Conférence TALN 1999, Cargèse, 12-17 juillet 1999

Construire un lexique dérivationnel : théorie et réalisations*

Georgette DAL — « SILEX », CNRS & Un. Lille3 — dal@univ-lille3.fr

Nabil HATHOUT — CNRS, INaLF Nancy & ERSS — hathout@univ-tlse2.fr

Fiammetta NAMER — LANDISCO & Université Nancy2 — namer@clsh.univ-nancy2.fr

Résumé

Le travail qui suit teste différentes façons de concevoir et de construire un lexique dérivationnel. Afin de mener à bien cette tâche, nous centrerons l'analyse sur les suffixations par *-able* et *-ité* du français (et les dérivés qu'elles forment), et nous les soumettrons à des éclairages différents : un éclairage proprement théorique et deux éclairages plus finalisés, DériF et DéCor, qui présentent des techniques différentes pour le traitement automatique de la morphologie. Au terme de ce travail, nous comparerons les résultats obtenus.

1. Introduction

Le présent travail s'inscrit dans le cadre du projet *FRANLEX* initié par G. Dal (UMR 5828 « SILEX », CNRS & Un. Lille 3) et Ch. Jacquemin (UPR 3251 « LIMSI », CNRS). *FRANLEX* résulte d'une réflexion scientifique collective à laquelle ont participé à ce jour une quinzaine de chercheurs [1]. Son objectif ? Construire (semi-) automatiquement une base de données morphologiques (BDM) associant à chaque unité lexicale construite du français une description structurelle et, à terme, sémantique, et combler ainsi une lacune dont pâtirait sinon le TALN. À notre connaissance, il n'existe en effet pour le français aucune base de ce type (d'autres langues en possèdent, par exemple CELEX pour le néerlandais, l'anglais et l'allemand). Or, un tel outil a des applications potentielles multiples en TALN : recherche documentaire, compréhension de textes, lexicographie mono- ou multilingue, recherche d'information (RI), etc. Dans ce travail, nous examinerons plus précisément les mots construits par les suffixations par *-able* et *-ité* du français, en les soumettant à des éclairages différents : un éclairage proprement théorique (§ 2), deux éclairages plus finalisés (§ 3), DériF et DéCor, qui mettent tous deux en œuvre des techniques pour le traitement automatique de la morphologie, mais selon des modalités différentes (DériF implémente des hypothèses linguistiques, DéCor adopte un point de vue statistique) [2]. La section 4 est dédiée aux évaluations : évaluation des résultats obtenus par DériF par rapport à la théorie, évaluation croisée des résultats obtenus par DériF et par DéCor, évaluation des avantages et inconvénients de chaque programme. Au terme de ces évaluations se profilera la BDM idéale.

* Un grand merci à Christian Jacquemin et à Marc Plénat qui ont eu la gentillesse de relire et de commenter des versions liminaires de ce travail. Merci également aux relecteurs anonymes de cette soumission. Il va de soi que le contenu de cet article et ses imperfections reste de notre seule responsabilité.

[1] Ces chercheurs sont nommés dans la version détaillée de *FRANLEX* aux URL http://www.limsi.fr/Individu/jacquemi/FRANLEX/ et http://www.univ-lille3.fr/www/silex/dal.html. Sans prétendre à l'exhaustivité, on citera ici D. Corbin (UMR 8528 « SILEX », CNRS & Un. Lille 3), B. Fradin (UMR 7546 « LLI », CNRS & Un. Paris 13), B. Habert (UPR 3251 « LIMSI », CNRS), F. Kerleroux (EA 272 « MLDL », Un. Paris X-Nanterre), M. Plénat (UMR 5610 « ERSS », CNRS & Un. Toulouse 2 le Mirail).

[2] Nous ne présentons ici que le traitement de deux affixes. Nous projetons sous deux ans le traitement complet de 12 affixes, ce qui permettra de traiter quelque 10.000 unités lexicales construites présentes dans le corpus de référence.

2. Les suffixes *-able* et *-ité* : ébauches d'analyse linguistique

2.1. Cadre théorique

2.1.1. Présentation générale

Les ébauches d'analyses linguistiques qui suivent se situent dans une perspective théorique générale concevant la morphologie dérivationnelle comme le calcul conjoint de la structure et du sens des unités lexicales. Dans une telle perspective dite « associative », un mot est analysable comme construit s'il a une structure construite et un sens construit qui soit calculable à partir de celle-là.

2.1.2. Définir un opérateur de construction d'unités lexicales

Quand il s'agit de définir un opérateur constructionnel, on pense généralement en premier lieu aux catégories lexicales qu'il met en relation. De fait, les opérateurs sont contraints catégoriellement, qu'il s'agisse de contraintes pesant sur l'appartenance catégorielle de leurs inputs ou sur celle de leurs outputs. S'il est nécessaire, ce seul appariement catégoriel n'est toutefois pas suffisant. S'il l'était, on n'expliquerait pas pourquoi *re-*, qui peut opérer sur des verbes (*part(ir)*$_V$ / *repart(ir)*$_V$), peut difficilement opérer sur tous (*vieill(ir)* / ?*revieill(ir)*), pourquoi *-erie*, qui peut opérer sur des adjectifs (*bizarre*$_A$ / *bizarrerie*$_N$), est rétif voire franchement hostile à d'autres (*atomique*$_A$ / **atomi(qu+c)erie*$_N$).

