Automatiser l'analyse morpho-sémantique non affixale : le système DériF

Fiammetta Namer*

Cet article est consacré au programme DériF, développé dans le cadre du projet MorTAL et de conception inspirée du modèle théorique de D. Corbin, qui effectue l'analyse morpho-sémantique des unités lexicales du français. Plus spécifiquement, cet article s'attache à la présentation d'un aspect peu pris en compte en traitement informatisé de la morphologie, à savoir l'automatisation de l'analyse morphologique non affixale. Tout d'abord, l'article décrit le fonctionnement de DériF en mettant en évidence les propriétés de celui-ci (récursivité, traitement hiérarchisé des procédés de construction de mot, gestion des résultats multiples, codage sémantique …). Ensuite, l'accent est mis sur la manière dont DériF traite les procédés de conversion et de composition dite « savante ».

This paper is devoted to the DériF program, developed within the framework of the MorTAL program, designed according to Danielle Corbin's theoretical model; Dérif carries out the morpho-semantic parsing of lexical units in French. Specifically, the paper seeks to present a rather neglected aspect of computational morphology , namely non affixal morphology parsing. First, the paper describes the way DériF works, and thus highlights the properties of this model, such as recursivity, a hierarchical account of word formation rules, handling of multiple parses, semantic encoding, etc. The paper goes on to focus on the way DériF deals with both conversion processes (also called nominal incorporation) and so-called neoclassical compounding.

* UMR « ATILF » et Université Nancy2.

Cahiers de Grammaire 28 (2003), « Morphologie et Lexique », pp. 31-48

1. Introduction : le projet MorTAL

La réalisation du projet MorTAL (Morphologie pour le TALN)[1] a trouvé sa justification dans la déficience du français en matière de bases de données morphologiques, contrairement à ce que l'on observe en anglais, allemand ou néerlandais par exemple avec la base CELEX (Baayen *et al.*, 1993). L'ambition de MorTAL a été de faire interagir théorie morpho-sémantique, *i.e.* les hypothèses linguistiques énoncées à l'origine par D. Corbin, et pratique, *i.e.* les applications TALN exploitant les aspects sémantiques, catégoriels et structurels de la construction lexicale, comme l'analyse documentaire, la recherche d'information, la compréhension de textes (cf. Dal *et al.*, à paraître). De ce fait, les objectifs du projet, ainsi que les moyens théoriques dont il s'est doté, différaient radicalement de ceux des systèmes de morphologie dérivationnelle existant pour le français. Ceux-ci ne prévoient en effet aucun calcul sémantique et sont consacrés pour l'essentiel (1) à la reconnaissance de mots inconnus (ces systèmes sont alors développés en amont de parseurs syntaxiques robustes), (2) à la structuration automatique de terminologie à partir de patrons graphémiques, ou encore (3) à la constitution automatique de familles morphologiques (pour un panorama des systèmes de traitement automatique de la morphologie du français, consulter *e.g.* Daille *et al.* 2002).

L'un des aboutissements concrets de ce projet est l'analyseur DériF (Dérivation du Français), dont l'élaboration a été pour l'essentiel le fruit d'une collaboration avec G. Dal. Avant d'examiner de façon plus détaillée le fonctionnement de DériF (§2) et sa prise en compte de la morphologie non affixale (§3), rappelons brièvement que les principes fondamentaux de la théorie qu'il implémente sont la régularité du lexique construit, l'associativité du sens et de la forme d'une unité lexicale (UL), et la compositionnalité du sens d'une UL construite par rapport à sa structure ; il en découle une grammaire de construction d'ULs constituée de contraintes phonologiques, sémantiques, catégorielles, etc. hiérarchisées.

2. DériF et la dérivation affixale

DériF s'applique à un lemme issu de la langue générale[2] muni d'une catégorie grammaticale. Les tests à grande échelle sont effectués sur un lexique de référence réunissant la nomenclature du *TLF* et celle du *Grand Robert électronique*, soit un total d'environ 90000 lemmes étiquetés. A ce

[1] Projet ACI « jeunes chercheurs », 1999-2002, financé par le MENRT et piloté par G. Dal, UMR SILEX.

[2] Dans la dernière version, DériF s'applique également à un vocabulaire issu de la terminologie bio-médicale, cf. §3.2.

jour, DériF effectue l'analyse morphologique et sémantique complète des unités lexicales suffixées par *-able, -ité, -et(te), -is(er), -ifi(er), -eur, -ment, -tion*, soit au total environ 10000 unités lexicales construites *;* les procédés morphologiques suivants sont, eux, partiellement couverts : suffixation en *-oir*, préfixation par *dé-, in-, re-, a- en-* (4000 unités lexicales), conversion et composition (cf. §3). Nous allons voir que l'utilisation des contraintes linguistiques mentionnées *supra* permet à DériF de produire en sortie de programme un résultat réunissant des aspects catégoriels et sémantiques, et également de prédire le sens et la structure de mots hors-dictionnaires.

2.1. Fonctionnement

Comme cela est souligné dans Namer (2002) et illustré dans la série d'exemples de la ***Fig. 1***, le système calcule l'arbre d'analyse d'un lemme étiqueté, cet arbre étant repris sous forme de famille ordonnant l'ensemble des bases successives reconnues par l'analyseur ; la troisième partie du résultat consiste en la représentation en langage naturel de la relation sémantique que tisse l'input avec sa base ; la quatrième et dernière partie est l'ensemble des traits sémantiques automatiquement acquis et affectant la base et/ou le dérivé en fonction des contraintes exercées par le procédé de construction.

