T.A.L., 2000, vol.

GÉNÉRATION ET ANALYSE AUTOMATIQUES DE RESSOURCES LEXICALES CONSTRUITES UTILISABLES EN RECHERCHE D'INFORMATIONS

Georgette DAL[*] **Fiammetta NAMER** [**]

Résumé - Abstract

En recherche documentaire, on se trouve sans cesse confronté à des mots absents des dictionnaires de langue générale. Or, très souvent, ces mots (i) sont construits, (ii) relèvent de langues dites de spécialité. Donner des outils pour les répertorier et les analyser peut par conséquent permettre d'enrichir une base terminologique. L'objectif de cet article est justement de présenter un système de génération et d'analyse automatiques d'unités lexicales construites non attestées par les dictionnaires, assorties d'informations constructionnelles et sémantiques.

A frequent problem in documentary research comes from large number of words encountered which are not listed in general language dictionaries. However, often, these words (i) are morphologically complex, (ii) belong to specialized language uses. Consequently, tools for listing and analyzing such words can enrich a terminological database. The purpose of this article is to present a system for automatic generation and analysis of morphologically complex lexical items which are not listed in dictionaries, which furthermore provides structural and semantic information about these items.

Mots clefs – Keywords

Morphologie - Racinisation à base de règles – Génération automatique de ressources lexicales

Morphology - Rule-based stemming - Derivational analysis – Automatic terms generation

* UMR 8528 " SILEX ", CNRS & Université de Lille 3 - 59653 Villeneuve d'Ascq cedex. Courriel : dal@univ-lille3.fr.

** " LANDISCO ", Université de Nancy 2 - BP 3397 - 54015 Nancy cedex. Courriel : namer@clsh.univ-nancy2.fr.

© copyright ATALA

Georgette DAL, Fiammetta NAMER

INTRODUCTION

En recherche documentaire, on se trouve sans cesse confronté à l'émergence de mots nouveaux, absents des dictionnaires de langue générale. Par exemple, alors que (*Le Monde* 1993), désormais *LM93,* et l'(*Encyclopedia Universalis* 1995), désormais *EU*, donnent à eux deux à observer 45 occurrences de l'adjectif *détectable*, c'est vainement qu'on cherchera ce mot dans le *Robert électronique* (*RE*), dans le *Trésor de la langue française* (*TLF*) ou dans le *Nouveau Petit Robert* (*NPR*), qui à eux trois assurent pourtant une bonne couverture du lexique attesté synchronique [1]. Par exemple encore, le nom *traçabilité,* également absent des trois dictionnaires cités, figure à 43 reprises dans (*Le Monde* 99), désormais *LM99*.

Or, très souvent, ces mots hors dictionnaires sont des unités lexicales construites [2] qui, en tant que telles, ont un sens prédictible à partir de leur structure. Par exemple, le sens construit de l'adjectif *détectable* est une fonction du sens instructionnel du suffixe *-able* appliqué au sens descriptif du verbe *détect(er)* - il marque la possession par le référent de son nom recteur d'une propriété latente activable par l'effectuation du procès qu'exprime le verbe *détect(er)* [3]-, le sens construit du nom *traçabilité* en est une du sens instructionnel du suffixe *-ité* appliqué au sens descriptif de l'adjectif *traçable* – à la manière d'un nom, il exprime la propriété qu'exprime sa base *traçable* -, et c'est bien comme tels que sont utilisés ces dérivés dans les citations suivante [c'est nous qui soulignons] :

> La théorie de l'information postule que, pour être détectable au mieux, un signal doit être émis dans une bande de fréquence extrêmement étroite. (*EU*, *s.v.* **exobiologie**)

> Le ministère entend organiser cette traçabilité totale des OGM, du champ jusqu'au produit fini. (*LM99*, 9 oct. 98, p. 10)

En plus d'être construits, ces mots relèvent majoritairement de langues dites de spécialité (technolectes scientifique, philosophique, médiatique, économique, etc.), si bien que donner des outils pour les répertorier et les analyser peut permettre d'enrichir une base terminologique.

L'objectif de cet article est précisément de présenter un système de génération et d'analyse automatiques d'unités lexicales construites *a priori* absentes les dictionnaires de langue générale, dans lequel chacune de ces unités s'assortit d'une double analyse, structurelle et sémantique.

Après avoir présenté le contexte global dans lequel s'inscrit le système de génération et d'analyse automatiques d'unités lexicales construites objet de cet article (§ 1.), nous ferons un état de l'art sur le traitement automatique

---

1 *Détectable* est en revanche reconnu par le vérificateur orthographique de *Word*97.
2 (Froissart C. & Lallich-Boidin G. 1996) notent que 32% des formes non reconnues par l'analyseur morphologique CRISTAL développé dans le cadre de l'action d'évaluation GRACE sont des mots construits ; si on exclut les erreurs typographiques et orthographiques, l'autre grand contingent des formes non reconnues est constitué de noms propres (cf. Maurel D. *et al.* 1996) et de sigles.
3 (Cf. Hathout N. *et al.* à paraître).

des unités lexicales construites absentes des dictionnaires (§ 2.) : cet état de l'art nous permettra de faire ressortir les motivations qui ont guidé la mise au point de notre système, baptisé GéDériF. Nous enchaînerons (§ 3.) en exposant la façon dont nous avons conçu notre générateur-analyseur en centrant notre propos sur trois opérations de construction d'unités lexicales : les suffixations par *-ité*, par *-able* et par *-is(er).* Nous terminerons le corps de cet article par une phase d'évaluation (§ 4.) avant de conclure (§ 5.).

1. SITUATION GÉNÉRALE

Le système de génération et d'analyse d'unités lexicales construites absentes des dictionnaires présenté ici est une émanation du projet *MorTAL* en cours de réalisation, ce dernier étant lui-même un enfant légitime de la plate-forme de réflexion constituée par le projet *FRANLEX* [4].

Le projet *MorTAL,* qui réunit Ch. Jacquemin (USR 705, CNRS-INaLF), N. Hathout (UMR 5610 " ERSS ", CNRS & Université de Toulouse-le Mirail) ainsi que les deux autrices du présent travail, bénéficie pour trois ans d'un financement accordé par le MENRT dans le cadre des actions concertées incitatives blanches 1999. Son objectif à terme est de construire de façon semi-automatique une base de données constructionnelles [5] du français pour le traitement automatique de la langue naturelle et la recherche d'information compilant l'ensemble des unités lexicales majeures attestées dans le *TLF* et dans le *RE* (pour une présentation circonstanciée du projet, cf. Hathout N. *et al*. à paraître, Dal G. *et al* soumission).

MorTAL part donc de l'attesté. Or, en commençant à constituer *MorTAL*, il nous est apparu qu'une partie des règles implémentées pour l'analyse automatique du lexique d'entrée était par ailleurs utilisable pour générer et analyser automatiquement des unités absentes de ce corpus : l'idée de notre générateur-analyseur automatique d'unités construites non attestées dans les dictionnaires était née.

2. ÉTAT DE L'ART, OU : QUEL TRAITEMENT POUR LES UNITÉS LEXICALES CONSTRUITES ABSENTES DES DICTIONNAIRES ?

2.1. Analyseurs basés sur dictionnaires

La plupart des - rares - systèmes automatiques fournissant des informations, même minimales [6], sur les unités lexicales construites prennent en entrée des lexiques finis. Citons par exemple pour le français le système mis au point par (Grabar N. *&* Zweigenbaum P. 1999) visant à construire par apprentissage une base de données morphologiques à partir de la

4 La version détaillée de *FRANLEX* figure sous l'URL : http://www.limsi.fr/Individu/jacquemi/FRANLEX/.

5 Comme D. Corbin (cf. Corbin D à paraître), nous préférerons l'adjectif *constructionnel* à *dérivationnel* parce qu'il inscrit explicitement le référent de son nom recteur dans le domaine de la **construction** des unités lexicales. Nous ne nous interdirons toutefois pas absolument l'adjectif *dérivationnel*.

6 Les ouvrages généralistes sur le TALN s'accordent à souligner la faible place accordée aux informations constructionnelles, considérées comme moins adaptées au domaine que les informations flexionnelles (cf. (Bouillon P. 1998 : 48 ; Fuchs C. éd. 1993 : 87)), certainement parce qu'elles sont réputées moins régulières que leurs homologues flexionnelles. Pour plus de détails, cf. (Sproat R.W. 1992 ; Fradin B. 1994).

nomenclature médicale SNOMED (*Systematized Nomenclature of Human and Veterinary Medecine*), ou encore l'analyseur mis au point par (Savoy J. 1993) : basé sur l'exploitation d'un gros lexique, son but est d'effectuer l'analyse morphologique complète des mots non étiquetés d'un texte soumis ensuite à des applications de RI. Le dernier des systèmes mentionné ici est le programme décrit dans (Gaussier E. 1999), qui acquiert par apprentissage des familles morphologiques hiérarchisées et des opérations de suffixation, à partir d'un lexique de formes fléchies étiquetées. Ce système est conçu pour servir à la fois en analyse (construire des procédures de racinisation) et en génération (assister le lexicographe dans le développement de nouvelles ressources lexicales).

