'avions pas répertoriés manuellement. Pour chacun de ces patrons, nous avons extrait l'ensemble des « radicaux » concernés (tout simplement ce qui reste de la base une fois le « suffixe » enlevé, comme *-le* dans l'exemple ci-dessus). Nous avons alors construit automatiquement une expression régulière aussi informative que possible et couvrant l'ensemble des radicaux dudit patron. Ainsi, nous avons extrait automatiquement que la transformation *-le* > *-iliser* n'était associée qu'à des

radicaux se terminant en `[aiu]b`, c'est-à-dire que la transformation ne s'applique qu'à des mots en *-able*, *-ible* et *-uble* (cf. par exemple *imperméable* > *imperméabiliser*, *crédible* > *crédibiliser* et *soluble* > *solubiliser*).

Sur les 2 246 entrées du lexique créé, 2 002 contiennent un identifiant de classe LVF nous en donnant la sémantique. Une analyse rapide de ces données confirme les assertions de (Willems 1979) et (Namer 2002) et montre qu'en majorité (1 103) les verbes en *-iser* et *-ifier* expriment un changement d'état de type *rendre*, *devenir* ou *muter* (*égaliser*, *vinifier*). Cela étant, on y trouve également un nombre significatif (119) de factitifs (*colostomiser*), de verbes ayant le sens de *munir* (232, dont *électrifier*), voire de *frapper* (*tyranniser*), *briser* (88).

5. Mise en œuvre sur corpus: pondération des entrées et extraction de nouvelles entrées

Afin d'enrichir notre lexique des verbes dérivés en *-iser* et *-ifier*, nous l'avons confronté à un corpus de taille importante. Nous avons utilisé pour cela un corpus journalistique (articles du *Monde diplomatique*) d'environ 20 millions de mots. Nous en avons extrait le nombre d'occurrences des formes fléchies de ces verbes dérivés, et donc le nombre d'occurrence des lemmes correspondant, ainsi que des formes susceptibles de représenter des lemmes verbaux dérivés en *-iser*/*-ifier* mais absents du lexique.

Les informations fréquencielles montrent une grande disparité dans la répartition des verbes en *-iser*/*-ifier*, avec une décroissance zipféenne. Les lemmes les plus fréquents sont, dans l'ordre, *utiliser* (4 896 occurrences), *organiser*[4] (4 657), *réaliser* (3 634), *favoriser* (2 856) et *qualifier* (2 225). Nous avons trouvé au moins une occurrence de 725 des 1 701 lemmes verbaux distincts couverts par notre lexique, soit moins de 43%, dont 190 sont des hapax (11%).

Pour détecter des lemmes candidats à constituer de nouvelles entrées, nous avons cherché les occurrences de formes correspondant à des étiquettes morphosyntaxiques fréquentes: infinitif, troisième personne du singulier et du pluriel du présent et de l'imparfait, participe passé. Nous avons éliminé certaines formes manuellement au moyen de motifs d'exclusion (par exemple, pas de lemme en *-criser* pour ne pas récupérer de candidats de type *anti-criser* ou *méga-criser*...). Nous avons ainsi détecté 141 lemmes candidats, qu'une validation manuelle a permis de répartir comme indiqué à la table 1. On constate que sur ces 141 candidats, 98 (70%) sont corrects (tout en étant absents de notre lexique et donc des trois lexiques de départ), dont 33 (21%) sont attestés par au moins une autre forme qu'un participe passé.

Une fois mis de côté les verbes à préfixe (avec ou sans tiret, cf. *sur-qualifier* et *hyperrationaliser*), notre mécanisme d'extraction de la base adjectivale ou nominale la plus probable, qui ne peut évidemment pas utiliser d'informations en provenance de LVF, propose toutefois une base dans 42 des 51 cas, dont seulement 3 sont incorrectes. C'est ainsi, par exemple, que le

[4] Il s'agit étymologiquement d'un verbe dérivé en *-iser* (*pourvoir d'organes*, selon le TLFi), bien que le sens ait glissé depuis.