1) explicabilité, NOM => [[[expliquer VERBE] able ADJ] ité NOM] (explicabilité/NOM, explicable/ADJ, expliquer/VERBE) :: "faculté d'être explicable"
explicabilité (abstrait,propriété,xxx)
explicable (xxx,inhérent,exogène,prédicatif)
expliquer (xxx,transitif, [agent,theme])

2a) introuvable, ADJ => [in [[trouver VERBE] able ADJ] ADJ] (introuvable/ADJ, trouvable/ADJ, trouver/VERBE) :: "non trouvable"
trouvable (xxx,inhérent,exogène,prédicatif)
trouver (xxx,transitif, [agent,theme])

2b) désossable, ADJ => [[dé [os NOM] VERBE] able ADJ] (désossable/ADJ, désosser/VERBE, os/NOM) :: "que l'on peut désosser"
désossable (xxx,inhérent,exogène,prédicatif)
désosser (xxx,transitif, [agent,theme])

3) desservir, VERBE => [dé1 [servir VERBE] VERBE] (desservir/VERBE, servir/VERBE) :: "(Enlever ce qui a pour effet de | Annuler l'état lié au procès) de servir"

desservir, VERBE => [dé2 [servir VERBE] VERBE] (desservir/VERBE, servir/VERBE) :: "Cesser de servir; servir fortement, intensément, jusqu'au bout, au loin"

4) benladenisation, NOM => [[[Benladen NPR] is(er) VERBE] tion NOM] (benladenisation/NOM, benladeniser/VERBE, Benladen/NPR) :: "action ou résultat de benladeniser"
benladenisation (abstrait,action/résultat,xxx)
benladeniser (causatif, transitif, [cause,thème])

Fig. 1 : *Exemples d'analyses de DériF*

Le mécanisme d'**acquisition de traits sémantiques** exploite les contraintes régulières exercées par les procédés de construction de mots sur les unités lexicales appropriées, que des études théoriques ont permis de mettre en évidence. Ainsi, la ***Fig. 2*** illustre la façon dont DériF implémente les informations syntactico-sémantiques décrites dans Corbin (1987), et, entre autres, dans les travaux sur les aspects sémantiques des déverbaux d'action (Kelling 2001, Jacquey 2002, Lecomte 1997), des noms de propriété (Dal *et al.* 1999), des collectifs (Aliquot-Suengas 1996), des verbes de changement d'état (Plag 1999, Roger 2003), des adjectifs dénominaux (Mélis-Puchulu 1991) ou déverbaux (Plénat 1988), les prédicats de dissociation (Aurnague & Plénat 1997, Gerhard 1998), etc. Les traits (ou `'xxx'` en cas de sous-spécification) sont organisés en listes. Comme il est expliqué dans Namer (2002), une même UL peut hériter de traits issus de plusieurs procédés morphologiques. La lecture de la ***Fig. 2*** est illustrée par l'exemple de la première ligne : quand DériF analyse un verbe comme affixé par *a-, é-, -is(er)* ou *-ifi(er)* sur base adjectivale, alors le verbe est automatiquement codé causatif transitif et ses arguments reçoivent les cas cause (pour le sujet) et thème (pour l'objet direct). L'adjectif base résultant de l'analyse est lui codé comme qualificatif, décrivant une propriété acquérable.

affixe	base	dérivé
a-, é-,-is(er), -ifi(er)	A = (xxx, acquérable, xxx, predicatif) *(tendre, court)*	V = (causatif, transitif, [cause, thème]) *(attendrir, écourter)*
dé1-	A= (transitoire, xxx,xxx, predicatif) *(las)*	V= (causatif ...) | (résultatif, intransitif, [thème]) *(délasser)*
-oir	V= (xxx, xxx,xxx) *(hacher, mourir)*	N=(concret, inanimé, lieu/instr) *(hachoir, mouroir)*
-able	V = (xxx, transitif, [agent,thème]) *(laver, encastrer)* V = (xxx, intransitif, [agent,sur(lieu)]) *(skier)*	A = (xxx, inhérent, exogène, prédicatif) *(lavable, encastrable, skiable)*
-aie	N= (concret, inanimé, végétal) *(bananier)*	N= (concret, inanimé, collectif) *(bananeraie)*
-aille	N= (concret, xxx, xxx) *(flic, fer, tripe)*	N= (concret, xxx, collectif) *(flicaille, ferraille, tripaille)*

Fig. 2 : *Acquisition de traits sémantiques*

En dehors de cette fonctionnalité, les caractéristiques principales de DériF sont (a) la **récursivité** : l'analyse d'un mot est réitérée jusqu'à l'obtention d'une base non analysable (ex. 1 de la ***Fig. 1***) ; (b) le **traitement hiérarchisé** de diverses opérations de construction (suffixation, préfixation, conversion, composition), en fonction de la portée respective des procédés en présence : ainsi, les exemples (2a) et (2b) de la ***Fig. 1*** illustrent l'ordre inverse d'analyse de la préfixation et la suffixation ; (c) la **gestion des analyses ambiguës** : DériF manipule des listes de données et de résultats, et est donc capable de générer autant de solutions qu'il y a d'ambiguïtés

éventuelles dans l'analyse d'un mot construit (ex. 3) ; (d) le **traitement des néologismes** : de par sa conception, DériF est capable d'analyser des mots inconnus, possibles et non attestés (ex. 4).

2.2. Illustration

Le mécanisme de DériF, illustré, dans la ***Fig. 3***, par l'analyse de l'adjectif *indéboulonnable,* consiste en l'application de l'une des deux méthodes générales suivantes, sur un lemme morphologiquement construit et muni d'une catégorie grammaticale, considéré comme l'input. Chaque méthode est pilotée par l'un des deux moteurs de DériF : le premier (MOTEUR1) s'occupant des lemmes suffixés, le second (MOTEUR2) prenant en charge les mots construits privés de marque suffixale.

Quand l'input possède un suffixe, alors l'analyse de ce lemme par rapport à ce suffixe est envisagée en priorité. Cette analyse débute par une recherche préalable d'éventuels préfixes portant sur la base suffixée. Cette recherche est bien entendu subordonnée à la valeur du suffixe courant : par exemple, pour le suffixe *-able*, certains préfixes sont examinés : *in-, auto-, super-, hyper-…* contrairement à d'autres, catégoriellement et/ou sémantiquement incompatibles : *re-, dé-, pré-*[3]. C'est également à ce stade que va se déclencher la recherche de lemmes morphologiquement ambigus (comme p.ex. *inversable, imposable, importable* qui peuvent donner lieu à deux analyses possibles). Dans la ***Fig. 3***, cette recherche conduit à la détection du préfixe *in-* et ramène la base adjectivale *déboulonnable* de *indéboulonnable.* Chaque étape d'une analyse comprend en outre systématiquement : l'appariement des formes allomorphiques, l'identification de la catégorie de la base, et les calculs sémantiques (cf. ***Fig. 1*** et ***Fig. 2***). C'est ainsi que le verbe *déboulonner* est produit, ainsi que la relation sémantique entre *indéboulonnable* et sa base et les traits lexicaux affectant *déboulonnable* et *déboulonner*. Cette **première méthode**, orchestrée par le premier moteur, est répétée tant qu'un suffixe est détecté.