Un opérateur constructionnel n'est donc pas définissable par les seules catégories lexicales qu'il met en relation. Deux autres types de contraintes, au moins, lui sont associés : des contraintes phonologiques, qui peuvent être générales comme la règle d'haplologie [3] ou spécifiques à l'opérateur en question, et des contraintes sémantiques : pour qu'un opérateur constructionnel puisse opérer sur une base, il faut qu'il y ait compatibilité entre le sens de la base et l'instruction sémantique de l'opérateur [4].

Les contraintes sémantiques et les contraintes catégorielles dont il vient de s'agir ne doivent pas être conçues indépendamment les unes des autres, les secondes étant subordonnées aux premières. Depuis peu en effet, tandis que l'on assiste à une réhabilitation progressive du sens en linguistique [5], on envisage que les contraintes catégorielles se font l'écho de l'instruction sémantique des opérateurs constructionnels : non que l'on considère désormais que la construction des unités lexicales est catégoriellement non contrainte, mais on pense maintenant que certains opérateurs ont une instruction sémantique se satisfaisant de bases polycatégorielles, tandis que l'instruction sémantique d'autres exige des bases monocatégorielles [6]. Une analyse linguistique, même minimale, des suffixes *-able* et *-ité* ne peut par conséquent pas se restreindre à étiqueter les catégories qu'ils associent : on dira ici également quelques mots de leur caractérisation sémantique (faute de place, et de réelles compétences dans le domaine, on laissera de côté les contraintes phonologiques).

---

[3] La règle d'haplologie interdit par exemple **candidatat* ou **lavetet(er)*, leur préférant *candidature* et *lavet(er)*. Sur cette contrainte, *cf.* Corbin & Plénat (1992).

[4] C'est par exemple une incompatibilité entre le sens de *atomique* et l'instruction de *-erie* qui explique l'agrammaticalité de **atomi(qu+c)erie* (*cf.* Dal (1997a)).

[5] Cette réhabilitation est nette en syntaxe : *cf.* Le Pesant & Mathieu-Colas éds (1998), p. 3.

[6] Du moins cette tendance se dessine-t-elle désormais dans le modèle théorique élaboré à SILEX.

2.2. Le suffixe -able

2.2.1. Caractérisation catégorielle

Le suffixe *-able* forme des adjectifs, à l'exclusion de tout autre type catégoriel de dérivés. Le lexique attesté donne certes à observer des noms en *-able* où *-able* est un suffixe (par ex., *dirigeable*, *imperméable*), mais ces derniers sont analysables comme les produits de l'application d'une conversion à des adjectifs comportant déjà *-able* dans leur structure [7]. S'il ne construit qu'un seul type catégoriel de dérivés, *-able* peut en revanche opérer sur deux types catégoriels de bases : des verbes (par ex., *abaiss-*$_V$ / *abaissable*$_A$; *accentu-*$_V$ / *accentuable*$_A$), mais aussi, même si c'est plus rare, des noms (par ex., *cardinal*$_N$ / *cardinalable*$_A$; *charité*$_N$ / *charitable*$_A$).

2.2.2. Caractérisation sémantique

Étant donnée la double catégorisation des bases que peut sélectionner *-able*, on doit en premier lieu se demander s'il s'agit là de deux instanciations d'un même suffixe ou s'il convient de démultiplier cette forme affixale en (au moins) deux suffixes homomorphes : on prendra ici la première option, pour des raisons sémantiques, d'autant plus que cette curiosité apparente se retrouve en anglais : *accept*$_V$ / *acceptable*$_A$; *fashion*$_N$ / *fashionable*$_A$ [8]. En effet, quelle que soit l'étiquette catégorielle de leur base, les adjectifs en *-able* expriment une propriété indépendante de sa mise en œuvre effective (= ils expriment une « possibilité »). Ils signifient plus précisément que le référent de leur nom recteur a ceci de particulier qu'il est susceptible :

– quand la base est verbale, de se voir appliquer le procès qu'elle exprime : *appelable*, *mangeable* (le nom recteur appartient à la classe sémantique des arguments internes du verbe de base),

– quand la base est nominale [9] : (i) de manifester ou de susciter la propriété qu'elle désigne (le type sémantique du nom recteur dépend de la propriété exprimée par le nom de base) : *charitable, confortable, épouvantable, effroyable*, *pitoyable,* (ii) de devenir le personnage social qu'elle désigne (le référent du nom recteur est alors une personne) : *cardinalable*, *ministrable,* etc. ; °*députable*, °*épiscopable*, etc., (iii) de se voir assujettir à l'impôt qu'elle désigne (le référent du nom recteur est de nouveau une personne) : *corvéable*, *mainmortable*, etc. ; °*CSGable*, °*IGFable*, etc., (iv) de permettre le déplacement au moyen de ce qu'elle désigne (le nom recteur désigne alors un type de voie) : *carrossable*, *cyclable* ; °*rollerable*, °*VTTable*.