Qu'ils produisent des ressources morphologiques ou qu'ils étiquettent des textes, ces divers systèmes fondés sur l'exploitation d'un lexique fini sont toutefois inopérants face à un terme absent de leur lexique d'entrée.

2.2. Analyseurs non-basés sur dictionnaires

Mais tous les analyseurs constructionnels ne se fondent pas sur l'exploitation d'un lexique fini. C'est notamment le cas des applications de désaffixation, dont le but est de constituer des familles morphologiques de mots caractérisés par une racine commune et des liens flexionnels et dérivationnels. On peut répartir ces applications en deux grandes familles : (i) les unes, basées sur règles, mettent majoritairement en œuvre l'algorithme de Porter ; (ii) les autres ont un fonctionnement purement ou principalement statistique, comme le programme Automorphology.

Ces systèmes ont les moyens de traiter les mots construits absents des dictionnaires. Toutefois, on peut en premier lieu reprocher à ces deux approches, qui ne distinguent pas flexion et dérivation lors du processus de troncation, d'être peu performantes en termes de lemmatisation. En outre, comme nous le montrons ci-dessous, ces approches génèrent soit du bruit, soit du silence lors de la constitution des familles constructionnelles.

2.2.1. Algorithme de Porter

L'algorithme de Porter (cf. (Porter M. 1980)) met en jeu un système de règles pondérées dont le mode d'application est fonction (i) du suffixe à supprimer (la notion de suffixe est ici purement géographique) [7], (ii) de caractéristiques de la base. Pour une langue à morphologie peu complexe comme l'anglais, cet algorithme a été jugé très performant en RI, à tel point que l'injection de traits linguistiques y a été jugée inutile (aucune amélioration en terme de résultats n'ayant été observée, notamment lors des expériences décrites dans (Lennon M. *et al.* 1981)). Pour des langues à morphologie plus complexe, comme le néerlandais, l'intégration de connaissances linguistiques y est au contraire préconisée (Kraaij W. *&* Pohlmann R. 1996).

En ce qui concerne l'adaptation au français, on observe que l'absence de connaissances linguistiques, et notamment l'absence de listes d'exceptions modulant l'activation des règles, conduit à la production de deux types d'erreurs :

1. Admettons que l'on pose les deux règles de désuffixation par *–aille*

7 À l'origine, la suppression ne concerne pas les préfixes.

(présent dans *cochonnaille, ferraille*) et par *–ite* (présent dans *aluminite*, *pierrite*). L'application de ces règles aux noms *ferraille* et *ferrite* aboutit licitement à la séquence initiale commune *ferr-*, qui va servir de clé au calcul de la famille constructionnelle {*ferraille*, *ferrite*}. Si, maintenant, on applique les mêmes règles aux noms *marmaille* et *marmite*, on voit le problème : ces deux noms, sur le seul partage de la séquence initiale *marm-*, se retrouvent constitués en une famille constructionnelle.

2. A l'inverse, les règles de désuffixation par *-ement* et *–er* [8], qui aboutissent logiquement à l'obtention de familles comme {*gonflement, gonfler*}, passent outre des paires comme {*achèvement*, *achever*}, {enlè*vement*, *enlever*}, dont les membres sont pourtant reliés.

Enfin, sans qu'il s'agisse là véritablement d'une erreur, le troisième inconvénient majeur de l'algorithme de Porter adapté au français est lié au traitement simultané des affixes flexionnels et dérivationnels, qui suppose la multiplication des règles (ainsi on duplique la règle de désuffixation de *-al* pour tenir compte de la désinence *-aux,* comme dans *national / nationaux*), et qui ajoute à la complexité de l'algorithme et nuit donc à son efficacité.

2.2.2. Le programme Automorphology

Contrairement au précédent, ce désaffixeur (cf. références en bibliographie) a un fonctionnement totalement probabiliste, prend en compte aussi bien les suffixes que les préfixes et analyse les textes de n'importe quel domaine et n'importe quelle langue européenne.

Le principe est le suivant : le système se fonde sur l'appariement de séquences identiques contenues dans des mots apparaissant dans le texte pris en entrée pour apprendre un certain nombre d'affixes. L'appariement est pondéré par des critères de fréquence : plus le texte est gros, plus les fréquences des mots contenant les séquences communes sont importantes et plus le résultat est fiable. Celui-ci est restitué sous la forme d'une mise en facteur de la séquence commune, associée à la famille des terminaisons regroupée en signature. Ainsi, l'analyse de *atomiser* et *atomique* aboutit à : *"atom" – "ique.iser"*.

D'après l'auteur, sur un texte d'au moins 100.000 mots, les résultats commencent à devenir satisfaisants. Cependant, un test que nous avons effectué sur un corpus de textes français de plus d'un million et demi de mots a montré qu'il subsiste un grand nombre d'analyses erronées : ainsi, *départ* et *département* sont analysés comme ayant une base commune (*départ*), de même que *me* et *mer*, ou *temps* et *temple* ; à l'inverse, des liens attendus ne sont pas effectués (par exemple, *région(s)* et *régional/aux*). Ces résultats confortent notre conviction qu'un analyseur constructionnel du français doit comporter un module linguistique.

2.3. Synthèse

On vient de voir qu'en général, les analyseurs constructionnels pour le français sont soit tributaires de l'utilisation d'un lexique (ils sont alors inopérants face à des mots en dehors de ce lexique), soit limités dans leurs résultats par l'absence de connaissances linguistiques : ils génèrent donc

8 On rappelle que l'algorithme de Porter traite à l'identique les suffixes constructionnels et flexionnels.

des résultats insuffisants ou incorrects. De plus, à notre connaissance, aucun ne propose d'analyse sémantique des mots construits. Or, une analyse sémantique, même sommaire, est une donnée majeure dans le cadre de la RI, puisqu'elle est un moyen d'offrir de nouveaux critères de recherche, en permettant l'expression des requêtes non seulement sous forme de termes mais aussi sous forme de prédicats subsumant ces termes : ainsi, avec un système qui calcule la relation sémantique entre les différents membres d'une famille morphologique, on peut imaginer une requête portant sur les "processus de transformation", qui engendrera une recherche à partir des mots construits en *Xifier* et *Xification*. De même, dans un texte scientifique portant sur les mathématiques ou la physique par exemple, on peut imaginer une requête qui fasse ressortir tous les noms exprimant une aptitude (propriété + qui peut être), de sorte à constituer un micro-lexique du domaine. À l'inverse, la présence de relations sémantiques constitue un filtre utilisable pour trier les documents résultants des recherches.

2.4. Ce qu'apporte notre approche

L'objectif de GéDériF est triple. Il est (i) de produire un lexique d'unités lexicales construites absentes des dictionnaires, et, comme le système DériF dont il émane (cf. (Namer F. 1999), (Dal *et al*. 1999), (ii) d'enrichir ce lexique d'informations constructionnelles et sémantiques, (iii) de constituer des micro-familles constructionnelles.

Comme les systèmes mentionnés sous 2.1., GéDériF part lui aussi d'un corpus fini puisque ce système emprunte une grande partie de ses entrées à *TLFnome* [9], complété systématiquement de façon manuelle par une consultation du *RE* et du *NPR*. Cependant, comme nous allons le faire apparaître dans la section suivante, le système que nous présentons ici est apte à analyser des mots hors dictionnaires, puisqu'à ces entrées entérinées par les dictionnaires viennent se greffer les propres sorties du système, et ce de façon possiblement récursive : le lexique qu'il prend en entrée est donc sans cesse renouvelé.

Contrairement aux systèmes précédemment cités, toutes approches confondues, GéDériF est en outre capable d'associer aux unités lexicales construites des informations constructionnelles et sémantiques linguistiquement motivées et contrôlées. Il offre enfin à l'utilisateur en RI une base de mots construits qu'il pourra exploiter du point de vue :

- de la famille du mot analysé, pour élargir ou filtrer ses requêtes,
- du sens induit par les opérations morphologiques mises en jeu, pour élargir ses possibilités de requêtes au moyen de nouveaux critères de recherche, par exemple les relations sémantiques.

Le fonctionnement et les résultats de GéDériF sont présentés et illustrés dans la section 3.2. Nous légitimerons auparavant linguistiquement dans la section 3.1. les choix de génération et d'analyse qui ont été faits.