Si l'input est non suffixé, ou dès que l'analyse produit un lemme non suffixé, comme c'est le cas pour déboulonner dans la ***Fig. 3***, une autre méthode, prise en charge par le second moteur, ne prévoit alors que la recherche et l'analyse de formes préfixées, converties ou composées, jusqu'à l'obtention d'un lemme morphologiquement indécomposable. Ici, la méthode ne s'applique qu'une fois, la base nominale *boulon* du verbe préfixé par le *dé-* formateur de verbes soit perfectifs (ou « ablatifs », selon Martinet (1985)) soit « privatifs » (Gerhard 1998) étant non construite.

[3] Cette liste de préfixes n'est pas exclusive, et pour chacun d'eux le programme répertorie un certain nombre d'exceptions : ainsi, *superposable* n'est pas construit sur *posable*, et à l'inverse, *défavorable* a pour base *favorable*.

Le bas de la Fig . 3 illustre enfin l'affichage des résultats obtenu à l'issue de l'analyse de *indéboulonnable*, suivant la structure décrite au § 2.1.

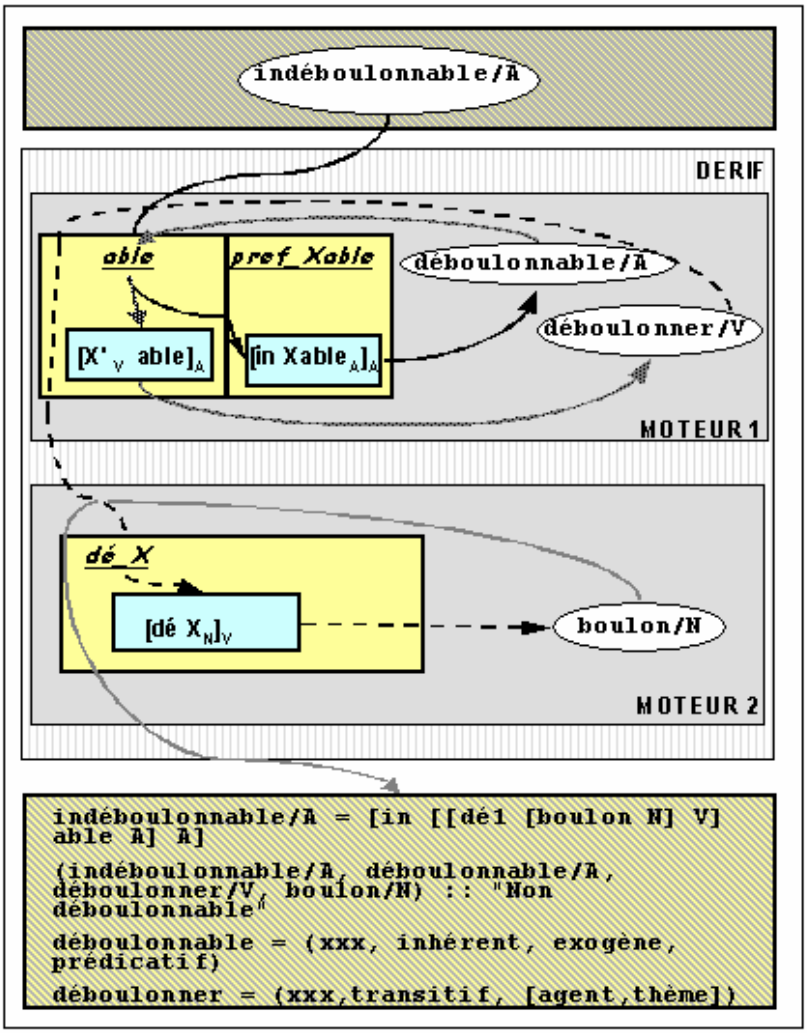

Fig. 3 : *Analyse par DériF de l'adjectif «* indéboulonnable *»*

3. DériF et la morphologie dérivationnelle non affixale

Très souvent, une vision réductrice de la morphologie constructionnelle associe à tort le lexique construit aux seuls résultats de la préfixation et de la suffixation, au détriment de la conversion et de la composition, que nombre d'auteurs classent comme des procédés syntaxiques, comme le relèvent, respectivement, Kerleroux (1996) et Villoing (2002). Ce point de vue est alimenté en grande partie par l'apparente hétérogénéité du fonctionnement de l'affixation d'une part, de la conversion et de la composition d'autre part. En effet, alors que l'affixation met en présence deux types de matériau lexical – une base et un affixe, le premier étant subordonné au second – l'input de la conversion est, lui, limité à une unité lexicale unique ; par ailleurs, si la composition fait effectivement intervenir deux composants, ceux-ci n'entretiennent à première vue pas le même rapport de subordination que celui qu'on observe en préfixation ou en suffixation.

Malgré leurs comportements apparemment hétérogènes, ces quatre types de procédés morphologiques obéissent aux mêmes principes fondamentaux de régularité, autonomie et associativité décrits entre autres dans Corbin (1987). Cette section est consacrée à la modélisation de la conversion et de la composition dans DériF. En tant que modules d'analyse dépendants des moteurs de la phase 1 et 2, ces deux types de procédés sont intégrés dans l'analyse d'un mot construit suivant les mêmes principes de hiérarchisation, récursivité, calculs sémantiques et gestion des résultats ambigus énoncés et illustrés au §2. Cependant, la spécificité de chaque type de procédé engendre des difficultés particulières en terme de formalisation, que nous détaillons ci-dessous, ainsi que les solutions proposées.