2.3. Le suffixe -ité

Comme *-able*, *-ité* forme un seul type catégoriel de dérivés – des noms – et accepte deux

---

[7] Ces noms convertis désignent diverses catégories d'entités dont le trait saillant est exprimé par l'adjectif de base : un dirigeable est un objet qui possède la propriété saillante 'dirigeable', un imperméable un objet qui possède la propriété saillante 'imperméable'. Sur cette conversion, *cf.* Corbin & Corbin (1991).

[8] C'est du reste pour cette raison que M. Aronoff, tributaire de l'hypothèse de la base unique (« Unitary Base Hypothesis ») qu'il met au jour dans son ouvrage de 1976, distingue deux suffixes *–able* anglais homomorphes dans ce même ouvrage.

[9] La quadripartition qui suit se sert de Plénat (1988). Dans les exemples ci-après, la rondelle en exposant devant certains dérivés indique qu'ils sont possibles mais, pour autant qu'on le sache, non attestés (la possibilité de former de nouveaux dérivés selon le patron X_N + *-able* semble restreinte aux types (ii)-(iv) ci-après).

types catégoriels de bases : des adjectifs (*absolu*$_A$ / *absoluité*$_N$; *absorptif*$_A$ / *absortivité*$_N$) et des noms, même si c'est plus rare (*bouddha*$_N$ / *bouddhéité*$_N$; *sœur*$_N$ / *sororité*$_N$). De nouveau, il convient donc de se demander si *-ité* est un de ces suffixes qui se satisfont d'une polycatégorialité des bases qu'ils sélectionnent, ou s'il subsume (au moins) deux suffixes homomorphes. L'option que l'on prendra sera analogue à celle qui a été prise pour *-able* : on dira que la grammaire du français comprend un seul suffixe *-ité*, en arguant de raisons sémantiques.

On tient en effet *-ité* pour former des noms de propriété, à l'instar de *-erie*, *-esse*, etc. Comme chacun de ces opérateurs, *-ité* configure à sa façon les noms qu'il construit : sa spécificité est de présenter les propriétés qu'expriment ces noms comme étant objectives, ou pour le moins, objectivables (*cf.* Corbin (à paraître)). Étant donné le type sémantique de noms que forme *-ité*, il opère de façon optimale sur des adjectifs : le noyau dur de la catégorie des adjectifs que sont les qualificatifs est en effet définissable par plusieurs critères, parmi lesquels l'expression d'une propriété. Pourtant, rien n'interdit à ces suffixes de sélectionner d'autres types catégoriels de bases : il suffit que ces dernières puissent exprimer des propriétés. Ils peuvent donc opérer sur noms, sous réserve qu'on ne les considère plus dans leur extension, comme référant à des substances, mais dans leur intention, comme référant aux propriétés de ces substances : c'est le cas dans *bouddhéité* dans lequel *bouddha* ne réfère pas à une entité mais à une manière d'être, une philosophie ou dans *matissité* [10] où *Matisse* ne réfère plus au peintre que l'on sait, mais à une technique réductible à quelques traits objectifs.

Si l'on se fonde sur ce qui précède, l'analyse linguistique d'un opérateur constructionnel n'a rien de trivial, puisqu'elle met en scène, outre des contraintes catégorielles, aussi (surtout ?) des contraintes sémantiques. On peut donc supposer que, même grossière, une analyse linguistique peut guider et enrichir les outils pour le traitement automatique de la morphologie dérivationnelle. C'est ce qu'on se propose d'évaluer dans les sections suivantes.

3. Traitement automatique de la morphologie dérivationnelle (TAMD)

Deux techniques pour le TAMD, DériF et DéCor, sont présentées dans cette section. Elles diffèrent à la fois dans leur complexité et dans leur objectif. DériF est un analyseur dérivationnel associant à chaque mot son arbre d'analyse, *i.e.* les découpages successifs base/affixe jusqu'à l'obtention d'une unité indivisible. DéCor, quant à lui, permet de calculer des bases possibles d'un point de vue statistique pour les mots construits. Avant de présenter chacune de ces approches ainsi que leurs résultats, nous donnons en § 3.1 un bref aperçu de quelques analyseurs morphologiques existants. (Les deux programmes ont été appliqués aux mêmes corpus, extraits de *TLFnome* [11]).

3.1. Quelques analyseurs morphologiques

Parmi les analyseurs récents, on distingue principalement les approches à base de corpus (comme DéCor) et celles à base de connaissances (comme DériF). Nous présentons ci-dessous les aspects innovants de ces deux analyseurs par rapports à d'autres travaux.

Dans le cadre de la linguistique de corpus (Habert *& al*. 1997), on peut citer pour l'anglais

---

[10] Pour une analyse moins rudimentaire de *matissité*, *cf.* Dal (1997b).

