Un Lexique Génératif de référence pour le français

Fiammetta NAMER[1], Pierrette BOUILLON[2], Evelyne JACQUEY[3]
[1]Université Nancy2 et ATILF
fiammetta.namer@univ-nancy2.fr
[2] ISSCO
pierrette.bouillon@issco.unige.ch
[3]ATILF,
evelyne.jacquey@atilf.fr

Résumé. Cet article propose une approche originale visant la construction d'un lexique sémantique de référence sur le français. Sa principale caractéristique est de pouvoir s'appuyer sur les propriétés morphologiques des lexèmes. La méthode combine en effet des résultats d'analyse morphologique (Namer, 2002;2003), à partir de ressources lexicales de grande taille (nomenclatures du TLF) et des méthodologies d'acquisition d'information lexicale déjà éprouvées (Namer 2005; Sébillot 2002). Le format de représentation choisi, dans le cadre du Lexique Génératif, se distingue par ses propriétés d'expressivité et d'économie. Cette approche permet donc d'envisager la construction d'un lexique de référence sur le français caractérisé par une forte homogénéité tout en garantissant une couverture large, tant du point de vue de la nomenclature que du point de vue des contenus sémantiques. Une première validation de la méthode fournit une projection quantitative et qualitative des résultats attendus.

Abstract. This paper describes an original approach aiming at building a reference semantic lexicon for French. Its main characteristic is that of being able to rely on morphological properties. The method thus combines morphological analyses results (Namer 2002;2003;2005) from large scale lexical resources (i.e. TLF word lists) with already tested acquisition methodologies on lexical information (Sébillot, 2002). The representation format, within the Generative Lexicon framework, has been chosen for its expressiveness and economy features. So, this approach allows us to consider building a reference lexicon for French, which is fundamentally homogeneous as well as of a large coverage. A feasability study of the described method provides a projection of expected results, from both quantitative and qualitative points of view.

Mots-clés : acquisition lexicale, lexique de référence du français, modèle du lexique génératif, morphologie constructionnelle, corpus, sémantique.

Keywords: lexical acquisition, reference lexicon for French, generative lexicon model, word formation, corpora, semantics.

1 Introduction : objectifs

Cet article présente une méthodologie[1], dont l'objectif est de construire automatiquement un lexique du français dans le format du Lexique Génératif (Pustejovsky 1995), désormais LG. La construction de ce lexique exploite deux sources d'acquisition : des règles linguistiques fondées sur des contraintes imposées par la morphologie (désormais règles de construction de lexèmes : RCL, cf. (Aronoff 1994; Fradin 2003)), et des règles d'apprentissage, à partir de définitions du TLFi[2] ou/et à partir de corpus. Ces dernières s'appliquent aux lexèmes non construits, c'est-à-dire non accessibles aux RCL, et permettent, en cas d'ambiguïté non soluble par les règles, d'acquérir des informations spécifiques, ou d'enrichir les schémas sous-spécifiés. Par cette combinaison d'approches, nous espérons arriver à une meilleure cohérence globale du lexique dérivé. En effet, les RCL définissent des structures sémantiques générales qui s'appliquent de façon systématique à des classes de lexèmes construits similaires. C'est pourquoi elles sont traduisibles sous forme de modèles lexicaux dans le LG. Par exemple, tous les adjectifs déverbaux dénotent une propriété attendue, qui se manifeste sous la forme de l'activation potentielle du prédicat de base. Cette propriété commune s'observe quelles que soient les caractéristiques du verbe de base, comme l'illustrent les exemples sous (1a) (le verbe est agentif) et (1b) (le verbe est inaccusatif) :

(1) a. [[pull lavable]] = λy, $\exists$e, $\exists$x ($\Diamond$ laver'(e, x,y) & pull'(y))
b. [[marchandise périssable]] = λy, $\exists$e ($\Diamond$ périr'(e, y) & marchandise'(y))

Dans la suite, nous présentons cette méthodologie en détail. Nous la situons d'abord dans un cadre plus général, pour montrer sa spécificité. Nous l'illustrons ensuite par plusieurs exemples concrets qui montrent l'apport respectif des RCL et des ressources lexicales et textuelles. Enfin, nous évaluons quantitativement les résultats escomptés.