3.1. La conversion : comment orienter l'analyse automatiquement ?

Comme l'explique F. Kerleroux (2000), le procédé morphologique de conversion pose des problèmes de choix théoriques (comment définir une morphologie concaténatoire quand il n'y a rien à concaténer, comme c'est le cas en conversion ?) et définitoires (qu'est-ce qu'une unité lexicale ?) qui constituent autant d'obstacles à l'identification de ce procédé par des auteurs comme Bailly (1965), MacNamara (1982) ou encore Spencer (1991). L'adoption du point de vue théorique défendu par D. Corbin permet de proposer une analyse de la conversion similaire à l'analyse des procédés d'affixation. Une opération de conversion associe une base et un dérivé, appartenant à deux catégories grammaticales nécessairement différentes, et entretenant une relation sémantique particulière. C'est ainsi que $V \rightarrow_{CONV} N$[4]

[4] Rappelons que le modèle théorique dont ce travail s'inspire considère le matériel flexionnel (et donc les terminaisons verbales) comme n'intervenant pas

sous-entend généralement que N décrit le procès véhiculé par V (*vol(er)* $\rightarrow_{CONV}$ *vol*), alors que N $\rightarrow_{CONV}$ V donne à voir le référent de N comme l'instrument du procès décrit par V (*balai* $\rightarrow_{CONV}$ *balay(er)*)[5]. L'application de ce principe d'orientation de l'analyse selon des critères sémantiques est cependant problématique pour un système dont l'input est dépourvu d'informations sémantiques. Pour éviter cependant un usage systématique des listes d'exceptions, le traitement par DériF de la conversion distingue deux cas de figures : on envisage d'une part les unités lexicales obtenues par conversion sur une base suffixée ou préfixée ; d'autre part, les UL converties dont la base est indécomposable.

3.1.1. UL convertie sur base suffixée

La détection d'une UL obtenue par conversion sur une base suffixée est assez simple, dans la mesure où le suffixe observé sur l'UL est incompatible avec la catégorie de celle-ci. Le premier moteur de DériF (cf. § 2.2.) vérifie alors que la catégorie qui correspond au suffixe pourrait se révéler celle de la base dans une opération de conversion donnant lieu à l'UL examinée. Dans ce qui suit, nous examinons cas par cas les différents types d'opérations de conversion effectuées par DériF et aboutissant à une base suffixée, en suivant le raisonnement ci-dessus.

Asuff → N : Les noms comme : *portable, maxillaire, burlesque, labiale, hollandais, américain* possèdent tous un suffixe d'adjectif dénominal. Or la conversion A → N (dite de « focalisation ») existe, le sens construit de N pouvant se gloser : "Ce(lui) dont la propriété vue comme saillante est d'être A".

Ndéverbal → V : On observe de nombreuses unités lexicales verbales qui se terminent apparemment par un suffixe formateur de noms déverbaux, comme *-tion* (*collectionner*), *-age* (*grillager*), *-ure* (*hachurer*), *-ment* (*ornementer*). Etant donné l'ambivalence de la conversion reliant N et V, se pose *a priori* la question de l'orientation : ainsi, *collectionner* est-il la base d'une opération

dans la construction du lexique, la morphologie flexionnelle servant au contraire à réaliser l'interface entre lexique et syntaxe. Les unités lexicales, sur lesquelles s'appliquent les procédés morphologiques, sont donc privées de toute marque flexionnelle, ce qui explique que les phénomènes de conversion N → V (*vol* → *vol(er)*), V → N (*sing(er)* → *singe*) et A → V (*vide* → *vid(er)*) soient envisageables.

5 Dans cette section sont modélisées les opérations de conversion recensées et analysées, entre autres, dans Corbin (2000), et Kerleroux (1996, 1997, 2000). Voir également Namer (2002), pour l'implémentation des contraintes sémantiques concernant base et mot construit dans le cadre de **A** $\rightarrow_{CONV}$ **V.**

de conversion dont le résultat est *collection*, ou est ce l'inverse qui se produit ? D. Corbin (1987), qui choisit la deuxième solution, argumente sa décision comme suit : quand le nom en jeu est lui-même déverbal, le parallélisme des opérations de construction de mot qui résulteraient de la première hypothèse aboutirait à une structure incohérente, où le nom *collection* aurait à la fois pour base le verbe *collecter* (ce qui est le cas) et le verbe *collectionner* :

collect(er) $\rightarrow_{\text{-tion}}$ collection
collectionner $\rightarrow_{\text{CONV}}$ collection

Les déverbaux sont les bases nominales les plus fréquentes que nous ayons observées pour la construction de V dénominaux par conversion. Leur rôle, dans la structure argumentale du verbe, qui est transitif, est le plus souvent celui d'instrument : *grillager, réglementer, bordurer, clôturer…* Les bases en *-tion* semblent, elle, jouer plutôt le rôle de But (*partitionner, collectionner*). Les contraintes sémantiques sur N liées à la conversion N → V (N = instrument, ou lieu) et la relation qu'entretiennent V et N (V = « agir sur qqc en se servant de N, ou dans le but de faire N ») font cependant apparaître quelques cas de bases dénominales (évaluatifs), comme *brumasser, brouillasser*. On n'observe pas, en revanche de bases nominales désadjectivales (noms de propriété), ce qui est sémantiquement explicable.

Vdéverbal → N : Les verbes construits au moyen d'un suffixe évaluatif sont des bases possibles pour des nominalisations par conversion, qui désignent alors le procès, ou le résultat du procès, décrit par V : *tremblote*, *parlote*[6]… Le même raisonnement qu'au paragraphe ci-dessus mène à la conclusion que l'orientation inverse de la conversion est inenvisageable : si *parloter* était issu de *parlote*, par N $\rightarrow_{\text{conv}}$ V, alors ce verbe aurait deux bases, puisqu'il est déverbal (*parl(er)* $\rightarrow_{\text{-ot(er)}}$ *parloter*).

Aconstruit → V : Le verbe obtenu désigne un changement d'état (*impatient*$_A$ → *impatienter*$_V$, *inactif*$_A$ → *inactiver*$_V$) ; à la fin du déroulement de ce procès le thème de V a acquis la propriété décrite par A (il est *impatient,* resp. *inactif*). Il est alors sémantiquement prédictible que seuls les adjectifs dénominaux, dont le codage sémantique consécutif à l'analyse de V est la liste : `(xxx, acquérable, xxx, prédicatif)` (cf. §2.1) constituent des candidats pertinents pour servir de base à de tels verbes, la relation entre base et verbe désadjectival étant identique quel que soit le mode de formation des verbes désadjectivaux de changement d'état (cf. Roger (2003) pour une remise en question de cette affirmation, qui est cependant suffisante à notre propos). En

[6] Je remercie F. Kerleroux de m'avoir suggéré ces exemples.

particulier (cf. Namer & Dal 2000), les adjectifs déverbaux en *-able* constituent des bases candidates très improbables. Les A évaluatifs semblent également rétifs à fonctionner comme base pour la construction de tels verbes.