[11] *TLFnome* est un lexique de formes fléchies construit à l'INaLF par M. Papin et J. Maucourt, à partir de la nomenclature du *Trésor de la Langue Française* (*TLF*). Ce lexique 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*. Cette extension constituera un corpus d'évaluation pour ces systèmes.

les travaux de Ch. Jacquemin (1997) et de J. Xu & W. B. Croft (1998), tous deux destinés à l'extension de requêtes en RI à l'aide de variantes morphologiques de termes complexes (dans la première étude) et de mots cooccurrents dans un corpus (dans la seconde). Dans les deux cas, aucune information linguistique n'est mise en œuvre. Jacquemin (1997) utilise une liste de termes et un corpus pour calculer des variantes morphologiques pour les termes complexes (*i.e.* comportant plus de deux mots) de la liste. Ce travail est plus élaboré que DéCor car il met en œuvre des techniques de classification pour regrouper les variations allomorphiques, les variantes graphiques et les fautes d'orthographe. Pour le français, citons Zweigenbaum & Grabar (1999) dont le but est de construire une BDM à partir du microglossaire médical SNODEM et de la Classification Internationale des Maladies CIM-10. Les principales différences entre ce travail et DéCor sont que le premier utilise une ressource structurée, comportant des indications sémantiques (relation de synonymie) qui garantissent la correction de ses solutions et qu'il ne distingue pas la flexion de la composition ou de la dérivation : il destine la BDM qu'il construit à l'extension de requêtes par mots-clés en RI, alors que DéCor, développé dans le cadre de *FRANLEX*, vise à construire une BDM générale pour le TALN.

En ce qui concerne la linguistique à base de connaissances, elle se subdivise en deux approches : (1) les formalismes basés sur **dictionnaire**, comme celui de J. Savoy (1993) dont l'objectif est de proposer une racine (et non pas une analyse dérivationnelle) et dont l'inconvénient principal est son inaptitude à traiter les néologismes ; (2) les formalismes basés sur règles, qui se réclament en général de la **morphologie à deux niveaux** (Koskenniemi 1983). Cette approche a été utilisée par le système de D. Clémenceau (1993) qui nécessite **une base lexicale complexe**, puisqu'il exploite les traits qui codent les entrées lexicales du Lexique-Grammaire (Gross 1975) pour définir, sous forme d'automates, les dérivations possibles entre les classes de mots. Citons encore les travaux de C. Gruaz & *al.* (1996), ainsi que le modèle proposé par V. Clavier (1996a, 1996b). Ce dernier s'inscrit dans le cadre formel **des grammaires régulières**, ne traite pour l'instant que les **suffixes**, et se sert, comme D. Clémenceau, d'un **lexique fortement structuré** comme référentiel, ce qui explique pourquoi son champ d'**expérimentation** est limité.

Contrairement aux travaux présentés ci-dessus, à l'exception du modèle de V. Clavier, DériF est un système basé sur règles. Il met en œuvre la théorie dérivationnelle développée à SILEX (pour un premier état du modèle, *cf.* Corbin (1987)) et n'implémente pas de modèle formel particulier ; contrairement au système de V. Clavier, DériF est déterministe [12] et produit des analyses de mots construits par suffixation et/ou préfixation à l'aide de règles et de listes d'exceptions. Le référentiel requis est simplement un ensemble, de taille quelconque, de mots étiquetés, ce qui garantit des expérimentations variées et à grande échelle.

3.2. DériF : un analyseur morphologique DERIvationnel du Français

Le programme DériF effectue l'analyse dérivationnelle d'un mot étiqueté au moyen (i) d'un ensemble de règles élaborées à partir des hypothèses linguistiques résumées en §2, (ii) d'une liste d'exceptions (non productives) et (iii) d'un référentiel. Les règles produisent plusieurs types de résultats : elles découpent le mot selon chacun de ses suffixes et/ou préfixes ; elles représentent dans un format crocheté la portée de chaque affixe analysé sur leur base respective ; elles rassemblent, au fur et à mesure que le mot est analysé, la famille de celui-ci

[12] L'input de DériF, comme celui de DéCor, est un corpus de lemmes étiquetés ; ceci ne constitue pas une limitation dans la mesure où des catégorisateurs du français sont aujourd'hui aisément disponibles (par ex. WinBrill*v*0.3 (1998) distribué par l'INaLF, qui effectue étiquetage et lemmatisation).

jusqu'à parvenir à une unité lexicale indivisible. À moyen terme, les mots analysés se verront également associer une description sémantique reflétant la dernière opération qui les a construits [13]. À terme, DériF permettra donc de dresser presque [14] automatiquement la fiche signalétique de tous les mots construits du français, concrétisant ainsi l'objectif que s'assigne le projet *FRANLEX*. Dans sa version actuelle, DériF analyse les suffixes *-able*, *-ité* et *-et(te)*, ainsi que quelques allomorphes du préfixe *in-* formant des dérivés niant la propriété qu'expriment leur base (*apte / inapte*, *confort / inconfort*).