9 *TLFnome* est un lexique de formes fléchies construit à l'INaLF à partir de la nomenclature du *Trésor de la Langue Française*. Il contient actuellement 63 000 lemmes, 390 000 formes et 500 000 entrées. Il est en cours de complétion grâce à 36 400 lemmes supplémentaires issus de l'index du *TLF*.

3. CONSTITUTION DU GÉNÉRATEUR-ANALYSEUR D'UNITÉS CONSTRUITES ABSENTES DES DICTIONNAIRES

Comme nous l'avons signalé plus haut, notre générateur-analyseur d'unités construites *a priori* absentes des dictionnaires de langue est un système construit parallèlement au projet *MorTAL* : s'il se nourrit des opérateurs de construction d'unités lexicales implémentés dans *MorTAL*, il en est par conséquent également tributaire. Cela explique qu'il ne peut travailler pour l'heure que sur les suffixes *-et(te), -ifi(er)*, *-able*, *-ité* et *-is(er)* ainsi que sur les préfixes *dé-* et *in-*. Pour éviter de nous éparpiller, dans la suite de cet article, nous centrerons nos propos sur les trois derniers suffixes mentionnés et sur les combinaisons qu'ils autorisent.

3.1. Légitimation linguistique

Un système de génération et d'analyse automatiques d'unités lexicales construites absentes des dictionnaires est par nature un mécanisme surgénérateur. Il convient toutefois de contrôler linguistiquement cette surgénération : ainsi, un générateur qui produirait **infabricagiste* (cf. Gruaz C. *et al*. 1996), qui est un monstre linguistique, serait trop puissant et par là même inadapté.

Aussi avons-nous veillé à ce que les résultats produits par GéDériF soient linguistiquement motivés, d'un double point de vue structurel et sémantique. C'est ce que nous ferons apparaître dans ce qui suit en commençant par caractériser les suffixations par *-able*, *-ité* et *-is(er)* (pour autant que cela intéresse notre propos), puis en exposant les combinaisons deux à deux possibles qu'autorisent ces trois opérateurs.

3.1.1. Les suffixations par *-able*, *-is(er)* et *-ité* : aperçu

Le suffixe *-able* français ne forme qu'un type catégoriel de dérivés, des adjectifs, mais peut sélectionner deux types catégoriels de bases : des verbes ($port(er)_V \rightarrow portable_A$) et des noms ($ministre_N \rightarrow ministrable_A$). Le commun dénominateur sémantique des $Xable_A$ est de marquer la possession par le référent de leurs noms recteurs d'une aptitude susceptible de se voir révéler par la réalisation d'un procès, exprimé par la base dans le cas d'une dérivation à partir d'un verbe (pour plus de détails, (cf. Dal G. *et al.* 1999 ; Hathout N. *et al.* à paraître)).

Le suffixe *-is(er)* construit des verbes, et opère lui aussi à partir de deux types catégoriels de bases : des adjectifs ($adverbial_A \rightarrow adverbialis(er)$) et des noms ($bémol_N \rightarrow bémolis(er)$). Comme son homologue anglais *-ize* (cf. Plag I. 1997), *-is(er)* se caractérise par un sens instructionnel au spectre large. Cette caractéristique est plus ou moins manifeste selon la catégorie de la base :

1. les verbes en *-is(er)* désadjectivaux expriment un changement d'état pour le référent de leur complément d'objet direct. Cependant, étant donné le moule catégoriel dans lequel se coulent ces verbes (A $\rightarrow$ V), cette caractérisation est entièrement prédictible, indépendamment de l'opérateur constructionnel *–is(er)* : tous les verbes dérivés d'adjectifs expriment en effet un changement d'état, quelle que soit l'opération

constructionnelle en jeu (conversion : *rouge*$_A$ → *roug(ir)*$_V$; suffixation par *-ifi(er)* : *acide*$_A$ → *acidifi(er)*$_V$; préfixation : par *a-* (*pauvre*$_A$ → *appauvr(ir)*$_V$), par *dé-* (*niais*$_A$ → *déniais(er)*$_V$) etc.), parce que c'est là la seule relation sémantique instaurable entre un adjectif et un verbe qui en dérivé.

2. Les verbes en *-is(er)* dénominaux peuvent exprimer des procès variés, selon le sens et/ou les caractéristiques référentielles du nom de base :
   - Si la base est un nom propre référant à un individu au comportement singulier, le verbe, alors intransitif, exprime le procès de se conduire à la manière du référent du nom de base (ex. *diogéniser* : " 1. Vivre dans un dénuement matériel absolu; ou professer le cynisme (comme Diogène). " (*RE*)).
   - Si la base est un nom propre référant à un individu réputé par son œuvre (littéraire, politique, etc.), le verbe, alors transitif, exprime le procès de donner au référent de son COD le caractère typique de l'œuvre en question (cf. (*La banque des mots* (désormais *BDM*) 7 :105) : " ici ou là on s'efforce de brechtiser […] Molière, ailleurs c'est Brecht qu'on moliérise ").
   - Si la base réfère à une action, à un utilitaire ou à un lieu, le verbe exprime le procès de soumettre le référent de son COD à l'action indiquée par la base, à l'utilitaire ou au lieu auquel elle renvoie (*bémoliser*; *hospitaliser*).
   - …

La liste qui précède ne prétend pas être exhaustive ; elle suffit toutefois à montrer que le suffixe *-is(er)* est un suffixe verbalisateur par excellence apte à opérer sur des types sémantiques de bases divers et, davantage que le suffixe *-ifi(er)* plus strict quant aux types sémantico-référentiels des bases qu'il sélectionne, c'est plutôt la conversion qu'il concurrence [10].

Enfin, le suffixe *-ité* ne forme aussi qu'un type catégoriel de dérivés (des noms) et peut s'appliquer à deux types catégoriels de bases : à des adjectifs (*acerbe*$_A$ → *acerbité*$_N$) et, même si c'est plus rarement, à des noms (*édile*$_N$ → *édilité*$_N$). D'un point de vue sémantique, *-ité* forme des noms de propriétés qu'il présente comme objectives (cf. Corbin D. à paraître) :

- quand il opère sur des adjectifs, cette propriété est exprimable au moyen de l'adjectif de base. *Xité*$_N$ et *X*$_A$ n'ont évidemment pas le même sens, même quand *X*$_A$ n'exprime qu'une (des) propriété(s) objective(s) (quand il exprime en plus une propriété subjective, par exemple *sensible*, *-ité* n'active que la(es) propriété(s) objective(s) de la base : cf. *sensibilité* vs *sensiblerie*). De façon générale, là où un adjectif exprime une propriété qui a besoin d'un support pour se réaliser, un nom de propriété permet en effet d'évoquer cette propriété indépendamment de ses supports, comme le fait apparaître la citation suivante extraite de l'*EU* [c'est nous qui soulignons] :

10 On peut voir un symptôme de cette concurrence dans l'ensemble des doublets attestés suivant : *adultériser* / *adultérer, alinéatiser* / *alinéter, anagrammatiser* / *anagrammer, budgétiser* / *budgéter, cabaliser* / *cabaler, coaltariser* / *coaltarer*, *cotoniser* / *cotonner, maniériser* / *maniérer, platiniser* / *platiner, silicatiser* / *silicater*, *texturiser* / *texturer, tuberculiniser* / *tuberculiner, voyelliser* / *voyeller*.

> La structure de la caisse doit assurer un bon compromis entre : rigidité en flexion et en torsion, protection des passagers lors de collisions, habitabilité et légèreté. (*s.v.* **automobile**)

i) quand il opère sur des noms, *-ité* extrait de leur sens (ou de leur contenu [11] s'il s'agit de noms propres, par exemple *matissité* (cf. Dal G. 1997a)) une (quelques) propriété(s) objective(s).

3.1.2. Combinaisons possibles

Comme nous l'avons annoncé plus haut, nous avons appliqué GéDériF aux combinaisons qu'autorisent les trois suffixes dont il vient de s'agir. Dans cette section, nous examinerons conjointement quelles combinaisons sont *a priori* possibles et lesquelles sont effectivement possibles d'après l'observation du lexique attesté dans les dictionnaires, et d'après une analyse sémantique (nous exclurons d'emblée les cas d'auto-concaténation, le français interdisant l'application contiguë de deux opérateurs de même forme).

3.1.2.1. [[*Xis(er)*] *able*] / *[[*Xité*] *able*]

Puisqu'il peut opérer à la fois sur des verbes et sur des noms (cf. supra, § 3.1.1.), le suffixe *-able* peut en première analyse sélectionner comme bases les produits de la suffixation par *-is(er)* ou ceux de la suffixation par *-ité*.