Nsuff → A : En général, la conversion N → A est moins fréquente que les précédentes. Ce procédé programme l'adjectif qu'il construit à désigner la propriété (généralement liée à une forme, couleur, comportement) qui caractérise le nom de base. Seuls très peu de noms, apparemment tous non construits (*carré, rond, marron, rose, vache…*), semblent concernés.

3.1.2. UL convertie : autres types structurels de base

Si DériF ne détecte aucune forme suffixoïde en contradiction avec la catégorie grammaticale, alors son second moteur (cf. § 2.2.) s'active de manière à vérifier si l'UL en présence est obtenue par conversion d'une base préfixée. Le principe est le même : DériF tente de détecter une incompatibilité entre la catégorie de l'UL à analyser et le préfixe ayant apparemment construit cette UL. Les exemples relevés montrent des cas de :

prefV $\rightarrow_{\text{CONV}}$ N : *dépose*$_{\text{N}}$, *entremise*$_{\text{N}}$. Les cas observés ne font état que de verbes préfixés déverbaux (*déposer*, *entremettre*).

prefA $\rightarrow_{\text{CONV}}$ N : *sous-marin*$_{N}$, *impur*$_{N}$. On remarque que l'adjectif préfixé a lui-même une base soit nominale (*mer* $\rightarrow_{\text{sous-}}$ *sous-marin*) soit adjectivale (*pur* $\rightarrow_{\text{in-}}$ *impur*).

prefA $\rightarrow_{\text{CONV}}$ V : *inactiver*$_{\text{V}}$. La base de l'adjectif préfixé est ici adjectivale (*actif* $\rightarrow_{\text{in-}}$ *inactif*), ce qui n'exclut pas *a priori* d'autres combinaisons catégorielles.

Enfin, on peut s'attendre à quelques cas de bases elles-mêmes obtenues par conversion, comme cela est souligné dans Corbin (2000), qui cite par exemple la chaîne constructionnelle :

orange$_{\text{N+concret}}$ → *orange*$_{\text{A}}$ → *orange*$_{\text{N+abstrait}}$.

Restent les UL converties à partir de bases non construites, dont la détection automatique présente un degré de complexité différent selon que l'orientation du procédé est réversible ou pas. La conversion **V $\rightarrow_{\text{CONV}}$ A** n'existant pas ce pour des raisons sémantiques (cf. Corbin 2000), la détection et le traitement des verbes convertis désadjectivaux (*pâlir*, *vider* …) ne pose pas de problème particulier. De même, le traitement de **A $\rightarrow_{\text{CONV}}$ N** est relativement simple (*bleu*$_{\text{N}}$, *idiot*$_{\text{N}}$ …), les cas de conversion inverse (**N $\rightarrow_{\text{CONV}}$ A**) étant suffisamment peu fréquents (*rose*$_{\text{A}}$, *cochon*$_{\text{A}}$, *carré*$_{\text{A}}$) pour

justifier leur inscription dans une liste d'exceptions. Le principe de reconnaissance de noms désadjectivaux convertis est déclenché par l'identification, dans le lexique de référence, d'un adjectif homographe[7] au nom en cours d'analyse. Reste la reconnaissance des conversions **V** $\rightarrow_{\text{CONV}}$ **N** et **N** $\rightarrow_{\text{CONV}}$ **V**, pour lesquelles on ne dispose quasiment d'aucun moyen automatique pour décider de l'orientation du procédé. Ainsi, en présence de *balay(er)*, ou de *vol(er),* et bien qu'ayant détecté dans le référentiel les noms quasi-homographes *balai* et *vol*, DériF n'est pas capable de décider si les verbes sont dénominaux, ou les noms déverbaux. Les rares exceptions à cet état de fait sont constituées par les paires (N,V) où N est porteur d'une marque sémantique (acquise automatiquement dans le cadre de l'analyse d'une autre UL dont il constitue un maillon de la chaîne constructionnelle, cf. §2.1) forçant l'orientation de la conversion : si le nom est abstrait, DériF l'analyse comme le résultat de la conversion **V** $\rightarrow_{\text{CONV}}$ **N** ; dans le cas contraire, comme la base de **N** $\rightarrow_{\text{CONV}}$ **V**[8]. Par exemple, le nom *flic*, base de *flicaille*$_{\text{N}}$, est codé dans le référentiel comme concret (cf. Aliquot-Suengas 1996), ce qui oriente le couple (*flic, fliquer*) dans le sens **N** $\rightarrow_{\text{CONV}}$ **V**, le nom concret jouant dans le prédicat verbal converti le rôle d'instrument ou d'agent. En l'absence de marque sémantique discriminante, seule l'utilisation de listes d'exceptions est envisagée pour l'analyse de **N** $\rightarrow_{\text{CONV}}$ **V** et **V** $\rightarrow_{\text{CONV}}$ **N**, mais cet emploi soulève une autre question : qui, des noms abstraits ou concrets, doivent être vus comme des exceptions, et selon quels critères ?

3.1.3. Exemples

Les exemples suivants illustrent l'analyse par DériF de différents cas de conversion présentés au § 3.1. Seuls les traits sémantiques acquis par le biais de l'opération de conversion sont indiqués.

[7] Cette identification s'effectue, en ce qui concerne le nom, aux variantes flexionnelles près : ainsi, l'analyse du nom féminin *idiote* comprend une phase préalable de neutralisation de la marque de genre.