La fonction principale examine la terminaison du mot à analyser M, pour déterminer si celle-ci est un suffixe S connu du système. Si c'est le cas, la fonction spécifique d'analyse de S est appelée et commence par examiner récursivement le(s) préfixe(s) qui porte(nt) sur la base suffixée de M, avant d'analyser S (règle d'appariement formel entre M et sa base B, calcul de la catégorie de B et instruction sémantique de S). Le résultat B est renvoyé à la fonction principale, qui teste sa terminaison à la recherche d'un nouveau suffixe.

À titre d'exemple, nous détaillons l'analyse de M=*inexplicable*. La règle de découpage de *-able* est activée. Elle détecte que la séquence initiale est un préfixe portant sur sa base suffixée : *inexplicable* ⇒ *in* + *explicable*. Etant donnée la double catégorisation *a priori* des bases potentielles des adjectifs en *-able* (noms ou verbes, *cf.* §2.2.1), la règle assigne ensuite systématiquement à chaque M'=B-able deux analyses : `Y1=[[M'₁ VERBE] able ADJ]` et `Y2=[[M'₂ NOM] able ADJ]` :

```
(1) a. explicable =>   [[expliquer VBE] able ADJ]
    b. cyclable =>     [[cycle NOM] able ADJ]
    c. ministrable =>  [[ministre NOM] able ADJ]
(2)    perméable =>    [[permé FWD] able ADJ]
(3) a. valable =>      [[val/valoir NOM/VBE] able ADJ]
    b. minable =>      [[mine/miner NOM/VBE] able ADJ]
    c. confortable =>  [[confort/conforter NOM/VBE] able ADJ]
    d. muable =>       [[mue/muer NOM/VBE] able ADJ]
```

Les résultats, illustrés par les exemples (1) à (3) ci-dessus, sont obtenus par application de règles qui calculent l'allomorphe M'_1 ou M'_2 de B, et sont filtrés ensuite automatiquement au moyen de *TLFnome*. Les exemples (1) illustrent le cas où seul M'_1 ou M'_2 apparaît dans le référentiel. Si, à l'inverse, ni M'_1 ni M'_2 n'appartiennent à *TLFnome*, on fait l'hypothèse que la base est d'origine étrangère (ex. (2)). Si, enfin, M'_1 et M'_2 sont dans le lexique, le programme fait pour l'instant l'hypothèse que les deux résultats sont sémantiquement liés et les affiche sans chercher à les hiérarchiser (ex. (3)). La sous-spécification de cette analyse ($M'_1 \rightarrow M'_2$ ou $M'_2 \rightarrow M'_1$) sera réduite par une étude linguistique plus poussée des couples (M'_1, M'_2) concernés. Ces résultats sont discutés en § 4. Dans l'état actuel de DériF le résultat final de l'analyse de M a la forme suivante :

```
inexplicable => [in [[expliquer VERBE] able ADJ] ADJ]
                (inexplicable, explicable, expliquer)
```

3.3. DéCor : morphologie DÉrivationnelle sur CORpus

DéCor est un ensemble de programmes destinés à construire, de façon semi-automatique, un lexique dérivationnel du français. L'option retenue consiste à dissocier totalement l'analyse

---

[13] Par exemple (*cf.* Namer (1999)), la description complète de *ministrable*, incluant la contribution sémantique au sens global du mot de l'affixe le plus périphérique de M sera représentée par : `ministrable ==> [[ministre NOM] able ADJ] (ministrable,ministre) :: susceptible d'être ministre`

[14] Nous disons « presque », en relation avec les interventions humaines nécessaires à l'amélioration des règles.

des dérivés et la validation manuelle des résultats. Cette validation peut ainsi être réalisée, de façon autonome, par des personnes n'ayant pas de compétences particulières en informatique. Elle peut être simple, si l'on ne dispose pas de moyens importants, ou croisée afin de limiter les divergences entre les interprétations de chaque opérateur. DéCor permet de construire, à partir d'un lexique comme *TLFnome*, un graphe orienté et valué dont les nœuds sont des lemmes, et dont les arcs relient les dérivés à leur base ; chaque arc porte deux valeurs : les règles de préfixation et de suffixation qui ont permis de produire le dérivé à partir de sa base. De ce fait, seuls les mots attestés dans le lexique de référence sont pris en compte : *perturber* est ainsi proposé comme base pour *imperturbable* car °*perturbable* n'est pas attesté [15].