Ces configurations sont en outre en première analyse sémantiquement légitimes (rappelons que, conformément au cadre linguistique théorique adopté (cf. Corbin D. à paraître), on considère ici que les opérations constructionnelles satisfont d'abord des contraintes sémantiques dont les contraintes catégorielles peuvent être vues comme des avatars) :

- **i**) Les *Xis(er)* sont majoritairement agentifs et transitifs. Or, quand il opère sur des verbes, *-able* sélectionne des verbes agentifs présentant au moins un argument interne (typiquement COD, mais aussi argument de type locatif : cf. *circulable, skiable*). Nous avons par conséquent décidé de programmer GéDériF pour qu'il génère des adjectifs de structure [[Xis(er)] able], sans restriction, et comme nous l'avons fait dans *MorTAL* pour les adjectifs déverbaux en *-able* attestés, de leur associer la paraphrase " (<Prep> lequel/Que l') on peut V ".
- **i**) Les noms en *–ité* sont des noms de propriété. Or, le lexique attesté donne à observer des adjectifs en *–able* construits sur des noms de propriété, en *–ité* (*charité* → *charitable ; équité* → *équitable*) ou non (*faveur* → *favorable* ; *misère* → *misérable*). Malgré cela, nous avons rejeté la structure [[Xité] able], pas tant parce que ce cas de figure ne concerne dans le lexique attesté par les dictionnaires qu'une poignée de dérivés que parce que le plus récent de ces *Xité* avec *X* = nom de propriété date du 16e siècle. Cette décision est en outre confortée par le caractère intuitivement monstrueux d'adjectifs comme **aviditable* ou

11 (Gary-Prieur M.-N. 1994) résout le problème épineux du sens du nom propre en parlant de contenu, défini comme l'" ensemble de propriétés attribuées au référent initial de ce nom propre dans un univers de croyance " (p. 51).

**rigiditable* construits sur des noms en *-ité* sémantiquement comparables à *charité* et *équité* (comme eux, ce sont des noms de sentiments).

3.1.2.2. *[[*Xable*] *is(er)*] / *[[*Xité*] *is(er)*]

Les dérivés en *-able* et en *-ité* satisfont les exigences catégorielles qu'a le suffixe *-is(er)* envers les bases qu'il sélectionne (adjectifs ou noms : cf. § 3.1.) : les deux structures sont donc catégoriellement licites.

D'un point de vue sémantique en revanche, on ne retiendra pas comme bases possibles de la suffixation par *–is(er)* les noms en *-ité*, dont le type sémantique ne correspond à aucun des types sémantiques inventoriés plus haut : le caractère monstrueux de séquences comme **acerbit(é)iser, *minéralit(é)iser* confirme le rejet de la structure [[Xité] is(er)].

Même si c'est moins immédiatement perceptible, l'application du suffixe *-is(er)* à des adjectifs en *-able* pose elle aussi un problème sémantique. En effet, les adjectifs en *-able* expriment des propriétés latentes, donc des propriétés endogènes au référent de leurs noms recteurs. Or, dans les *Xis(er)* dérivés d'adjectifs, X_A précise dans quel état se trouve l'entité sur laquelle a porté le procès. Il s'ensuit que X_A doit pouvoir exprimer une propriété susceptible de résulter de la réalisation d'un procès, donc une propriété exogène. La pénurie dans le lexique attesté de *Xabilis(er)* où *Xable* exprime une propriété latente [12] serait ainsi due à une incompatibilité sémantique entre les *Xable* et les exigences que fait peser *-ise(er)* envers les adjectifs qu'il sélectionne.

Devant la rareté dans les dictionnaires de verbes en *-is(er)* dérivés d'adjectifs en *–able* exprimant clairement une aptitude, et nous fondant sur le blocage sémantique que nous avons cru déceler, nous avons décidé d'interdire à GéDériF d'appliquer le suffixe *-is(er)* à des adjectifs en *–able*, quelle que soit la structure de ces derniers. Pour vérifier la validité de ce choix, nous avons interrogé automatiquement le moteur de recherche yahoo.fr à l'aide de 1287 verbes en *-abilis(er)* créés pour les besoins de la démonstration (par exemple, *abolissabilis(er)*, *abordabilis(er)*, etc.). Seules 6 de ces 1287 créations (approximativement 0,5%) ont eu des résultats positifs (*commutabilis(er)*, *notabilis(er)*, *portabilis(er)*, *potabilis(er)*, *sociabilis(er)* et *variabilis(er)*). Le silence que nous engendrons en refusant cette structure est par conséquent négligeable eu égard au bruit que nous engendrerions si nous la retenions.

3.1.2.3. [[*Xable*] *ité*] / *[[*Xis(er)*] *ité*]

Le suffixe *-ité*, on l'a dit, peut opérer sur des adjectifs et sur des noms. D'un point de vue catégoriel, sont donc d'entrée exclus les outputs de la suffixation par *-is(er)*.

Les produits de la suffixation par *-able* sont, eux, à la fois catégoriellement et sémantiquement licites : catégoriellement parce que ce sont des adjectifs ; sémantiquement parce qu'il y a unification sémantique

12 Moins de 2% des 700 *Xis(er)* dérivent d'adjectifs en *–able* : *amabilis(er)*, *comptabilis(er)*, *culpabilis(er)*, *fiabilis(er)*, *malléabilis(er)*, *navigabilis(er)*, *perméabilis(er)*, *probabilis(er)*, *rentabilis(er)*, *respectabilis(er)*, *responsabilis(er)*, *stabilis(er)*, *viabilis(er)*.

possible entre les adjectifs en *-able* et le suffixe *-ité*, qui exige des bases exprimant des propriétés objectives. En première analyse, tout adjectif en *-able* peut donc donner lieu à un nom de propriété en *-ité* dès lors que la propriété qu'il exprime est concevable indépendamment de ses supports, comme le confirme le nombre de dérivés en *-abilité* attestés par les dictionnaires qui nous servent de référents : 214 des 1409 des noms en *-ité* que nous avons collectés (= 15,2 %) dérivent d'adjectifs en *-able*.

Avant d'autoriser GéDériF à appliquer automatiquement le suffixe *-ité* à tout *Xable* quelle qu'en soit la structure, nous nous sommes penchées sur le cas particulier des adjectifs en *–isable*.

En effet, si ce qui vient d'être dit est juste, d'un point de vue purement linguistique, les *Xisable* , qui ne sont qu'un cas particulier des *Xable*, peuvent servir d'inputs à la suffixation par *-ité*. Toutefois, curieusement, on s'aperçoit que seuls 2 des 214 noms de propriété en *-abilité* collectés dans les dictionnaires (*hypnotisabilité* et *polarisabilité*) ont pour base un adjectif comportant le suffixe *-is(er)* dans sa structure, alors que ces mêmes dictionnaires offrent quelque 70 candidats au rôle de base (par exemple, *cicatrisable*, *fertilisable*). Nous avons donc cherché des explications à cette absence quasi-chronique de noms en *-isabilité* dans les dictionnaires, de façon à déterminer si la génération automatique de ce type de noms devait ou non être autorisée.

La seule raison que nous soit apparue relève de la performance : nous faisons l'hypothèse que la faible proportion dans les dictionnaires de langue générale de *Xisabilité* est imputable au fait que ces mots mettent en jeu trois opérations de construction d'unités lexicales successives, voire quatre si la base adjectivale du verbe en *–is(er)* est elle-même construite (par la suffixation par *–al,* par *–ien*, etc.). Cette multiplicité d'opérateurs peut provoquer un problème au niveau du calcul sémantique, si bien qu'on peut lui préférer des stratégies palliatives : à " La fertilisabilité d'une terre dépend de son ensoleillement ", on préférera peut-être " La possibilité de fertiliser une terre […] ". Cependant, comme la raison du blocage n'est pas à proprement parler linguistique, et comme ce problème de calcul de sens est visiblement soluble (pour preuve, *hypnotisablité* et *polarisabilité*), nous avons décidé de générer également les *Xisabilité*, à juste raison semble-t-il (cf. § 4.1.).

Comme leurs homologues attestés déjà analysés, les noms de structure [[*Xable*] *ité*] générés reçoivent automatiquement la glose " Propriété de ce qui est Xable ".

3.2. Le système GéDériF

GéDériF est un système à deux composantes, dont l'entrée est un lexique d'environ 70 000 formes non fléchies étiquetées au moyen du jeu d'étiquettes de Brill (cf. (Brill E. 1994) et (Lecomte J. *&* Paroubek P. 1996)). La première composante génère automatiquement de nouvelles entrées lexicales en fonction d'un certain nombre de critères (cf. § 3.2.1), et la seconde (cf. § 3.2.2) effectue l'analyse constructionnelle des mots du nouveau lexique ainsi constitué, en produisant en sortie un lexique enrichi d'informations constructionnelles et sémantiques.