[8] Il s'agit d'une première approximation, et la comparaison pour N et V de la date d'observation de leur première occurrence servira à mettre à jour les contre-exemples.

```
imperméable, N => [[ in- [ [permé(er) V] -able A] A] N] (imperméable/N, imperméable/A,
perméable/A, perméer/V) :: "Ce(lui) dont la propriété vue comme saillante est d'être imperméable".

hachurer, V => [[[hach(er) V] –ure N] V] (hachurer/V, hachure/N, hacher/V) :: "faire quelquechose au
moyen de hachure(s)"
        hachurer            (xxx, transitif, [agent, theme, instrument])

bougeote, N => [[[boug(er) V] –ot(er) V] N] (bougeote/N, bougeoter/V, bouger/V) :: "action de
bougeoter"
        bougeote            (abstrait, action/resultat, xxx)

blanchir, V => [[blanc A] V] (blanchir/V, blanc/A) :: "Rendre/devenir blanc"
        blanchir            (causatif,transitif,[cause,thème]) / (résultatif, intransitif, [thème])
        blanc               (xxx,acquérable,xxx, prédicatif)
```

3.2. Les mots composés « savants »

De nombreux auteurs nient à la composition le statut de procédé de construction lexicale. Ainsi, certains (Di Sciullo & Williams 1987, Lieber 1992, Zwanenburg 1992, Barbaud 1991, 1994) y voient un procédé purement syntaxique, alors que d'autres, comme Aronoff (1994) lui accordent un statut certes intermédiaire entre syntaxe et morphologie, mais où la syntaxe joue un rôle prépondérant ; pour Anderson (1992), la nature quasi-syntaxique de la composition se justifie en raison des relations argumentales mises en jeu, identiques à celles observées en syntaxe. On trouvera dans Corbin (1992, 1997), Villoing (2002) et Lesselingue (en préparation) de nombreux arguments contredisant ces affirmations, et justifiant de ce fait l'adoption de méthodes d'analyse automatique de la composition similaires à celles utilisées pour la suffixation, c'est à dire reposant sur les principes théoriques fondamentaux de construction du lexique rappelés au § 1.

D. Corbin (2000) propose une classification des opérations de composition reposant sur la position de l'élément recteur X, en attribuant le statut de « mot composé savant » aux mots composés pour lesquels X est quasiment toujours à droite[9] : *phonophobe*$_A$, *électrochoc*$_N$. Du point de vue de l'analyse automatique, ces constructions savantes présentent un double intérêt. Tout d'abord, en raison de leur créativité lexicale : on dénombre, dans les domaines spécialisés comme la médecine, 45% de « composés savants » parmi les néologismes. L'absence de traitement systématique de la composition savante constitue l'autre motivation : à ce jour, les analyses

[9] On trouve quelques exceptions à cette règle, telles que *philosophe*$_A$ ou *myocarde*$_N$, où la tête (*philo, myo*) précède le modifieur : 'qui aime la science', 'muscle du cœur'.

proposées sont plus pragmatiques que linguistiques, et relèvent essentiellement de la morphologie concaténatoire (Baud *et al.,* à paraître, Lovis *et al.,* 1998). En associant le calcul sémantique à la décomposition structurale, les analyses de DériF, dont nous détaillons ci-dessous les spécificités du mécanisme, permettent *a priori* à la fois de reconnaître les mots nouveaux, et d'en calculer le sens, et de ce fait d'identifier les synonymes.

3.2.1. Aspects structurels, catégoriels et sémantiques

Un mot composé savant, de catégorie A ou N, possède une structure [YX], X constituant l'élément recteur. Aussi bien X que Y sont soit des unités lexicales (UL), soit des bases supplétives non autonomes (que D. Corbin appelle unités infra-lexicales, et que nous abrégeons par UIL). Y est soit un N, soit un A, X pouvant également être de catégorie V[10]. A chaque type de [YX] correspond un sens construit, calculable en fonction de X, de Y et du type catégoriel du résultat, et généralement indépendant du statut (UL ou base supplétive) des composants. La ***Fig. 4*** illustre les différents cas observés mettant en jeu les combinaisons autorisées de catégories de X, Y, [YX], le statut de X et Y, les sens construits (la catégorie soulignée correspond à l'élément recteur). On remarque que quand X est un verbe, alors il ne peut être que de type UIL.

$[YX]_{CAT}$	Statut de X et Y	instruction sémantique	Exemple
$[N\underline{N}]_N$	X = UIL, Y=UIL X = UL, Y = UIL X = UIL, Y = UL X = UL, Y = UL	« <u>N</u> particulier en rapport avec N »	gastralgie gynogénèse marconigramme parasitophobie
$[A\underline{N}]_N$	X = UIL, Y=UIL X = UL, Y = UIL X = UIL, Y = UL X = UL, Y = UL	« <u>N</u> qui est A »	mégalithe similimatière chimiotaxie anglo-catholicisme
$[N\underline{V}]_N$	X = UIL, Y=UIL X = UIL, Y = UL	« ce/celui qui <u>V</u> (le) N »	homicide bactérioscope
$[A\underline{N}]_A$	X = UIL, Y=UIL X = UL, Y = UIL X = UIL, Y = UL	« qui possède un <u>N</u> qui est A »	microcéphale rectiforme berbérophone
$[A\underline{A}]_A$	X = UIL, Y=UIL X = UL, Y = UIL X = UL, Y = UL	« qui est d'un sous-type de <u>A</u> caractérisé par (des propriétés) A »	paridigitidé mésocrânien libano-égyptien

[10] Si X (ou Y) est une base supplétive, elle n'appartient à aucune des catégories N, A, V (cf. § 3.2.2.) ; on dira plutôt qu'elle s'identifie étymologiquement à un N, un V ou un A (ainsi *hydro*, apparenté à *eau*, s'identifie à un nom).

$[N\underline{V}]_A$	X = UIL, Y=UIL X = UIL, Y = UL	« qui V (le) N »	gémellipare sexophobe
$[N\underline{N}]_A$	X = UIL, Y=UIL X = UL, Y = UIL X = UIL, Y = UL X = UL, Y = UL	« qui a un <u>N</u> possédant la caractéristique N »	anthropomorphe filiforme palmipède palmiforme

Fig. 4 : *Sens et structure des composés* $[YX]_{A/N}$

3.2.2. Analyse automatique

DériF est confronté à deux difficultés dans l'analyse des mots composés savants : le stockage des bases supplétives (qui ne sont pas des éléments du lexique de référence) et le déroulement de l'analyse proprement dite, ainsi que la manière dont celle-ci s'insère, relativement aux deux phases d'activation, dans une chaîne constructionnelle.