	Lemmes simples	Lemmes à préfixe avec tiret
Candidats valides	20 (14%) *muséifier*	13 (9%) *co-organiser*
Candidats valides attestés seulement par des participes passés	57 (40%) *roumaniser*	8 (6%) *sur-qualifier*
Candidats provenant d'une faute d'orthographe	28 (20%) *rationnaliser*	–
Candidats erronés	10 (7%) *chalandiser*	4 (3%) *ex-coloniser*
Cas particuliers (créole)	1 (<1%) *mékaniser*	–

TAB. *1. Répartition des candidats lemmes verbaux dérivés en* -iser/-ifier *extraits du corpus et absents des ressources existantes.*

néologisme *littérariser*, absent des ressources, est correctement identifié à partir de la forme *littérarisé*, et correctement considéré par nos outils comme un dérivé désadjectival de *littéraire* selon le patron de dérivation *-ire* > *-riser* (dont nos outils ont su extraire qu'il ne s'appliquait qu'à des bases en *-aire* et *-oire*, ce qui est bien le cas ici).

6. Conclusion et perspectives

Nous avons l'intention, à très court terme, de valider manuellement le lexique obtenu par fusion et les entrées découvertes sur corpus, à la fois en morphologie et en syntaxe, puis de les intégrer au Le*fff*, afin d'en améliorer la couverture et la qualité. Un travail à plus long terme serait de les intégrer au lexique sémantique WOLF (WOrdnet Libre du Français) (Sagot & Fišer 2008), en utilisant par exemple le parallèle (supposé) avec les verbes en *-ize* et *-ify* du Princeton WordNet (Fellbaum 1998).

Plus généralement, cette étude a démontré que l'exploitation du LVF avec un objectif de traitement automatique est prometteuse, car la ressource est très couvrante. Il faut cependant pour cela contourner certains problèmes qu'elle pose, comme le fait de ne pas distinguer les compléments prépositionnels des circonstants ou encore le traitement indifférencié des différents types de constructions pronominales.

Bibliographie

Blanche-Benveniste C., Delofeu J., Stefanini J. et Eynde K. v. d. (1984), *Pronom et syntaxe. L'approche pronominale et son application au français*, SELAF, Paris.

Dal G. et Namer F. (2000), « Génération et analyse automatiques de ressources lexicales construites utilisables en recherche d'informations », in: *TAL*, vol. 41-2.

Danlos L. et Sagot B. (2007), « Comparaison du Lexique-Grammaire des verbes pleins et de DICOVALENCE : vers une intégration dans le Lefff », in: *Actes de TALN'2007*, Toulouse, France.

Danlos L. et Sagot B. (2008), « Constructions pronominales dans Dicovalence et le lexique-grammaire – Intégration dans le Le*fff* », in: *Actes du 27ème Colloque Lexique et Grammaire*, L'Aquila, Italie.

Dubois J. et Dubois-Charlier F. (1997), *Les verbes français*, Larousse-Bordas, Paris, France.

Fellbaum C. (1998), *WordNet: An Electronic Lexical Database*, MIT Press.

Namer F. (2002), « Acquisition automatique de sens à partir d'opérations morphologiques en français: études de cas », in: *Actes de Traitement Automatique du Langage Naturel (TALN)*, Nancy, France : 235–44.

Namer F. (2003), « Automatiser l'analyse morpho-sémantique non affixale: le système DériF », in: *Cahiers de Grammaire*.

Sagot B., Clément L., de La Clergerie É. et Boullier P. (2006), « The Le*fff* 2 syntactic lexicon for French: architecture, acquisition, use », in: *Proceedings of the 5th Language Resource and Evaluation Conference*, Lisbonne, Portugal.

Sagot B. et Danlos L. (2008), « Méthodologie lexicographique de constitution d'un lexique syntaxique de référence pour le français », in: *Actes du colloque Lexicographie et informatique: bilan et perspectives*, Nancy, France.

Sagot B. et Fišer D. (2008), « Building a free French wordnet from multilingual resources », in: *Actes de Ontolex 2008*, Marrakech, Morocco, (à paraître).

Sagot B. et Fort K. (2007), « Améliorer un lexique syntaxique à l'aide des tables du lexique-grammaire – Adverbes en *-ment*. », in: *Actes du 26ème Colloque Lexique et Grammaire*, Bonifacio, France.

van den Eynde K. et Mertens P. (2006), *Le dictionnaire de valence* DICOVALENCE *: manuel d'utilisation*, <http://bach.arts.kuleuven.be/dicovalence/manuel_061117.pdf>.

Willems D. (1979), « Syntaxe, morphosyntaxe et sémantique. Les verbes dérivés », in: *Cahiers de Lexicologie Paris*, vol. 35.