3.2.1. Génération du lexique de mots possibles

Le générateur construit des unités lexicales par affixation d'unités lexicales existantes. Cette tâche suppose que les unités de départ soient sémantiquement compatibles avec l'opérateur d'affixation. Etant donné l'absence actuelle de traits sémantiques dans les entrées lexicales monoconstituées, le générateur ne peut donc choisir la base des unités qu'il va générer automatiquement que parmi les mots déjà construits. En effet, ces derniers ont un sens calculable à partir de l'instruction sémantique de l'affixe qui les a construits.

Dans le cas qui nous intéresse, le générateur sélectionne les mots du lexique en *-able* et en *-is(er),* et leur concatène [13] toutes les combinaisons licites de *-able, -is(er)* et *-ité* qui sont complémentaires à leur terminaison [14]. Ainsi, parmi l'ensemble des combinaisons qu'autorisent *a priori* les unités lexicales de la forme *Xis(er)* et *Xable*, seules les possibilités entourées ci-dessous sont valides (pour une justification, cf. § 3.1.) :

	**is(er)*		**is(er)*
X(is)able	**able*	**Xis(er)**	***able Xisable***
	ité X(is)abilité		**ité*

On ne conserve ensuite parmi les résultats obtenus que ceux qui ne sont pas déjà dans le lexique d'entrée.

À partir des unités lexicales de la forme *Xable* et *Xis(er)* et des suffixes *-able* et *-ité*, GéDériF a ainsi produit un ensemble de 2691 mots construits absents du lexique d'entrée et néanmoins linguistiquement plausibles.

3.2.2. Fonctionnement de l'analyseur DériF

Le second composant de GéDériF est le module d'analyse constructionnelle DériF. Ce module s'applique sur les mots construits apparaissant dans le lexique d'entrée grossi des termes obtenus par génération automatique.

Le résultat de l'analyse d'un mot M est un triplet constitué de (1) l'arbre d'analyse morphologique de M sous forme crochetée, (2) l'ensemble des lemmes appartenant à la famille morphologique de M, et (3) le sens induit par le dernier affixe ayant opéré pour construire M, sous forme d'une relation sémantique entre M et sa base.

Ainsi, pour M = *absorbabilité*, le résultat est :

```
(1) [ [ [ absorber VERBE] able ADJ] ité NOM]
(2) (absorbabilité, absorbable, absorber)
(3) :: Propriété de ce qui est absorbable
```

13 L'opération de concaténation affecte éventuellement un allomorphe du morphème, obtenu au moyen de la règle d'appariement appropriée, e.g. : *-able* > *-abil-*. Ainsi, la génération d'*absorbabilité* se fait en deux étapes à partir d'*absorbable : absorbable* > *absorbabil-* → *absorbabilité*.

14 Les unités lexicales en *-ité* ne peuvent servir de base, car, on le rappelle, ce suffixe est sémantiquement incompatible avec les suffixation par *-able* et *-is(er)* (cf. supra, § 3.1.).

DériF a été déjà décrit (cf. (Namer F. 1999), (Dal G. *et al.* 1999)), aussi n'en ferons-nous ici qu'une brève présentation.

Comme l'illustre la Figure 1, le moteur examine le mot M en entrée, et appelle le cas échéant la fonction $\mathbf{F_S}$ spécifique de l'analyse du suffixe S de M. $\mathbf{F_S}$ recherche d'abord l'existence de préfixes portant sur la base suffixée de M : la fonction **PrefixeAvant** effectue récursivement la décomposition de M selon le préfixe pref et base suffixée B, avec, le cas échéant, le calcul de l'allomorphe B_M de B. Puis $\mathbf{F_S}$ tronque B_M selon son suffixe S, procède aux calculs sémantiques, catégoriels, aux appariements entre allomorphes et renvoie le résultat au moteur. Celui-ci réitère son application jusqu'à l'obtention du primitif (*i.e.* de la séquence structurellement simple), et affiche les résultats.

Figure 1 : fonctionnement de DériF

La tâche la plus importante incombe donc aux fonctions $\mathbf{F_S}$, que nous illustrons ci-dessous.

3.2.3. Analyse de *Xité* et *Xis(er)*

Le fonctionnement détaillé de la fonction spécifique du suffixe -*able* a été décrit par ailleurs. Aussi ne présentons-nous ici brièvement que l'algorithme sous-jacent aux fonctions $F_{ité}$ et $F_{is(er)}$ analysant respectivement les unités lexicales construites de structure *Xité* et *Xis(er)*.

Les suffixes –*ité* et -*is(er)* peuvent opérer sur des bases adjectivales qui ont des propriétés sémantiques proches (dans le lexique attesté par les dictionnaires, un même adjectif sert d'ailleurs parfois d'input à ces deux suffixations : par exemple, *absolu*, *adverbial* qui donnent lieu à respectivement *absoluité* et *absolutis(er), adverbialité* et *adverbialis(er)*). Les bases que sélectionnent ces suffixes sont en outre soumises à des variations allomorphiques semblables. C'est ainsi que les fonctions $F_{ité}$ et $F_{is(er)}$ ont en commun une grande partie du système d'appariement des variantes allomorphiques de la base X résultant de la troncation de -*ité* ou -*is(er)*, en fonction de la séquence finale de X, *modulo* un certain nombre d'exceptions. Certaines de ces variations sont illustrées ci-dessous :

Xitser ou Xité	base	Règle d'appariement
régular#iser / régular#ité	régulier	ar <-> ier
sonor#iser / sonor#ité	sonore	or <-> ore
african#iser / african#ité	africain	an <-> ain

D'autres variations mettent en lumière une unité infralexicale (désormais, UIL) identifiant une base allogène (par exemple, -*virgin*- dans *virginité* ou *virginis(er)*). Le programme la retrouve dans une table spéciale qui inventorie l'ensemble des bases infralexicales absentes du référentiel venant du latin, du grec, de l'allemand, etc., leur traduction, éventuellement approximative, ainsi que la catégorisation grammaticale de cette traduction :

UIL (catégorie = FWD)	traduction de l'UIL	catégorie de la traduction
-virgin-	*vierge*	A

L'UIL est alors conservée dans la partie analyse du résultat affiché (avec l'étiquette FWD : "mot étranger"), alors que sa traduction apparaît dans la seconde et troisième partie du résultat (cf. § suivant).

3.2.4. Résultats

Nous concluons cette brève présentation de GéDériF par la description des différents types de résultats obtenus.

3.2.4.1. Arbre d'analyse

Dans DériF, l'arbre d'analyse des unités décrites prend la forme d'une structure crochetée étiquetée. Par exemple,

a) `biodégradabilité =>[[[bio NOM][[dégrader VB]able ADJ]ADJ] ité NOM] (biodégradabilité, biodégradable, dégradable, dégrader)` [15]
b) `revaloriser =>[re [[ valeur NOM] is(er) VB] VB] (revaloriser, valoriser, valeur)`
c) `immuabilité => [ [ in [[ muer VB] able ADJ] ADJ] ité NOM] (immuabilité, immuable, muable, muer)`
d) `dévirginiser =>[dé+ [-virgin- FWD] +is VB] (dévirginiser, -virgin- = 'vierge')`

Les représentations sous formes de structures crochetées évitent de réduire les unités lexicales construites à de simples concaténations d'opérateurs constructionnels et d'unités à sens descriptif (ce que ferait une représentation plate : par exemple, *immuabilité/NOM* : *in* + *muer* + *able* + *ité*) : DériF hiérarchise l'application des opérateurs ou opérations constructionnel(le)s, et restitue en clair cette hiérarchisation dans la famille constructionnelle entre parenthèses constituée en parallèle.

Ainsi, la structure crochetée (a) associée à *biodégradabilité* met en évidence une opération de suffixation par *–ité* postérieurement à une opération de composition appliquée à un adjectif suffixé.

Les deux analyses suivantes (b) et (c) font apparaître l'ordre relatif des opérations de suffixation et de préfixation : suffixation PUIS préfixation pour *revaloris(er)* (le schéma crocheté (b) rend compte du fait que le préfixe *re-* porte sur le verbe suffixé *valoris(er)* et qu'il constitue l'affixe le plus périphérique) ; préfixation PUIS suffixation pour *immuabilité* (dans ce dérivé, le suffixe *-ité* porte sur la base préfixée *immuable*).