2 Cadre Théorique

Aujourd'hui, il existe différents lexiques sémantiques pour le français, plus ou moins librement disponibles, par exemple l'adaptation et/ou l'extension en français de ressources comme EuroWordnet (Vossem 2001). Cependant, aucun d'entre eux n'a réussi à rendre explicite l'ensemble des propriétés suivantes : les liens morphologiques entre les entrées lexicales (comme ceux tissés en (1) entre LAVABLE et LAVER, PÉRISSABLE et PÉRIR) ; d'autres relations permettant d'étoffer la structure hiérarchique du lexique (par exemple, le lien nom-verbe entre COUTEAU et sa fonction prototypique, *i.e.* COUPER) ; la polysémie systématique de certains lexèmes (cf. exemples (2)) ; les interactions syntaxe-sémantique (comme la relation mise en jeu, en (2a), entre les deux emplois de COULER et les fonctions grammaticales réalisées dans chaque cas). Or c'est l'acquisition de ces propriétés qui constitue la motivation fondamentale de notre approche.

(2) (a) le bateau coule versus les pirates coulent le bateau
COULER : est-il transitif/causatif ou intransitif/inchoatif ?

[1] développée dans le cadre d'une proposition de projet qui réunit les auteurs de cet article ainsi que C. Fabre (ERSS), P. Sébillot et V. Claveau (IRISA).

[2] TLFi : "Trésor de la Langue Française informatisée": version informatisée du 'Trésor de la langue Française', 16 volumes, cf. URL : http://atilf.atilf.fr/tlf.htm

(b) *commencer* à écrire/lire un livre versus *commencer* un livre
COMMENCER : sélectionne-t-il un objet direct nominal ou un syntagme verbal ?
(c) la salle a applaudi, la salle contient 200 personnes, la salle a été repeinte récemment
SALLE : désigne-t-il un collectif, un espace ou un objet physique ?
(d) *un chien* rapide versus *une route* rapide
RAPIDE : qualifie-t-il la nature (CHIEN) ou la fonction (ROUTE) du nom modifié ?
(e) un livre *rouge* versus un livre *intéressant*
LIVRE : dénote-t-il un objet physique ou de l'information ?

Son originalité tient de fait en deux points : pour extraire ces différentes informations, nous réutilisons un analyseur morphologique existant (DériF, cf. (Namer 2002, 2003, 2005)). Nous tirons ainsi parti des propriétés sémantiques des lexèmes reliés morphologiquement pour dériver un lexique cohérent et de grande taille à partir des traits morphologiques et sémantiques au moyen desquels l'analyseur Dérif annote automatiquement les lexèmes. Ensuite, le modèle choisi pour l'encodage des informations est le LG. Par rapport aux autres théories lexicales, LG présente différents attraits. Il se caractérise fondamentalement par un principe d'économie relayée par un système de types cohérent et des mécanismes génératifs qui factorisent l'information lexicale sémantique et rendent ainsi compte des cas d'ambiguïtés lexicales, comme celles illustrées sous (2). Dans LG, le sens est représenté sous la forme du produit de quatre rubriques, chacune remplissant un rôle particulier. Ces quatre rôles sémantiques, appelés FORMEL, CONSTITUTIF, AGENTIF, et TELIQUE, constituent la structure des qualia. Ils sont illustrés dans la Figure (1) pour le nom COUTEAU.

ROLE	**Fonction**	**Exemple : M = couteau**
FORMEL	place de M dans la taxinomie	x de type *artefact*
CONSTITUTIF	relations (partie-tout) entre M et ses constituants	Relation = Composant-assemblage , entre x (ie le couteau) et z (de type manche) et entre x et y (de type lame)
AGENTIF	conditions nécessaires (présupposées) à l'existence de M	Evénement (accomplissement) : fabriquer, faisant intervenir un agent u et le résultat x
TELIQUE	décrit la finalité de M	Evénement (accomplissement) : couper, faisant intervenir un agent u', l'instrument x et l'objet à couper : y

Figure 1 : Structure de qualia d'une entrée lexicale M dans le LG

Dans cette structure, chaque rôle définit un prédicat qui relie entre eux différents paramètres, et est caractérisé d'un point de vue événementiel, comme l'illustre la troisième colonne. Les paramètres manipulés par le prédicat sont typés. C'est aussi le cas de la structure de qualia, qui, sous la forme du Paradigme Lexical Conceptuel, organise le lexique en classes sémantiques. Toute structure de qualia peut ainsi être interprétée et reformulée dans différents modèles logiques (Pustejovsky 2001). Par exemple, la structure de qualia de COUTEAU peut être traduite dans la formule (3). Cette propriété générale rend le LG particulièrement adéquat pour le calcul logique (Pustejovsky 2001).