DéCor permet donc d'apparier les éléments d'un ensemble *D* de dérivés potentiels et ceux d'un ensemble *B* de bases potentielles. (Chaque suffixe est traité séparément.) Il utilise des règles de préfixation et de suffixation apprises par **findaffix**[16] à partir de ces corpus de mots. Ces règles sont appliquées à *D* et *B* pour associer à chaque $d \in D$ un sous-ensemble de *B* de bases candidates, $c(d)$, parmi lesquelles sera choisie une solution. Dans le modèle sous-jacent à notre démarche, chaque dérivé ne peut avoir qu'une seule base pour une dérivation donnée. Certains dérivés peuvent néanmoins avoir plusieurs bases comme *inactivable* que l'on peut dériver de *activable* et de *inactiver*, mais le plus souvent, un dérivé n'a qu'une base. En s'appuyant sur cette approximation, le programme **meilleure-base** de DéCor trie l'ensemble des bases candidates associées à un dérivé afin de déterminer celle qui est statistiquement la plus probable.

meilleure-base privilégie les candidats obtenus par préfixation seulement lorsqu'il y en a. Quatre stratégies basées sur des fonctions de tri ont été testées pour comparer ces candidats. Les fonctions de tri simple utilisent trois critères pour comparer les candidats : favoriser la règle de fréquence maximale et de taille minimale ; pénaliser les préfixations. Lorsque *D* peut contenir des dérivés à la fois préfixés et suffixés, la combinaison de deux fonctions de tri, l'une pénalisant les candidats à la fois préfixés et suffixés et l'autre ne les pénalisant pas, permet de tirer profit des points forts de chacune d'entre elles. L'efficacité de ces stratégies peut être améliorée en réitérant le tri sur des ensembles de candidats de plus en plus précis : avant l'itération *i*, on élimine de $[c(D)]_{i-1}$ les candidats obtenus par des règles qui n'ont produit aucun meilleur candidat à l'itération *i*-1 ; on remplace les fréquences calculées par **findaffix** par celles des règles de préfixation et de suffixation des meilleurs candidats de l'itération *i*-1 [17].

4. Résultats, adéquation aux hypothèses théoriques

Dans cette section nous faisons une évaluation quantitative et qualitative des résultats obtenus automatiquement par chaque analyseur. Les résultats de DériF sont confrontés aux analyses préconisées en §2 puis comparés à ceux de DéCor. Nous dressons enfin, pour chaque démarche, un bilan de ses avantages et inconvénients. DériF et DéCor ayant des finalités différentes, leur comparaison est nécessairement limitée : les seuls résultats que l'on ne puisse comparer sont les bases proposées. Les deux systèmes divergent entre autres sur la nature de leurs résultats : DériF fournit des bases qui sont en relation avec le dérivé du point de vue

---

[15] Cette analyse est sous-spécifiée pour la portée de la préfixation et de la suffixation :

[im[[perturber $_{VBE}$] able $_{ADJ}$] $_{ADJ}$], *[[im [perturber $_{VBE}$] $_{VBE}$] able $_{ADJ}$] ou *[im [perturber $_{VBE}$] able $_{ADJ}$] ?

[16] Script UNIX destiné à la compression des dictionnaires de formes en remplaçant certaines entrées par des règles d'affixation qui permettent de calculer les formes supprimées à partir de celles qui sont conservées.

[17] Les résultats atteignent un point fixe rapidement : quatre itérations au maximum pour les corpus *-able* et *-ité*.

morphologique et **sémantique** ; cet objectif est atteint en particulier grâce aux **listes d'exceptions**. DéCor ne peut proposer comme base, pour un dérivé, qu'un **lemme attesté** ; par exemple, il ne peut pas traiter correctement les emprunts.

4.1. Résultats de DériF par rapport aux hypothèses linguistiques

Sur l'ensemble du corpus traité (2043 mots non fléchis en *-able* et *-ité*), 1754 bases calculées sont exactes (soit 86%), 233 sont dans *TLFnome* mais posent des problèmes linguistiques (*cf.* § ci-dessous), et 56, absentes du *TLFnome*, sont repérées par le symbole « * ». Ce symbole signale donc *a priori* une erreur de calcul sur la base. Certaines bases sont effectivement erronées, par absence de traitement de certains affixes. Ainsi, *copaternité*↔*copaternel** (ADJ) car la portée du préfixe *co-* (qui dans le cas présent s'applique sur sa base suffixée par *-ité*) n'a pas encore été intégrée dans le programme. Mais d'autre part, le référentiel est (forcément) incomplet, et « * » signale alors des entrées manquantes. Enfin, l'analyse aboutissant parfois à une base possible non attestée (*imperturbable*↔*perturbable**) ou d'origine latine (*perméable*↔*permé**), celle-ci est forcément absente du lexique de référence sans que cela soit une erreur.

Les résultats qui figurent dans le référentiel peuvent poser divers types de problèmes linguistiques ; ainsi, rappelons-le, la catégorie de la base des mots en *-able* n'est pas unitaire (*cf.* § 2.2.1) ; la stratégie utilisée (*cf.* § 3.2) consiste à calculer les deux bases M'_1 et M'_2 possibles et à ne garder que celle qui est dans le référentiel. Si les deux s'y trouvent, cela ne signifie pas nécessairement qu'elles aient un lien morphologique : (1) l'exemple 3 a) montre un couple (M'_1, M'_2) = (*val*, *valoir*) dont les éléments sont formellement proches sans avoir de lien sémantique ; (2) contrairement à ce que propose DériF, même si M'_1 et M'_2 sont en relation, il se peut que seule l'une des analyses soit linguistiquement acceptable : soit base verbale (ex. 3 b) : *minable* ← $miner_V$, *$mine_N$), soit nominale (ex. 3 c) : *confortable* ← $confort_N$, *$conforter_V$) ; (3) enfin, l'exemple 3 d) pose le problème du sens de la conversion NOM ⇔ VERBE, qui dépend entre autres des propriétés sémantiques de M'_1 et M'_2 [18].