Enfin, l'exemple (d) illustre un cas de préfixation déclenchant l'apparition d'un " marqueur de classe " (cf. notamment Corbin D. à paraître), c'est-à-dire d'une finale suffixoïde dépourvue de sens instructionnel, n'entrant par conséquent pas dans la procédure du calcul du sens du dérivé dans lequel elle apparaît : dans *dévirginis(er)*, *-is(er)* n'est pas un suffixe (le sens de ce verbe est une fonction de l'opérateur préfixal *dé-* appliqué à l'adjectif infralexical *-virgin-*). Son seul rôle est de marquer iconiquement

15 Des considérations sémantiques, que nous ne développerons pas ici, nous incitent à analyser *biodégradabilité* comme un nom de propriété dérivé par suffixation de l'adjectif composé *biodégradable*.

l'appartenance du produit de la préfixation par *dé-* de l'adjectif *–virgin-* à la classe des verbes (dans ce cas précis, on observe d'ailleurs que *dévirgin(er)* est également attesté dans le *RE*, et que le procès qu'il exprime est donné comme coréférent de celui de *dévirginis(er)*). Dans DériF, les préfixes déclenchant l'apparition d'un marqueur de classe sont suivis du signe " + ". Ce même signe est repris à l'initiale du suffixoïde concerné.

3.2.4.2. Glose

Dans le modèle linguistique sous-jacent à DériF, les unités lexicales construites reçoivent des définitions métalinguistiques (pour les raisons de ce choix, cf. notamment (Corbin 1993) et (Dal 1997b)). Les gloses assignées automatiquement aux entrées de DériF sont, elles, délibérément formulées en langue naturelle, pour que leur exploitation en TAL et en RI soit possible, quelles que soient les décisions formelles prises.

Dans DériF, les gloses reflètent le rôle sémantique de l'opération constructionnelle la plus périphérique à partir de la catégorie lexicale de la base (ou de la traduction de celle-ci, dans le cas de base marquée FWD : cf. l'exemple de *virginité* au § précédent) : en ce sens, elles relèvent des définitions dites "dérivationnelles" (Martin 1992). Bien que les gloses ne traduisent le rôle sémantique que de la dernière opération constructionnelle intervenue, les informations sémantiques des niveaux inférieurs sont le cas échéant récupérables et exploitables.

On prendra ici comme exemple l'unité construite *fertilisabilité*, qui met en jeu les trois suffixes étudiés dans cet article.

Fertilisabilité est une unité lexicale de type X_{ADJ}-ité/NOM. La glose qui lui correspond est par conséquent une instance de la glose générique "propriété de ce qui est X_{ADJ}" associée aux noms produits par l'application de *-ité* à des adjectifs. L'analyse complète de *fertilisabilité* dans GéDériF est donc :

```
fertilisabilité => [[[[ fertile ADJ]is(er) VB]able ADJ]ité NOM]
(fertilisabilité, fertilisable, fertiliser, fertile)
      :: Propriété de ce qui est fertilisable
```

Fertilisable est une unité lexicale de type X_{VBE}-able/ADJ. Les dérivés de ce type expriment la possibilité, pour le référent du nom qu'ils qualifient, de se voir appliquer le procès qu'exprime *X*, ce que traduit en première approximation la glose "que l'on peut V". Appliquée à *fertilisable*, cette glose prend la forme :

```
fertilisable/ADJ : [ … ]  ( … ) :: Que l'on peut fertiliser
```

Enfin, la glose correspondant à une unité lexicale de type X_{ADJ}-is(er)/VBE reflète en langue naturelle l'instruction sémantique associée à la suffixation par *–is(er)* quand elle opère sur des adjectifs (cf. § 3.1.1.). Une façon de traduire le sens construit du dérivé est "rendre X_{ADJ}". Dans le cas particulier de *fertilis(er)*, la glose est :

```
fertiliser/ADJ : [ … ],  ( … ), :: Rendre fertile
```

De proche en proche, on reconstruit de la sorte la relation sémantique

existant entre le nom *fertilisabilité* et son primitif *fertile*. Cette relation est représentable par la notation semi-formelle suivante :

```
sens_de(fertilisabilité)=
propriété_de_ce(que_l_on_peut(rendre(fertile))))
```

Le système GéDériF n'est donc pas un simple générateur d'unités lexicales construites *a priori* absentes des dictionnaires de langue générale. Il permet aussi d'associer, de façon également automatique, une structure étiquetée et une glose à chaque unité qu'il génère.

4. ÉVALUATION DES RÉSULTATS

Nous avons procédé à une double série de tests (quantitatifs et qualitatifs) pour évaluer les résultats qui viennent d'être présentés. D'une part, nous avons chiffré la proportion des 2691 termes inventés au moyen des combinaisons licites des suffixes *-ité, -able* et *–is(er)* qui apparaissent réellement dans les documents. D'autre part, nous avons vérifié manuellement la validité des analyses (formelles, structurelles et sémantiques) produites automatiquement par l'analyseur de GéDériF sur l'ensemble de ces unités lexicales construites.

4.1. Evaluation quantitative

4.1.1. Deux types de recherches

Afin de valider le lexique préfiltré linguistiquement (section 3.1.2.) obtenu en sortie du générateur de GéDériF (section 3.2.1.), nous avons procédé à deux séries de vérifications, qui permettent d'établir dans quelle proportion ces 2691 termes se retrouvent effectivement dans les documents.

Tout d'abord, nous avons vérifié systématiquement la présence des éléments de ce lexique dans une version électronique de l'*Encyclopedia Universalis*, et dans l'ensemble des numéros-papier de la revue de terminologie la *Banque des Mots*, qui puise dans des sources variées (revues scientifiques, économiques, etc.). Le choix de ces deux ressources a été fait en raison de leur caractère représentatif pour des genres différents, ce qui est propice à l'apparition de technolectes divers. Pour nous assurer que l'*EU* et la *BDM* favorisent effectivement l'émergence de termes construits non attestés dans des dictionnaires, nous avons procédé à une vérification automatique témoin sur deux corpus de textes de 8Mo chacun, contenant respectivement les articles du journal *Le Monde* de l'année 1992 et des notices bibliographiques appartenant au domaine de l'agroalimentaire, extraites de la base PASCAL [16].

En parallèle, nous avons conçu un script qui a interrogé automatiquement le moteur de recherche www.yahoo.fr avec chacun des termes générés. Le message fourni en réponse à ces requêtes mentionne le cas échéant combien d'occurrences (soit sous forme d'URL, soit sous forme de catégories) ont été retrouvées, ce qui permet dans une certaine mesure de pondérer la validité du terme.

16 base documentaire scientifique développée à et maintenue par l'INIST-CNRS.

4.1.2. Résultats

Les recherches menées sur les corpus de *Le Monde* et de l'agroalimentaire ont donné des résultats quasiment nuls (0,9% de succès, sur les 2691 candidats), ce qui confirme que (1) les mots à tester présentent un caractère trop spécialisé pour un corpus journalistique et que (2) le sens de ces mots couvre un ensemble trop large de spécialités pour que leur présence soit remarquable dans un corpus spécialisé dans un seul domaine [17].

En ce qui concerne les résultats des recherches menés sur l'*EU*, la *BDM* et sur w3, ils sont consignés dans le tableau ci-dessous :

Listes	**Quantité (total = 2691)**	***Expérience 1*** **Présents dans *EU* et *BDM* (Nb)**	***Expérience 2*** **Présents sur w3**			**Présents globalement (w3+EU+BDM) Nb / (pourcent)**
			Succès : Nb / (pourcent)	**Nb : moins de 10 occurrences**	**Nb : plus de 10 occurrences**	
Xisable	755	39	101 / (13,4%)	94	7	**112 / (14,8%)**
Xisabilité	833	2	18 / (2,1%)	15	3	**18 / (2,1%)**
Xabilité	1103	56	232 / (21%)	197	35	**246 / (22,3%)**

Tableau 1 : résultats quantitatifs

Les deux premières colonnes décrivent la répartition des termes générés en fonction de la combinaison de suffixes. La 3ème colonne indique combien de ces termes ont été retrouvés, par recherche manuelle dans l'*EU* ou la *BDM* (*Expérience 1* de validation). Les trois colonnes suivantes résument les résultats positifs obtenus sur w3 (*Expérience 2* de validation). Nous avons décomposé ces résultats (colonnes 5 et 6) selon le nombre d'occurrences annoncées par www.yahoo.fr. Enfin, la dernière colonne additionne les résultats des deux expériences (les résultats communs n'étant bien entendu comptabilisés qu'une fois). Ces résultats appellent certains commentaires :

1. Le faible pourcentage des termes en *Xisabilité* (2,1%) utilisés dans les corpus corrobore l'hypothèse qu'une construction dépassant la combinaison de deux suffixes peut poser des problèmes de performance, et que, bien que les dérivés de structure *Xisabilité* soient possibles et parfaitement interprétables, l'usage va plutôt adopter des stratégies palliatives pour éviter la formation de tels termes (cf. § 3.1.2.3.).
   Une autre illustration du phénomène se retrouve lors de la consultation d'*EURODICAUTOM* [18] : par exemple, la traduction attestée du nom anglais *graphitizability* n'y est pas *graphitisabilité*, mais "*aptitude à la*

17 On notera quand même que certains termes absents de corpus couvrant des genres textuels larges sont présents dans des corpus journalistiques : c'est le cas de *insoupçonnabilité* et *adoptabilité*, présents dans *LM99* et absents de w3, de la *BDM* et de l'*EU* (du moins dans la version de 1996 consultée).