(3) λx [couteau'(x) : formal(x) = λx [artefact'(x)] $\wedge$ constitutif(x) = $\exists y,z$ [composant-assemblage'(y, x) $\wedge$ composant-assemblage'(z,x)] $\wedge$ agentif(x) = $\exists e,u$ [fabriquer'(e, u, x)] $\wedge$ telique(x) = $\lambda e'$ $\exists uy'$ [$\Diamond$ couper'(e', u', y, avec'(x))]

Dans le passé, d'autres projets se sont donné un objectif similaire (notamment Acquilex, cf. (Copestake et al. 1993), Simple (Busa et al. 2001) et Clips, cf. (Calzolari et al. 2003)). Cependant aucun d'entre eux n'a véritablement exploité les liens morphologiques pour en

extraire les informations sémantiques pertinentes à grande échelle, faute d'un analyseur morphologique adéquat. En prenant comme source l'analyse fournie par Derif, nous évitons deux pierres d'achoppement : d'abord, la représentation LG découle ici du procédé morphologique impliqué, ce qui la motive théoriquement. D'autre part, tous les lexèmes construits de la même manière reçoivent une représentation LG similaire, ce qui devrait assurer une meilleure cohérence globale de la ressource. Nous espérons ainsi deux retombées principales sur le plan théorique. D'une part, nous complétons l'apport de l'analyseur DériF en le couplant à une sémantique plus profonde. Rappelons que la contribution de la morphologie à la construction du sens lexical, que DériF formalise, n'est que partielle (cf section 3.3) : une RCL ne fournit en effet que les éléments fondamentaux de l'interprétation d'un lexème, qui seront spécifiés ultérieurement, notamment par le contexte d'utilisation (cf. (Aronoff 1980)). D'autre part, nous confirmons les hypothèses théoriques sur lesquelles repose le LG, et notamment nous répondons à la question suivante : LG fournit-il un cadre suffisant pour décrire les différents aspects de la sémantique lexicale et ceci, à grande échelle, pour différentes familles de lexèmes ? Idéalement, toutes les propriétés prédictibles par les RCL (par exemple : la base verbale des verbes préfixés par *dé-* , e.g. DÉCOUDRE, DÉFAIRE, dénote un *accomplissement*, dont le résultat désigne une propriété *réversible*, cf. (Amiot à paraître)) devraient pouvoir être représentées formellement au niveau de la structure des qualia. De cette structure, différentes informations peuvent ensuite être extraites, dont on a déjà montré l'intérêt sur le plan pratique. (Bouillon et al. 2000; Claveau et al. 2001; Claveau et al. 2003) ont notamment montré comment tirer parti des liens Nom-Verbe exprimés dans la qualia pour la recherche documentaire. Une question ouverte est de savoir si cette ressource sera plus utile que les précédentes basées sur le même formalisme. Nous pensons en tout cas que le fait de lier cette ressource à un analyseur morphologique existant la rend potentiellement plus apte à répondre à la question fondamentale de la créativité de la langue, DériF étant conçu pour l'analyse des lexèmes inconnus. Dans la suite, nous décrivons plus en détail la méthodologie d'acquisition.

3 Méthodologie d'acquisition du lexique

L'acquisition des entrées lexicales fait collaborer deux approches décrites dans ce qui suit. La première se fonde sur l'utilisation de connaissances morphologiques (section 3.1), la seconde sur l'emploi de méthodes d'apprentissage à partir de corpus (section 3.2). La combinaison des deux approches (section 3.3), enfin, confirme, infirme ou affine les contributions propres à chaque technique.

3.1 Acquisition d'entrée par règles morphologiques

Cette approche exploite les résultats fournis par l'analyseur morphologique DériF (Namer 2002, 2003, 2005), qui sont reformatés de manière à être en conformité avec les notations du LG, suivant l'extension du modèle proposé dans (Jacquey et al. à paraître; Namer et al. 2003, soumission).

3.1.1 Résultats de DériF

DériF produit la décomposition d'un lexème L muni d'une catégorie grammaticale, selon sa base morphologique B. Cette analyse s'accompagne d'annotations, reflétant les contraintes de la RCL appliquée, et pouvant porter sur L et B. Par exemple, la Figure2 illustre au moyen de

l'analyse de DESSOULER$_{VERBE}$ (construit sur SOUL$_{ADJ}$) le format dans lequel s'affiche le résultat de l'analyse par la RCL en *dé-* des verbes désadjectivaux. En dehors de l'analyse elle même (lignes 1 et 2), la règle attribue aux lexèmes reliés les traits suivants : l'adjectif doit être toujours **qualificatif** et décrire une propriété **transitoire** (ligne 3). Le verbe est soit **transitif causatif** ("le café salé dessoule Max") soit **intransitif résultatif** (ou anticausatif) ("Max dessoule") (ligne 4).

```
1  dessouler/VERBE ==> soul,ADJ/dé:prefixe
2  "(Supprimer - Faire perdre) le caractère soul"
3  soul/ADJ:(prédicatif,_,temporaire)
4  dessouler/VERBE:  (causatif,transitif,[cause,theme])||
                     (resultatif,intransitif,[theme])
```