La représentation des bases non autonomes nécessite la constitution d'un lexique, sous forme de tables de correspondances, où sont appariées, comme l'illustre la ***Fig. 5***, la base, sa traduction (approximative) et la catégorie de celle-ci. Souvent, dans le jargon bio-médical tout du moins, deux bases supplétives (l'une grecque, l'autre latine) alternent pour une même UL. C'est ce qui se passe notamment avec les noms : *aqua/hydro* ↔ *eau*, *pédi/podo* ↔ *pied*, plus rarement avec les adjectifs : *recti/ortho* ↔ *droit*. C'est par le biais de la traduction commune des bases en alternance que va pouvoir s'effectuer la reconnaissance des synonymes.

UIL	UL correspondante (Cat)	UIL	UL correspondante (Cat)
gémel	jumeau (N)	phage	se nourrir (V)
hydro	eau (N)	vore	se nourrir (V)
aqua	eau (N)	recti	droit (A)
pare	engendrer (V)	ortho	droit (A)

Fig. 5 : *Table de correspondances UIL → Trad (Cat)*

L'analyse des composés savants est du ressort du moteur de la phase 1 ou 2 de DériF, selon qu'un suffixe a été détecté ou non dans la séquence à analyser. La consultation de la table de correspondance permet d'identifier, le cas échéant, les bases en présence, ainsi que la catégorie de leur traduction. La valeur de ces catégories, ainsi que leur position relative, donne accès à la glose paraphrasant le sens construit du composé (ex. 1)[11]. Si le composé sert à son tour de base à une UL complexe, l'analyse de cette dernière s'effectue

[11] Une seule étiquette catégorielle, FWD (foreign word) identifie toutes les bases non autonomes.

normalement, et la hiérarchie des opérations correspond à ce qui est relaté au § 2. (ex. 2) :

1) rectident, A => [[recti FWD] [dent N] A] (rectident/A, dent/N) :: "qui a le dent recti=droit"
2) gémelliparité, N => [[[gémel- FWD][pare FWD] A] -ité N] (gémelliparité/N, gémellipare/A, pare/FWD=engendrer) :: "faculté d'être gémellipare"

3.2.3. Détection de la synonymie

L'analyse de termes comme *rectident*$_A$ */orthodonte*$_A$, *aérophage*$_A$ */aérivore*$_A$ ou encore *archicérébron*$_N$*/archencéphalon*$_N$ conduit à l'identification de sens construits identiques (respectivement : "qui a les dents droites", "qui se nourrit d'air" et "cerveau primitif"), qui traduisent la synonymie des composés. Bien sûr, cette synonymie ne s'observe pas uniquement dans le cadre de l'analyse des composés : toute UL construite incluant une base supplétive peut avoir un équivalent sémantique, *e.g. gastrique*$_A$ / *stomacal*$_A$, ou *hémateux*$_A$ / *sanguin*$_A$.

4. Perspectives, conclusions

DériF est d'ores et déjà intégré dans une chaîne d'analyse de corpus permettant le traitement complet de documents en ligne en vue de la conception d'une base de données lexicales interrogeable en morphologie (cf. Namer, à paraître) ; en outre, la mise en pratique de ce système prévoit, dans le cadre des projets UMLF[12] et VUMEF[13], l'analyse et l'annotation morpho-sémantique par DériF de termes médicaux construits, avec pour objectif la création à terme d'un lexique francophone de la langue médicale (cf. Zweigenbaum *et al.* 2003, Darmoni *et al*., à paraître).

L'évolution de DériF, qui est de conception modulaire, prévoit naturellement l'adjonction de la grammaire d'analyse des procédés de construction non encore pris en compte à ce jour. Leur ordre d'intégration va être décidé à partir de critères quantitatifs, où les procédés les plus créatifs seront privilégiés dans un premier temps. Enfin, signalons que parmi les améliorations de DériF envisagées à court terme, un effort particulier va être consacré à la gestion des analyses multiples, telles qu'elles sont illustrées dans l'exemple 3 de la ***Fig. 1*** au § 2.1. En effet, les caractéristiques de DériF en matière de récursivité font que le système, dès qu'il détecte une UL morphologiquement ambiguë, déclenche systématiquement la procédure

[12] Projet ACI n° 02C0163, 2002-2004, piloté par P. Zweigenbaum, STIM/DSI, Assistance Publique, Hôpitaux de Paris.

[13] Projet labellisé par le Ministère de la Recherche dans le cadre du programme RNTS 2003 (Réseau National des Technologies de la Santé), piloté par la société Vidal, et coordonné par S. Darmoni, L@STICS, CHU de Rouen.

d'analyse multiple. Ainsi, par conséquent, une liste de résultats est – justement – produite lors de l'analyse *e.g.* de $dépiler_V$, construit soit sur $pile_N$, soit sur $poil_N$. Cependant, l'ambiguïté peut disparaître, dans certains cas, quand le verbe est lui-même la base d'autres mots construits. Et c'est ce que DériF est à ce jour incapable de détecter : ainsi, si *dépiler* est la base de $dépilage_N$, les deux interprétations sont valides, ce qui justifie l'analyse ambiguë. Mais si *dépiler* résulte lui-même de l'analyse de $dépilation_N$, seul le nom *poil* reste une racine valide alors qu'à l'inverse, pour $dépilement_N$, la chaîne constructionnelle ne doit aboutir qu'à $pile_N$. Dans un cas comme dans l'autre, où les contraintes sémantiques sur le nom d'action ne sont pas clairement établies, seule la méthode des listes d'exceptions est actuellement envisageable pour bloquer la mauvaise analyse. Ajoutons pour finir que l'exemple de *dépiler* ne constitue pas un cas isolé : quasiment toutes les ULs constructionnellement ambiguës ne sont employées que dans un seul sens lorsqu'elles interviennent à leur tour dans un procédé de construction morphologique.

Références bibliographiques

Anderson, S.R. (1992), *A-morphous morphology*, Cambridge, Cambridge University Press.

Aronoff, M. (1994), *Morphology by itself*, Cambridge (Mass.), MIT Press.

Aliquot-Suengas, S. (1996), *Référence collective/sens collectif. La théorie du collectif dans les noms suffixés du lexique français*, thèse de doctorat, Univ. de Lille III.