18 *Eurodicautom* est le dictionnaire automatique multilingue de la Commission des Communautés européennes à Luxembourg. Il couvre tous les aspects des activités de la CE dans 11 langues, et est consultable de manière interactive (cf. URL http://eurodic.echo.lu/cgi-bin/edicbin/EuroDicWWW.pl)

graphitisation".

2. Pour les deux autres types de constructions, on obtient en revanche de bons résultats. La distinction, arbitraire, dans l'*Expérience 2* entre "plus de 10 occurrences" et "moins de 10 occurrences" est en outre une indication possible de la multiplicité des domaines dans lesquels le terme est employé (une recherche plus poussée permettrait sans doute de mettre en évidence quels domaines sont les plus créateurs de termes ; cependant, une telle expérience est hors de notre propos, notre objectif étant de montrer simplement l'existence dans l'usage de ces mots hors dictionnaires).
3. Enfin, la comparaison des résultats obtenus lors des *Expériences 1* et *2* confirme que l'absence de résultat sur w3 ne prouve pas que le terme n'est pas usité : en témoignent par exemple *théâtralisable* ou *interdéfinissabilité,* recueillis dans l'*EU*, ou encore *égouttabilité* ou *pluralisable*, recueillis dans la *BDM*. Nous pouvons donc affirmer que les pourcentages obtenus indiquent des quantités **minimales** de termes existant parmi les unités générées : en fait, parmi les termes générés automatiquement, et quelle que soit la combinaison licite de suffixes choisie (rappelons que *-is(er)*, *-able* et *-ité* ne constituent qu'une illustration de notre système) nous pouvons nous attendre **au moins** à 15 - 20 % de mots déjà utilisés dans un ou plusieurs domaines de spécialité.

Reste maintenant à déterminer si ces termes, dont la génération automatique est justifiée par leur utilisation démontrée ou probable, reçoivent de la part de GéDériF une analyse constructionnelle et sémantique appropriée, condition *sine qua non* à leur exploitabilité dans la base par l'utilisateur en RI. C'est justement ce qui est présenté dans le paragraphe ci-dessous.

4.2. Evaluation qualitative

Notre deuxième série de tests a en effet porté sur l'évaluation de la qualité du lexique généré, d'un triple point de vue formel, structurel et sémantique.

Avant de générer notre lexique, nous nous étions munies d'un certain nombre de garde-fous linguistiques (cf. § 3.1.2.). En outre, le lexique produit ne rencontre pas de problèmes catégoriels (n'ont été générées que des unités lexicales catégoriellement licites), et *a priori* peu de problèmes formels d'allomorphie. Les risques d'erreur sont donc minimisés d'entrée, et se cantonnent dans le domaine du sens.

De fait, les résultats obtenus automatiquement sont globalement bons, mais néanmoins perfectibles pour un certain nombre d'entre eux : le lexique généré et analysé automatiquement hérite en effet des imperfections des analyses déjà implémentées sur les unités lexicales attestées, auxquelles viennent se greffer le cas échéant certains problèmes d'analyse qui lui sont propres. Les points perfectibles que nous avons repérés sont les suivants (la vérification manuelle systématique a été l'occasion de corriger certaines erreurs ponctuelles, qu'il n'est pas opportun de signaler puisqu'elles n'existent plus) :

1. DériF ne traite actuellement pas les cas d'ambiguïté structurelle qui, parmi les suffixes étudiés à ce jour, touche principalement des dérivés

comportant *-able* dans leur structure. Sont notamment structurellement ambigus les adjectifs de forme *inXable*, quand *inX* et *Xable* sont respectivement un verbe et un adjectif attestés ou possibles : par exemple, *inversable* peut être dérivé d'*invers(er)* (il dit alors du référent de son nom recteur qu'il peut être inversé) ou de *versable* (il dit alors du référent de son nom recteur qu'il ne peut pas être versé) [19]. Le choix fait dans DériF, donc dans GéDériF, est actuellement de voir dans tout adjectif de la forme *inY* un antonyme de *Y* (parce que c'est le cas le plus fréquent) : ainsi, *inversable* par exemple ne reçoit qu'une structure ([in [versable ADJ] ADJ]) et qu'une glose (" non versable "). L'option actuellement faite se répercute naturellement sur les noms de propriété correspondants : quand il applique *-ité* à un adjectif en *inXable* pour former un nom hors dictionnaires, GéDériF ne propose qu'une chaîne dérivationnelle ($X \rightarrow Xable_A \rightarrow inXable_A \rightarrow inXabilité_N$) alors qu'une seconde chaîne est parfois possible (($X \rightarrow$)$inX_V \rightarrow inXable_A \rightarrow inXabilité_N$). Cette imperfection peut être facilement gommée ; elle n'engendre en outre que du silence.

2. L'autre type de dérivés dont l'analyse peut être parfaite est constitué des unités lexicales de structure *Xisable* et, de façon liée, de structure *Xisabilité*, avec *X* = nom propre (par exemple, *pantagruélisable* / *pantagruélisabilité*). Le problème que posent ces séquences a été entrevu au § 3.1.1. : *-is(er)* peut certes opérer sur des noms propres mais, selon le contenu du nom propre, le verbe en *–is(er)* est tantôt intransitif, tantôt transitif. Or, ce n'est que dans ce dernier cas qu'il peut se voir appliquer le suffixe *–able*). Comme les entrées de DériF ne comportent pas d'informations sémantico-référentielles et, donc, ne développent pas le contenu des noms propres (cf. § 3.2.1.), les résultats de GéDériF pâtissent de cette absence d'informations : dans l'état actuel du système, nous ne pouvons pas faire automatiquement le départ entre des unités comme *brechtisable* / *brechtisabilité*, possibles parce que le contenu du nom propre *Brecht* s'y prête (*brechtis(er)* exprime le procès de donner à une œuvre les caractéristiques de l'œuvre de Brecht : il est donc transitif), et *pantagruélisable* / *pantagruélisabilité*, peu plausibles parce que *pantagruélis(er)* est un verbe intransitif exprimant le procès de se comporter à la manière de Pantagruel. Même si cette décision engendre parfois du bruit, nous choisissons de continuer de produire automatiquement les adjectifs en *–able* et noms en *–ité* à partir de verbes en *–is(er)* dérivés de noms propres, pour deux raisons : d'une part, la base de ces verbes a plus souvent pour référent un individu connu pour son œuvre qu'un individu connu pour son comportement singulier (pour une banale raison pragmatique : les gens connus le sont plus souvent pour leur œuvre que pour leur comportement singulier) ; d'autre part, parce que le calcul du sens d'un adjectif comme *pantagruélisable* ne pose en fait pas un problème linguistique : on lui associe régulièrement la glose " que l'on peut pantagruéliser ". En outre, certains *Xis(er)* où *X* est un anthroponyme présentent les deux acceptions dans le lexique attesté : par exemple, *diogénis(er)*, qui, selon le *RE*, peut exprimer le procès de se

19 Voici d'autres dérivés en *–able* (entérinés par les dictionnaires ou possibles) qui présentent la même caractéristique : *importable, imprécisable, inactivable, incitable, indisposable, infiltrable*, *infléchissable*, *informable, ingérable, inhumable, intaillable, insonorisable, invalidable.*

comporter à la manière de Diogène (il est alors intransitif), mais aussi celui de rendre quelqu'un comparable à Diogène (il est alors transitif).