Figure 2 : Analyse du verbe DESSOULER par DériF

3.1.2 Traduction des résultats dans le LG

En fonction des résultats produits par DériF, il est au mieux possible de générer automatiquement les deux entrées au format LG (celles du lexème analysé L et de sa base B) reliées morphologiquement par la règle ayant produit le résultat. Le niveau de spécification de chaque entrée dépend des informations produites lors de la phase d'analyse morphologique. Dans l'exemple pris plus haut, il est ainsi possible de prédire la distribution sous forme de rôles de qualia des étapes de l'enchaînement causal qui constituent l'événement complexe défini par le verbe DESSOULER (Figure 3a), à savoir la succession des situations suivantes : (1) l'individu y est soul (état initial, présupposé dans le rôle AGENTIF) ; (2) l'agent x dessoule l'individu y ou y dessoule (accomplissement, dans le rôle AGENTIF) ; (3) y n'est plus soul (état final, rôle FORMEL). En ce qui concerne l'adjectif SOUL, en tant que base de DESSOULER, la seule information inférable à partir de la règle de préfixation en *dé-* est qu'il s'agit d'une propriété identifiée d'un point de vue événementiel sous la forme d'un état transitoire e (ou stage level predicate, cf. (Carlson 1977)) affectant un individu y (cf. Figure 3b) (dans l'ontologie du rôle FORMEL, e caractérise SOUL comme sous-type du type 'état'). La valeur des autres traits (Structure Argumentale, Structure Evénementielle) constituant une entrée lexicale dans le LG est ensuite instanciée à partir du contenu de la Structure de qualia.

ROLE	**(V) dessouler**
FORMEL	not (soul' (e1 :état_trans,y :individu))
AGENTIF	soul'(e0 :etat_trans,y) ET dessouler_acte'(e2 :accompl., x :agent, y) OU soul'(e0 :etat_trans,y) ET dessouler_acte' (e2 :accompl., y)

ROLE	**(A) soul**
FORMEL	soul'(e :état_trans, y :patient)

Figure 3 : Décomposition des sens de DESSOULER (a) et de SOUL (b) sous forme de rôles de qualia, à partir de l'analyse par DériF

3.2 Acquisition par apprentissage sur corpus

En dehors des cas où les lexèmes simples sont des bases d'autres lexèmes construits, DériF ne fournit aucune information à traduire dans le format LG. Dans ces cas de silence, la méthodologie présentée s'appuie sur un apprentissage de corpus, et plus précisément sur un

corpus issu des données lexicographiques du TLFi. Une expérience préliminaire, effectuée sur la version XML catégorisée du TLFi, montre par exemple que les définitions de ce dictionnaire sont suffisamment régulières pour permettre de détecter les substantifs ayant une fonction prototypique, c'est-à-dire une facette télique dans leur contenu sémantique. En recherchant dans le corpus des définitions des expressions comme « servir, permettre, destiné à/au/aux/de », on repère près de 14% de substantifs ayant un emploi télique (4279 des 30544 substantifs du TLFi). Ceci constitue donc un premier résultat non négligeable. De plus, les expressions verbales ci-dessus s'accompagnent de substantifs (« appareil, organe, instrument ») qui permettent d'identifier des rôles formels (« balai = Ustensile$_{\text{FORMEL}}$ de ménage servant au **nettoyage**$_{\text{TELIQUE}}$ »). Ces substantifs servent à leur tour à la recherche de nouvelles expressions verbales, permettant ainsi de détecter de nouveaux prédicats téliques, et ainsi de suite… Cette première expérience montre donc l'exploitabilité du corpus des définitions catégorisées du TLFi dans le cadre de la construction automatique d'un lexique du français.

3.3 Croisement des deux approches

Le croisement des deux méthodes d'acquisition présentées *supra* peut servir à préciser certaines informations laissées sous-spécifiées par l'application des RCL. La sous-spécification a deux origines. Soit la morphologie ne dispose pas d'indices suffisants pour préciser un sens (cf section 3.3.1), soit le lexème construit est intrinsèquement ambigu (cf. section 3.3.2). Dans les deux cas, la contribution de l'approche par corpus doit servir à apporter l'information manquante.