4.2. Résultats de DéCor, divergences avec DériF

Les corpus de travail des deux systèmes étant quasi-identiques, nous avons évalué DéCor par rapport à DériF. Sur les 836 entrées de l'intersection des deux corpus, DéCor et DériF divergent pour 127 dérivés (15%) qui se répartissent en 67 résultats corrects (sous-spécification de DéCor du type *imperturbable*↔*perturber* ; ambiguïté *coupable*↔*couper*/*coupable*), 26 non dérivés (ex. *friable*), 21 dérivés de noms et 13 erreurs. En considérant les non dérivés et les dérivés de noms comme des silences (ils ne sont pas reconnus en tant que tels), on obtient ainsi un rappel de 94% et une précision de 98%. Pour les dérivés en *-ité*, les résultats sont moins bons car cette dérivation est moins régulière (au sens où elle implique plus de variations allomorphiques) que la précédente. Les résultats des deux systèmes coïncident pour 857 des 1 030 entrées (83%). Les 173 divergences correspondent à 26 résultats corrects, 42 non dérivés, 15 dérivés dont la base n'est pas dans *TLFnome* (ex. *grec*↔*grécité*), 21 dérivés de noms et 57 erreurs. On a donc un rappel de 92% et une précision de 93%.

[18] Contrairement à (2) et (3), dont la solution est linguistique, la résolution de (1) dépend uniquement de l'enrichissement du programme, qui mettra en évidence que *val* ne dérive pas de *valoir*, et inversement.

4.3. Comparaison des deux approches

DériF et DéCor ayant des finalités différentes (rappelons que le premier, basé sur des règles, vise à terme à établir une fiche signalétique complète des mots construits du lexique tandis que le second recourt à des méthodes statistiques pour relier automatiquement un dérivé à sa base), leur comparaison ne peut porter que sur les bases qu'ils proposent.

Couverture. La couverture d'un système – *i.e*. la partie du lexique qu'il est capable de traiter – a deux propriétés : (1) elle est inversement proportionnelle aux connaissances linguistiques que le système met en œuvre ; (2) elle est en relation avec la capacité de ce système à traiter les mots inconnus.

(1) DériF reflète les hypothèses linguistiques théoriques, dans des limites en terme de corpus et dans celles imposées par les contraintes de programmation. L'analyse d'un nouvel affixe revient donc à la définition d'une nouvelle fonction reflet de nouvelles hypothèses linguistiques [19]. Par conséquent, non seulement l'élaboration d'une fonction demande un travail important, mais la quantité de celui-ci est multipliée par le nombre de suffixes du français. À l'inverse, DéCor produit pour chaque dérivé sa base la plus probable d'un point de vue statistique. Ses analyses étant réalisées sans connaissances, il dispose d'une couverture large : aucune modification n'est à apporter au système si l'on souhaite traiter un nouvel affixe (ex. les substantifs en *-tion* dérivés de verbes, les adjectifs en *-ique* dérivés de substantifs, etc.). DéCor fournira pour ces « nouveaux » corpus des résultats d'une qualité comparable à celle des résultats déjà obtenus.

(2) DériF permet de traiter les mots inconnus selon deux aspects. D'une part l'approche par règles permet d'envisager à moindre coût l'analyse des **néologismes**, selon le principe simple consistant à décider que tout mot inconnu a un comportement régulier correspondant au schéma de construction de mots le plus fréquent pour l'affixe considéré. D'autre part, la production de bases absentes du référentiel est possible, car on ne veut pas dépendre des limitations de celui-ci, et est signalée par un « * »[20]. C'est ainsi que **des bases construites non attestées** sont obtenues, comme lors de l'analyse de °*impéchabilité* qui fournit l'arbre : `[[ impéchable* NOM] ité NOM]`, où *impéchable* est absent du référentiel, comme en atteste l'« * », tout en constituant une étape indispensable à l'obtention de *péchable*, ce qui justifie qu'on conserve cette base.

Complémentarité. Une caractéristique intéressante de DéCor est sa systématicité : étant indépendant des connaissances de tout humain en particulier, il est capable d'associer des bases et de proposer des bases auxquelles un humain n'aurait pas pensé, même après avoir consulté un dictionnaire, comme *arer* pour *arable*. DériF, en revanche, dépend directement des compétences linguistiques de ses auteurs et de leurs interprétations. Ce fait apparaît en particulier dans le choix des exceptions. DériF et DéCor présentent, de ce point de vue, une complémentarité certaine : DéCor permet à l'informaticien-linguiste qui souhaite implémenter dans DériF le traitement d'un nouvel affixe, de dégrossir de manière importante le travail d'analyse en lui fournissant des bases correctes pour la plupart les dérivés de son corpus. Ces résultats permettent également de repérer aisément les emprunts (ex. *conductivité* est emprunté à l'anglais ; *conducteur*, la solution proposée par DéCor, est sémantiquement correcte, mais morphologiquement fausse) ainsi que les dérivations pour lesquelles une analyse linguistique

[19] L'ajout d'un suffixe, donc d'une fonction de découpage n'altère toutefois pas le comportement des autres, ce qui permet d'affirmer que la conception de DériF est modulaire.