5. CONCLUSION

La présentation qui a été faite sur les combinaisons constructionnellement licites des suffixes *-able*, *-ité* et *–is(er)* est aisément reproductible pour d'autres combinaisons productives dans divers domaines du vocabulaire technique, savant, etc. et quasiment absentes des dictionnaires. La démarche suivie est par exemple reconductible :

- pour la combinaison *-al* (*–el*)+ *-is(er)* : alors que le *RE*, le *TLF* et le *NPR* attestent à eux trois environ 80 verbes de forme *Xalis(er)* où *-al* correspond au suffixe *–al* (*théâtre* → *théâtral* → *théâtralis(er)*) ou à l'allomorphe du suffixe *–el* (*profession* → *professionnel* → *professionnalis(er)*), une recherche rapide dans le seul *LM99* enrichit le corpus de près de 10% : *compartimentaliser* (1 occurrence) ; *contextualiser* (12 occ.) ; *contractualiser* (21 occ.) ; *convivialiser* (1 occ.) ; *originaliser* (1 occ.) ; *procéduraliser* (1 occ.) ; *proportionnaliser* (1 occ.) ;
- pour la combinaison *–is(er)* + *-ation* (on opposera ainsi les 496 *Xisation* que livre le *RE* aux 744 *Xisation* présents dans le seul *LM99*), éventuellement combinée à la précédente (149 *Xalisation* dans le *RE* contre 200 dans le seul *LM99*) [20] ;
- pour la combinaison *–ation* + *-el* (25 *Xationnel* dans le *RE* contre 40 dans le seul *LM99*), éventuellement combinée à la précédente (1 *Xisationnel* contre 2 dans *LM99*) ;
- etc.

À terme, le système GéDériF est donc capable de générer, d'analyser et de gloser un ensemble potentiellement infini d'unités lexicales construites contrôlées linguistiquement, constituant ainsi des ressources dont peuvent tirer profit la recherche d'informations et la terminologie.

RÉFÉRENCES

AUTOMORPHOLOGY. URL : http://humanities.uchicago.edu/faculty/goldsmith/

BDM = La banque des mots, revue de terminologie française publiée par le conseil international de la langue française, Paris, Conseil international de la langue française.

BOUILLON, Pierrette (1998) : *Traitement automatique des langues naturelles*, Paris-Bruxelles, Duculot.

BRILL, Eric (1994) : "Some Advances in Transformation-Based Part of Speech Tagging*", in Proceedings of the AAAI Conference, Volume1,* Seattle, 722-727.

CORBIN, Danielle (1993) : "Morphologie et lexicographie : la représentation du sens dans le *Dictionnaire dérivationnel du français*", *in* Hulk A., Melka F. & Schroten J. éds, 63-86.

CORBIN, Danielle (à paraître) : *Le lexique construit*, Paris, Armand Colin.

20 On ne donne ici et dans les configurations suivantes que des chiffres bruts issus de la consultation du *RE* et de *LM99*, l'étude des opérateurs cités n'ayant pas encore menée à ce jour.

DAL, Georgette (1997a) : “Du principe d’unicité catégorielle au principe d’unicité sémantique : incidence sur la formalisation du lexique construit morphologiquement”, *in* P.-A. Buvet, S. Cardey, P. Greenfield & H. Madec éds, *Actes du colloque international Fractal’97,* “ Linguistique et informatique : théories et Outils pour le traitement automatique des langues ”, *BULAG* numéro spécial, 105-115.

DAL, Georgette (1997b) : *Grammaire du suffixe* –et(te), Paris, Didier Erudition.

DAL, Georgette, HATHOUT, Nabil & NAMER, Fiammetta (1999) : “Construire un lexique dérivationnel : théorie et réalisations”, *in Actes de la VIe conférence sur le Traitement Automatique des Langues Naturelles (TALN'99), Institut d'Etudes Scientifiques de Cargèse*, Corse, 12 -17 juillet 1999, 115-124.

DAL, Georgette, HATHOUT, Nabil & NAMER, Fiammetta (soumission) : “Morphologie constructionnelle et traitement automatique des langues : le projet MorTAL”, *Lexique* 16.

EU = CD-ROM *Encyclopaedia Universalis*, version 2.0, Paris, Encyclopaedia Universalis, 1995.

FRADIN, Bernard (1994) : “L’approche à deux niveaux en morphologie computationnelle et les développements récents en morphologie”, *T.A.L.* 35/2, 9-48.

FROISSART, Christel & LALLICH-BOIDIN Geneviève (1996) : “Morphologie robuste et analyse automatique de la langue : étude réalisée à partir des corpus de l’évaluation GRACE”, *in Actes du séminaire Lexique. Représentations et Outils pour les bases lexicales. Morphologie robuste*, Grenoble, CLIPS-IMAG, les 13 et 14 nov. 1996, 88-96.

FUCHS, Catherine éd. (1993) : *Linguistique et traitements automatiques des langues*, Paris, Hachette supérieur.

GARY-PRIEUR Marie-Noëlle (1994) : *Grammaire du nom propre*, Paris, Presses Universitaires de France.

GAUSSIER Eric (1999) : “Unsupervised learning of derivational morphology from inflectional lexicons“, *Workshop on Unsupervised methods for natural language processing*, *ACL* 1999.

GRABAR, Natalia & ZWEIGENBAUM, Pierre (1999) : “Acquisition automatique de connaissances morphologiques sur le vocabulaire médical”, *in Actes de la VIe conférence sur le Traitement Automatique des Langues Naturelles (TALN’99), Institut d’Etudes Scientifiques de Cargèse, Corse, 12 -17 juillet 1999*, 175-184.

GRUAZ, Claude, JACQUEMIN, Christian & TZOUKERMAN, Evelyne (1996) : “Une approche à deux niveaux de la morphologie dérivationnelle du français”, in *Actes du séminaire Lexique. Représentations et Outils pour les bases lexicales. Morphologie robuste*, Grenoble, CLIPS-IMAG, les 13 et 14 nov. 1996, 107-114.

HATHOUT, Nabil, NAMER, Fiammetta & DAL, Georgette (à paraître) : “Une base de données constructionnelles expérimentale : le projet MorTAL“, *in* P. Boucher ed., *Morphology book*, Cambridge Mass, Cascadilla Press.

KRAAIJ, Wessel & POHLMANN, Renée (1996) : “Viewing stemming as recall enhancement”, in *Proceedings of ACM-SIGIR 96, Conference on Research and Development in Information Retrieval*, 40-48.

LECOMTE, Josette & PAROUBEK, Patrick (1996) : “Le catégorisateur d'Eric Brill. Mise en œuvre de la version entraînée à l'INaLF”, rapport

technique, Nancy, INaLF-CNRS.

LENNON, M. ; PIERCE, D. ; TARRY, B. & WILLETT, P. (1981) : “An evaluation of some conflation algorithms for information retrieval”, in *Journal of Information Science*, n° 3, 177-183.

LM93 = CD-ROM du journal *Le Monde* pour 1992, Le Monde / Research Publications International, 1993.

LM99 = *Le Monde sur CD-ROM*, SA Le Monde (Paris) – CEDROM-SNi inc. (Montréal), 1999.

MARTIN, Robert (1992) : *Pour une logique du sens*, 2e éd. revue et augmentée, Paris, Presses Universitaires de France ; 1e éd., 1983.

MAUREL, Denis ; BELLEIL, Claude ; EGGERT, Elmar ; PITON, Odile (1996) : “Réalisation d’un dictionnaire électronique relationnel des noms propres du français”, in *Actes du séminaire Lexique. Représentations et Outils pour les bases lexicales. Morphologie robuste*, Grenoble, CLIPS-IMAG, les 13 et 14 nov. 1996, 164-175.

NAMER, Fiammetta (1999) : “Le traitement automatique des mots dérivés : le cas des noms et adjectifs en *–et(te)*”, *in* D. Corbin, G. Dal, B. Fradin, B. Habert., F. Kerleroux, M. Plénat & M. Roché éds, *La morphologie des dérivés évaluatifs*, *Silexicales* 2, Université de Lille III, 169-179.

NPR = Le Petit Robert. Dictionnaire de la langue française. Version électronique du *Nouveau Petit Robert*. Disque optique compact CD-ROM, Paris, Dictionnaires Le Robert / van Dijk, 1996.

PLAG, Ingo (1998) : “The polysemy of *–ize* derivatives : on the role of semantics in word formation”, *in* G. Booij & J. Van Marle eds, *Yearbook of Morphology 1997*, 219-242.

PORTER, Martin (1980) : “An algorithm for suffix stripping”, in *Program*, n°14, 130-137.

RE = Le Robert électronique DMW, Disque optique compact CD-ROM, Paris, Dictionnaires Le Robert, 1994.

SAVOY, Jacques (1993) : “Stemming of French Words Based on Grammatical Categories”, *JASIS: Journal of the American Society for Information Sciences,* vol. 44 : 1, 1-9.

SPROAT, Richard William (1992) : *Morphology and Computation*, Cambridge, Massachusetts / London, England, The MIT Press.

TLF = Trésor de la langue française. Dictionnaire de la langue du 19e et du 20e siècle (1789-1960), 16 vol., Paris, Éditions du CNRS (t. 1-10) / Gallimard (depuis le t. 11), 1971-1994.