3.3.1 Morphologie et sous-spécification : la préfixation en dé-

Considérons le cas des verbes en *dé-* sur base nominale. Deux types de sens sont possibles pour ces verbes. En effet, soit le nom de base décrit la localisation initiale de ce que dénote l'objet direct du verbe ; c'est ce que l'on observe avec DÉTERRER : « faire sortir qqch de terre ». Soit, au contraire, le nom de base décrit l'entité qui subit le changement de lieu ; c'est ce que l'on observe avec DÉSOSSER : « faire sortir les os de qqch/quelquepart ». L'analyse automatique par DériF est incapable de distinguer les deux cas de figure, qui se différencient sur la base de caractéristiques des noms de base qui sont d'ordre extralinguistique. Par conséquent, DériF fournit systématiquement deux définitions à chaque analyse de ce type de verbe. Ainsi, pour DÉTERRER, on obtient :

```
1  déterrer/VERBE ==> terre,NOM/dé:prefixe
2  "Faire sortir qqc de terre || Faire sortir la terre de qqc"
3  déterrer/VERBE:  (dynamique,trans. ,[cause,theme], causatif)
```

Figure 4 : Analyse de DETERRER par DériF

Cette analyse illustre un certain nombre de faits. Tout d'abord, aucune contrainte n'est généralisable pour la base. En effet, le nom sélectionné par la règle en *dé-* est soit concret (TERRE : DETERRER), soit abstrait (COURAGE : DÉCOURAGER). Par contre, la règle impose que le verbe résultant (DÉTERRER, DECOURAGER) soit transitif et décrive un accomplissement (plus précisément, un changement de localisation). En résumé, les contraintes liées à la formation de verbes dénominaux en *dé-* sont les suivantes : (1) le verbe est transitif, causatif, faisant intervenir une cause **x** et un thème **y**; (2) le procès désigne l'acte du causateur sur le thème (« x cause qqc à y »); (3) l'état final affecte soit **y** (c'est le cas avec DÉTERRER), soit

l'entité décrite par le nom de base (e.g. avec DÉCLOUER). Avec DÉTERRER c'est **y** qui est une entité délocalisée par rapport à son site[3] initial : TERRE (*Max déterre le coffre*). Par contre, avec DÉCLOUER (ex : *Max décloue le coffre*), **y** joue le rôle du site initial alors que CLOU est la cible (voir note précédente) délocalisée par rapport au site. La RCL est donc incapable de déterminer qui, du thème ou de la base d'un verbe quelconque préfixé par *dé-*, joue le rôle de cible (et de site), et donc, en conséquence, de définir le verbe construit de façon univoque par rapport à sa base. Voilà pourquoi deux définitions sont proposées pour DÉTERRER, comme l'indique la ligne 2 de la Figure 4. Face à ce type d'ambiguïté que des règles linguistiques ne peuvent pas résoudre, deux cas de figures peuvent se présenter : soit un seul des deux sens est attesté, soit les deux emplois existent, mais avec des fréquences différentes. Dans les deux cas, l'analyse de corpus et de définitions du TLFi permet, selon le cas, soit de lever l'ambiguité, soit de pondérer chacun des emplois. Avec 'DÉTERRER X', par exemple, l'interprétation « faire sortir la terre de X » est improbable : elle est absente des définitions du TLFi et des 100 premiers résultats renvoyés par Google.

3.3.2 Morphologie et exceptions : la suffixation en -oir

Alors que l'exemple précédent montre comment les définitions dictionnairiques et les connaissances textuelles précisent l'information lexicale que la morphologie ne sait qu'esquisser, voyons maintenant un cas inverse : l'analyse de corpus au service de la détection des exceptions à une RCL, à savoir la formation de noms déverbaux en *–oir*. A l'exception de quelques noms désignant le patient prototypique du verbe de base (TIROIR), ou son agent (CONSOLOIR [4]), la RCL *–oir* construit des noms faisant référence à des lieux ou des instruments [5]. La différence entre ces deux concepts est parfois ténue, dès lors qu'un objet possède la taille requise pour occuper les deux fonctions : c'est le cas de la majeure partie des noms déverbaux en *–oir* (ABREUVOIR, BALANÇOIRE, ÉGOUTTOIR désignent chacun l'objet qui sert à la fois d'instrument aidant au déroulement du procès décrit par le verbe, et de lieu où ce procès se déroule). D'autres noms en *–oir* ont la particularité d'être polyréférentiels : ils dénotent deux objets distincts, comme HACHOIR qui désigne soit un couteau soit une planchette, ou SOULOIR, qui en argot fait référence à un verre ou à un débit de boisson. Enfin, certains noms identifient clairement un lieu (DORTOIR, FUMOIR) ou un instrument (RATISSOIR, RASOIR). Comme le montre la Figure 5, l'analyse par DériF des noms déverbaux en *–oir* reflète la polysémie qui caractérise la plupart de ces noms. Presque rien, par contre, n'est prédictible pour le verbe de base : il peut être transitif (ABREUVOIR), ergatif (DESSOULOIR), inergatif (TROTTOIR), inaccusatif (MOUROIR). En bref, à quelques exceptions près (on relève CRÊCHOIR et VIVOTOIR sur la Toile, dont le verbe de base est analysable comme statif), la seule propriété verbale identifiable est 'dynamique'.

```
abreuvoir/NOM ==> abreuver,VERBE/oir:suffixe
```