Aurnague, M. & Plénat, M. (1997), « Manifestations morphologiques de la relation d'attachement habituel », *in* Corbin, D. *et al.* (éd.), *Mots possibles, Mots existants,* Villeneuve-d'Ascq, *Silexicales* 1, pp. 15-24.

Baayen, R.H., Piepenbrock, R. & van Rijn, H. (1993), *The CELEX Database (CD-ROM)*, Philadelphia, Univ. of Pennsylvania, Linguistic Data Consortium.

Barbaud, P. (1991), « Fondements grammaticaux de l'acquisition des mots composés », *Canadian Journal of Linguistics* 36-3, pp. 215-253.

Barbaud, P. (1994) « Conversion syntaxique », *Linguisticae Investigationes* XVIII-1, pp. 1-26.

Baud, R. (à paraître), « Morphosemantems for Words Composition », soumis à *AMIA2003,* Washington, DC.

Bailly, C. (1965), *Linguistique générale et linguistique française*, Berne, Francke.

Corbin, D. (1987), *Morphologie dérivationnelle et structuration du lexique*, 2 vol., Tübingen, M. Niemeyer Verlag, (2ème éd. PUL, 1991).

Corbin, D. (1992), « Hypothèses sur les frontières de la composition nominale », *Cahiers de grammaire* 17, pp. 26-55.

Corbin, D. (1997), « Locutions, composés, unités polylexématiques : lexicalisation et mode de construction », *in* M. Martins-Baltar (éd.), *La locution, entre langue et usages,* Fontenay-aux-Roses, ENS Editions, pp. 55-102.

Corbin, D. (2000), « French (Indo-European: Romance) », *in* G. Booij, C. Lehmann & J. Mugdan (eds), *Morphology. An International Handbook on Inflection and Word Formation*, Berlin / New York, Walter de Gruyter, vol 1, art. 121.

Daille, B., Fabre, C. & Sébillot, P. (2002), « Applications of computational morphology » *in* P. Boucher (ed.), *Many Morphologies*, Somerville, MA, Cascadilla Press, pp. 210-234.

Dal, G., Hathout, N. & Namer, F. (1999), « Construire un lexique dérivationnel : théorie et réalisations », *in TALN 1999*, Cargèse, pp. 115-124.

Dal, G., Hathout N. & Namer, F. (à paraître), « Morphologie constructionnelle et traitement automatique des langues : le projet MorTAL », *Lexique* 16.

Darmoni *et al.* (à paraître), « VUMeF : Extending the French Part of the UMLS », in *AMIA2003*, Washington, DC.

Di Sciullo, A.M & Williams, E. (1987), *On the definition of Words*, Cambridge (Mass.), MIT Press.

Gerhard, F. (1998), « Le préfixe *dé-* dit négatif et la notion d'éloignement », *Scolia* 11, pp. 68-90.

Jacquey, E. (2002), « Les déverbaux d'action en français : quel type d'ambiguïté et quelle catégorie conceptuelle », *in Représentations du sens linguistique,* Lincom Europa.

Kelling, C. (2001), « Agentivity and suffix selection », *in Proceedings of the LFG 2001 conference*, CSLI Publ., pp. 147-162.

Kerleroux, F. (1996), *La coupure invisible,* PUS, Villeneuve-d'Ascq, Presses Universitaires du Septentrion.

Kerleroux, F. (1997), « De la limitation de l'homonymie entre noms déverbaux convertis et apocopes de noms déverbaux suffixés », *in* Corbin, D. *et al.* (éd.), *Mots possibles, Mots existants,* Villeneuve-d'Ascq, *Silexicales* 1, pp. 163-172.

Kerleroux, F. (2000), « Identification d'un procédé morphologique : la conversion », *Faits de Langue* 14, pp. 89-100.

Lecomte, E. (1997), « Tous les mots possibles en *-ure* existent-ils ? », *in* Corbin, D. *et al.* (éd.), *Mots possibles, Mots existants,* Villeneuve-d'Ascq, *Silexicales* 1, pp. 191-200.

Lesselingue, C. (en préparation), *Proposition pour un traitement différentiel des unités NN en français*. Thèse de doctorat.

Lieber, R. (1992), *Deconstructing Morphology: Word Formation in Syntactic Theory*, Chicago, University of Chicago Press.

Lovis, Ch. *et al*. (1998), « Medical dictionaries for patient encoding systems: a methodology », in Elsevier *Artificial Intelligence in Medecine* (14), pp. 201-214.

MacNamara, J. (1982), *Names for Things*, Cambridge (Mass.), MIT Press.

Martinet, J. (1985), « Variantes et homonymie affixales, le cas du français *dé-* », in *La linguistique*, pp. 239-250.

Mélis-Puchulu, A. (1991), « Les adjectifs dénominaux : des adjectifs de 'relation' », *Lexique* 10, pp. 33-60.

Namer, F. (2002) « Acquisition automatique de sens à partir d'opérations morphologiques en français », in *TALN 2002*, Nancy, pp. 235-244

Namer, F. (à paraître), « Productivité morphologique, représentativité et complexité de la base : le système MOQUÊTE », *in* G. Dal (éd.), *La productivité en questions et en expérimentations*, *Langue Française*, 137.

Namer, F. & Dal, G. (2000), « GéDériF : automatic generation and analysis of morphologically constructed lexical resources », in *LREC*, Athens, pp. 1447-1454.

Plag, I. (1999), *Morphological Productivity*, Berlin/New York, M. De Gruyter.

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

Roger, C. (2003), « Pour une individualité sémantique des affixes : rejet de la notion de paradigme de procédés morphologiques », *in* Fradin, B. *et al.* (éd.), *Les unités morphologiques,* Villeneuve-d'Ascq, *Silexicales* 3, pp. 179-187.

Spencer, A. (1991), *Morphological Theory*, Cambridge, Basil Blackwell.

Villoing, F. (2002), *Les mots composés [VN] en français : réflexions épistémologiques et propositions d'analyse*, thèse de doctorat, Paris 10.

Zwanenburg, W. (1992), « Compounding in French », *Rivista di linguistica* 4/1, pp. 221-240.

Zweigenbaum, P. *et al*. (2003), « Towards a Unified Medical Lexicon for French », in *MIE 2003*, Saint-Malo, pp. 415-420.