[20] C'est par validation manuelle que la base notée « * » sera ou pas intégrée dans le référentiel.

plus approfondie est nécessaire.

5. Conclusion

Les comparaisons qui précèdent ont montré que, davantage que concurrents, DériF et Dé-Cor sont complémentaires. Elles ont également montré que, bien qu'implémentant des hypothèses linguistiques, DériF produit une part de résultats erronés ou perfectibles. Au vu de cela, nous conclurons que le lexique dérivationnel idéal (idéal en ceci qu'il engendre le moins de bruit linguistique possible tout en étant le moins coûteux en temps) résulte du passage successif de trois grains : le dégrossissement de la masse des données à traiter est confié à un analyseur basé sur des méthodes statistiques ; l'affinage des données est confié à un analyseur basé sur des règles et sur des listes d'exceptions ; les finitions ultimes sont humaines.

Références

ARONOFF M. (1976), *Word Formation in Generative Grammar*, Linguistic Inquiry, Monograph One, Cambridge, Massachusetts / London, England, The MIT Press.

CLAVIER V. (1996), *Modélisation de la suffixation pour le traitement automatique du français. Application à la recherche d'information*, Thèse de doctorat, Grenoble.

CLAVIER V, WARREN K., LALLICH-BOIDIN G. & STEFANINI M-H. (1996), « Intégration de la morphologie dérivationnelle dans un système distribué d'analyse du français écrit », *Actes de ILN96*, pp. 103-120.

CLEMENCEAU D. (1993), *Structuration du lexique et reconnaissance des mots dérivés*, Thèse de doctorat, Université Paris 7.

CORBIN D. & CORBIN P. (1991), « Un traitement unifié du suffixe *-ier(e)* », *Lexique* 10, pp. 61-145.

CORBIN D. & PLENAT M. (1992), « Note sur l'haplologie des mots construits », *Langue française* 96.

CORBIN D. (1987), *Morphologie dérivationnelle et structuration du lexique*, 2 vol., Tübingen, Max Niemeyer Verlag ; 2e éd., Villeneuve d'Ascq, Presses Universitaires de Lille, 1991.

CORBIN D. (à paraître), *Le lexique construit. Méthodologie d'analyse*, Paris, Armand Colin.

DAL G. (1997a), « Tous les mots existants sont-ils possibles ? », *in* D. Corbin., B. Fradin., B. Habert, F. Kerleroux & M. Plénat éds, *Mots possibles et mots existants, Silexicales* 1, Université de Lille III, pp. 91-100.

DAL G. (1997b), « 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 , pp. 105-115.

GRUAZ C., JACQUEMIN Ch. & TZOUKERMANN E., (1996), « Une approche à deux niveaux de la morphologie dérivationnelle du français », *Journées Lexique*, Grenoble.

HABERT B., NAZARENKO N. & SALEM A. (1997), *Les linguistiques de corpus*, Armand Colin, Masson, Paris.

JACQUEMIN Ch. (1997), « Guessing morphology from terms and corpora », *Proceedings, 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97),* Philadelphia, PA.

KOSKENNIEMI K. (1983), « Two-level model for Morphological Analysis », 8th IJCAI Conference, Karlsruhe.

LE PESANT D. & MATHIEU-COLAS M. éds (1998), « Présentation », *Langages* 131, pp. 3-5.

NAMER F. (1999), « Traitement automatique des mots construits : le cas des noms et adjectifs en –et(te) », *Silexicales* 2, Toulouse.

PLENAT M. (1988), « Morphologie des adjectifs en *-able* », *Cahiers de grammaire* 13, pp. 101-132.

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

XU J. & CROFT W.B. (1998), « Corpus-Based Stemming using Co-occurrence of Word Variants », *ACM Transaction on Information Systems*, vol. 16 :1, pp. 61-81.

ZWEIGENBAUM P. & GRABAR N. (1999), « Acquisition automatique de connaissances morphologiques sur le vocabulaire médical », *soumis à TALN99*.

TLF : *Trésor de la langue française. Dictionnaire de la langue du XIXe et du XXe siècle (1789-1960)*, 16 vol., Paris, Éditions du Centre National de la Recherche Scientifique (t. 1-10) / Gallimard (t. 11-16), 1971-1994, Ciaco éditeur.

WinBrillv0.3 = SOUVAY G. (1998), Version 0.3 du Catégoriseur de Brill pour Windows95/98, INaLF-CNRS, Nancy, consultation et téléchargement : http://jupiter.inalf.cnrs.fr/WinBrill.