[3] Les termes de cible et site sont empruntés à (Vandeloise 1986). D'autres linguistes du courant cognitiviste utilisent les termes de, respectivement, trajector et landmark (Langacker 1987) ou figure et ground (Talmy 1983).

[4] "Elle est mon refuge, mon consoloir", (Yahoo).

[5] Entre autres définitions des rôles thématiques, celle de (Fillmore 1968) dit qu'un instrument est "la force ou objet inanimé impliqué dans et à l'origine de l'événement" et un lieu désigne "la localisation ou l'orientation spatiale de l'événement"

```
"Instrument de abreuver || Lieu de abreuver"
abreuver/VERBE:     (dynamique,-,-,-)
```

Figure 5 : Analyse de ABREUVOIR par DériF

La représentation dans le LG des propriétés de ABREUVOIR prédites par la RCL *–oir* exprime donc les faits suivants: (1) le nom désigne une entité, qui possède un type pointé (instrument•lieu), mécanisme du LG pour exprimer les différentes facettes des lexèmes polysémiques ; (2) il est l'un des participants du prédicat (dynamique) potentiel décrit par le verbe de base et qui définit le rôle TELIQUE (c'est à dire la fonction prévue de l'entité décrite par le nom). Tout nom construit en *–oir* est codé suivant ces indications, qui révèlent l'ambiguïté supposée par défaut de ce type de noms. Cette hypothèse doit alors être vérifiée dans les corpus : le nom apparaît-il derrière la préposition "avec" ? dans un complément locatif ? En fonction de la réponse, soit la polysémie est confirmée, soit le codage est revu (instrument ou lieu seul) en fonction de l'échec à l'un ou l'autre des tests.

4 Evaluation

Nous sommes dès à présent en mesure de prévoir quantitativement et qualitativement une partie des résultats qui seront produits selon cette approche, à savoir les entrées lexicales générées à partir des analyses de DériF. Actuellement, au moins 35,5% (i.e. 35263/99445) des lexèmes du TLFi sont analysés comme construits par DériF, sous la forme de 45478 étapes constructionnelles[6]. Ce résultat correspond à l'activation d'environ 85 RCLs. Le pourcentage de 35,5% a deux motifs principaux : (1) toutes les RCL ne sont pas encore implémentées, (2) une grande partie des entrées lexicales du TLFi ne sont pas morphologiquement construites. La Figure 6 montre les règles les plus fréquemment appliquées.

Type d'opération de dérivation morphologique	Catégorie du Construit : Règle (Catégorie de la Base)
Suffixation	A :el, ique, if, al,eux, aire, ien, iste (N), **able**(V), N : **eur** (V), ie, **ité** (A), V :**ifier**, **iser**(N,A), re (V)
Préfixation	V :**en**, a (A,N), **é, dé** (A,N,V), pré (V) A :**in**, hyper,sub,non (A), sur,anti,sub,sous, mono, poly, auto (N)
Conversion	N ->V, **A->V,** V->N, A -> N

Figure 6 : Règles s'appliquant à plus de 80 lexèmes du TLFi

Parmi celles-ci, les règles en gras sont d'ores et déjà associées à des contraintes pour l'identification de leur entrée ou de leur sortie prototypique, à l'image de ce qui est déjà illustré par les Figures 2, 4, 5 (cf. *supra*) ainsi que la Figure 7, qui reprend ci-dessous les conditions de formation des adjectifs déverbaux en *–able*, ébauchées dans la section 1, cf. exemple (1).

```
lavable/ADJ ==> laver,VERBE/able:suffixe
  "Que l'on peut laver || PREP que l'on peut laver"
```

[6] Soit 39028 opérations de dérivation, et 6450 opérations de composition dite néoclassique (cf *infra*).

```
laver/VERBE: (dynamique,-,[-,theme], -)
lavable/ADJ :     (prédicatif, latent,-)
```

Figure 7 : Analyse de LAVABLE par DériF

La formation d'adjectifs déverbaux en *–able* illustre d'ailleurs un autre cas de collaboration nécessaire entre approche morphologique et recherche en corpus. En effet, la contrainte par défaut stipule que le nom recteur de l'adjectif s'identifie avec le patient du verbe de base de celui-ci (*laver un pull / un pull lavable*). Cependant, (Hathout et al. 2003) ont montré que selon l'adjectif (et donc son verbe de base) n'importe quel participant du prédicat pouvait occuper cette fonction : « *un poisson, une saison, un étang pêchable* ». Ce n'est donc qu'en retrouvant en corpus la construction verbale (« *pêcher un poisson, **dans** un étang, **pendant** une saison* ») que l'on pourra affiner l'entrée lexicale de l'adjectif et du verbe reliés par la RCLable. Pour finir, rappelons que l'avantage de DériF est qu'il s'applique à n'importe quel lexique, qui peut ainsi servir d'entrée à la méthodologie et accroître la taille de la ressource LG. Par exemple, la substitution d'un lexique spécialisé biomédical à la liste des nomenclatures du TLFi fait varier sensiblement les chiffres présentés dans la Figure 6 : ici, 59% des lexèmes (soit 17297 des 29273 entrées) sont analysés comme construits, essentiellement par composition néoclassique[7] (13237 des 21757 opérations morphologiques analysées). Une évaluation quantitative récente de DériF par rapport à un "Gold Standard" ((Namer et al. 2007)) a montré un score d'au moins 77,6% de bonnes analyses de la part du programme. Une autre expérience ((Namer 2007b),(Namer et al. 2007),(Deléger et al. 2007)) a, elle, prouvé que, appliqué aux lexiques spécialisés du domaine biomédical, les RCL de DériF sont facilement transposables dans d'autres langues; en particulier, des règles d'analyses des composés néoclassiques ont été traduites en anglais avec succès ((Deléger et al. 2007)).

5 Conclusion

Dans cet article, nous avons présenté une méthodologie pour dériver un LG. Son originalité repose surtout sur l'exploitation des propriétés morphologiques du lexique. Sur le plan pratique, cette approche nous permet de garantir la cohérence des informations dérivées ; sur le plan théorique, elle montre comment la morphologie peut collaborer avec d'autres connaissances (dictionnairiques et textuelles) pour dériver des représentations profondes du sens des mots. Nous ne connaissons pas d'autre tentative du même type pour exploiter ensemble l'apport de ces différentes disciplines. Les résultats attendus comprennent le lexique LG et une série d'outils pour extraire dynamiquement de nouvelles entrées, en particulier Dérif, des modèles pour convertir la sortie de Dérif en des modèles LG et des règles d'extraction pour le corpus/TLFi.

Références

AMIOT, D. (à paraître). La catégorie de la base dans la préfixation en *dé-*. *La raison morphologique. Hommage à la mémoire de Danielle Corbin*. B. Fradin. Amsterdam/Philadelphia, John Benjamins.
ARONOFF, M. (1980). Contextuals. *Language* **56**(4): 744-758.
ARONOFF, M. (1994). *Morphology by Itself*. Cambridge, MIT Press.

[7] Un nom ou adjectif composé néoclassique est formé par une règle de composition, et diffère à plusieurs égards des composés dits 'standard' ou 'ordinaires': leur sens, leur structure, les composants impliqués, les domaines textuels et/ou le registre de langue concernés, etc. cf. (Namer 2007a).

BOUILLON, P., C. FABRE, P. SÉBILLOT and L. JACQMIN (2000). Apprentissage de ressources lexicales pour l'extension de requêtes. *TAL* **41**(2): 367-393.
BUSA, F., N. CALZOLARI and A. LENCI (2001). Generative Lexicon and the Simple model, developing Semantic Resources for NLP. *the Language of Word Meaning*. P. Bouillon and F. Busa. Cambridge, CUP**:** 333-349.
CALZOLARI, N., F. BERTAGNA, A. LENCI and M. MONACHINI (2003). New perspectives for lexical web resource in the Semantic Web Scenario. *Generative Approaches to the Lexicon*, Geneva.
CARLSON, G. (1977). Reference to Kinds in English, University of California. **Ph.D. Dissertation**.
CLAVEAU, V., P. SÉBILLOT, P. BOUILLON and C. FABRE (2001). Acquérir des éléments du lexique génératif : quels résultats et à quels coûts ? *TAL* **42**(3): 729-753.
CLAVEAU, V., P. SÉBILLOT, C. FABRE and P. BOUILLON (2003). Learning Semantic Lexicons from a Part-of-Speech and Semantically Tagged Corpus Using Inductive Logic Programming. *Journal of Machine Learning Research* **4**: 493-525.
COPESTAKE, A., S. ANTONIO, T. BRISCOE and V. DE PAIVA (1993). The ACQUILEX LKB : an introduction. *Inheritance, defaults and the lexicon*. T. Briscoe, A. Copestake*et al*. Cambridge, CUP**:** 148-163.
DELÉGER, L., F. NAMER and P. ZWEIGENBAUM (2007). Analyse morphosémantique des composés savants : transposition du français à l'anglais. *TALN*, Toulouse.
FILLMORE, C. (1968). The case for case. *Universals in Linguistic Theory*. E. Bach and R. Harms. New-York, Holt, Rinehart, and Winston**:** 1-88.
FRADIN, B. (2003). *Nouvelles approches en morphologie*. Paris, Presses Universitaires de France.
HATHOUT, N., M. PLÉNAT and L. TANGUY (2003). Enquête sur les dérivés en *-able*. *Cahiers de Grammaire*. N. Hathout, M. Rochée*t al*. Toulouse, ERSS. **28:** 49-91.
JACQUEY, E. and F. NAMER (à paraître). Morphosémantique et modélisation : le cas des verbes dénominaux préfixés par *é-*. *Actes du Colloque "Le sens en linguistique" (25-27 mai 2003)*, Montréal.
LANGACKER, R. (1987). *Foundations of Cognitive Grammar*. Standford, Standford University Press.
NAMER, F. (2002). Acquisition automatique de sens à partir d'opérations morphologiques en français : études de cas. *Actes de Traitement Automatique du Langage Naturel (TALN) 2002*, Nancy, France, ATALA-ATILF.
NAMER, F. (2003). Automatiser l'analyse morpho-sémantique non affixale : le système DériF *Cahiers de Grammaire*. N. Hathout, M. Rochée*t al*. Toulouse, ERSS. **28:** 31-48.
NAMER, F. (2005). *La Morphologie Constructionnelle du Français et les Propriétés Sémantiques du Lexique - Mémoire présenté dans le cadre de l'habilitation à diriger des recherches*. UFR Sciences du Langage. Nancy, Université de Nancy2.
NAMER, F. (2007a). Composition néoclassique : est-on dans l' "hétéromorphosémie" ? *Morphologie à Toulouse - Actes du colloque international de Morphologie 4èmes Décembrettes*. N. Hathout and F. Montermini. München, Lincom Europa. **(LSTL 37):** 185-206.
NAMER, F. (2007b). Morphosémantique pour l'appariement de termes dans le vocabulaire médical : approche multilingue. *T.A.L.* **46**(2): 157-181.
NAMER, F. and R. BAUD (2007). Defining and relating biomedical terms : towards a cross-language morphosemantics-based system. *International Journal of Medical Informatics* **76**: 226-233.
NAMER, F. and E. JACQUEY (2003). Lexical Semantics and derivational morphology: the case of the popular *é-* prefixation in French *GL 2003 : 2nd International Workshop on Generative Approaches to the Lexicon (May, 15-17 2003)*, Geneva.
NAMER, F. and E. JACQUEY (à paraître). Word Formation Rules and the Generative Lexicon: Representing noun-to-verb versus verb-to-noun Conversion. *Generative Lexicon Book*. P. Bouillon, H. Isahara*et al*. Dordrecht, Kluwer.
PUSTEJOVSKY, J. (1995). *The Generative Lexicon*. Cambridge, MA, MIT Press.
PUSTEJOVSKY, J. (2001). Type Construction and the Logic of Concepts. *The Syntax of Word Meanings*. P. Bouillon and F. Busa. Cambridge, Cambridge University Press**:** 91-123.
SÉBILLOT, P. (2002). *Apprentissage sur corpus de relations lexicales sémantiques - La linguistique et l'apprentissage au service d'applications du traitement automatique des langues. Mémoire présenté dans le cadre de l'habilitation à diriger des recherches.* Rennes, Université de Rennes 1.
TALMY, L. (1983). How language structures space. *Spatial orientation, Theory, Research, and Applications*. H. Pick and L. Acredolo. New York, Plenum**:** 225-282.
VANDELOISE, C. (1986). *L'espace en français: sémantique des prépositions spatiales* Paris, Les éditions du Seuil.
VOSSEM, P. (2001). Condensed Meaning in EuroWordnet. *The language of Word Meaning*. P. Bouillon and F. Busa. Cambridge, CUP**:** 363-383.