TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Construction d'un lexique des adjectifs dénominaux

Jana Strnadová[1,2] & Benoît Sagot[3]
(1) LLF, CNRS & Univ. Paris 7, 5 rue Thomas Mann, 75205 Paris Cedex 13, France
(2) Univerzita Karlova, Filozofická Fakulta, nám. J. Palacha 2, 116 38 Prague, Rép. Tchèque
(3) Alpage, INRIA & Univ. Paris 7, Rocquencourt, BP 105, 78153 Le Chesnay Cedex, France
strnadjana13@gmail.com, benoit.sagot@inria.fr

Résumé. Après une brève analyse linguistique des adjectifs dénominaux en français, nous décrivons le processus automatique que nous avons mis en place à partir de lexiques et de corpus volumineux pour construire un lexique d'adjectifs dénominaux dérivés de manière régulière. Nous estimons à la fois la précision et la couverture du lexique dérivationnel obtenu. À terme, ce lexique librement disponible aura été validé manuellement et contiendra également les adjectifs dénominaux à base supplétive.

Abstract. After a brief linguistic analysis of French denominal adjectives, we describe the automatic technique based on large-scale lexicons and corpora that we developed for building a lexicon of regular denominal adjectives. We evaluate both the precision and coverage of the resulting derivational lexicon. This freely available lexicon should eventually be fully manually validated and contain denominal adjectives with a suppletive base.

Mots-clés : Adjectifs dénominaux, dérivation morphologique, lexique dérivationnel.

Keywords: Denominal adjectives, morphological derivation, derivational lexicon.

1 Introduction

La morphologie constructionnelle est la partie de la morphologie qui permet de créer des lexèmes à partir d'autres lexèmes, et d'augmenter ainsi le lexique d'une langue. Il s'agit donc d'un procédé de structuration du lexique qui relie des lexèmes entre eux et un procédé d'extension du lexique. À ce double titre, la disponibilité de ressources lexicales dérivationnelles est cruciale à la fois pour la description linguistique et pour le traitement automatique des langues (TAL). En linguistique, les ressources lexicales enrichies de liens dérivationnels permettent une approche systémique du lexique et des procédés productifs de construction de lexèmes. En traitement automatique des langues, de nombreux travaux antérieurs ont ainsi montré l'importance de la prise en compte de la morphologie constructionnelle pour l'analyse des mots inconnus (néologismes, termes techniques) et l'enrichissement de ressources lexicales (Dal *et al.*, 1999; Hathout & Tanguy, 2005), l'analyse syntaxique (Bourigault & Frérot, 2004), mais également dans des contextes plus applicatifs tels que les systèmes de question-réponse (Bernhard *et al.*, 2011) ou de traduction automatique (Cartoni, 2009).

Pour le français, les noms déverbaux ont été étudiés de façon systématique et font l'objet d'une ressource lexicale librement disponible, VerbAction (Tanguy & Hathout, 2002). Cette ressource a été développée en partie grâce au système Webaffix (Hathout & Tanguy, 2005), qui utilise le Web comme un corpus pour y détecter des lexèmes construits néologiques. Nous nous penchons ici sur les adjectifs dénominaux, avec pour objectif de construire à terme une ressource du même type que VerbAction, mais qui s'en distinguera entre autres par des entrées plus détaillées (définitions, procédé morphologique dérivationnel détaillé…) et par la prise en compte de la dérivation non régulière (*école* vs *scolaire*, cf. partie 2).

Dans cet article, nous décrivons la première phase des travaux entrepris en ce sens. Nous nous sommes pour l'instant concentrés sur les adjectifs dérivés à partir d'une base nominale par affixation régulière, tout en prenant en compte les nombreuses variantes possibles selon les affixes et les bases. Après une étude linguistique préliminaire (section 2), nous décrivons nos expériences destinées à produire des liens de dérivation formels entre entrées lexicales connues de lexiques de référence (section 3.2) puis entre noms connus et dérivés adjectivaux inconnus (section 3.3). Nous estimons enfin la précision et la couverture des résultats obtenus (section 4). La ressource obtenue est librement disponible sous licence LGPL-LR comme complément au lexique Le*fff* (cf. plus bas).

2 Adjectifs dénominaux

Notre travail repose sur la notion de lexème telle qu'introduite par Matthews (1974). Le lexème est une unité abstraite multiplanaire, définie par sa forme sur les plans phonologique et graphémique, par sa signification sur le plan sémantique et par sa catégorie sur le plan *syntactique* (Fradin, 2003). Les règles morphologiques de construction de lexèmes doivent ensuite donner une spécification pour chacun de ces plans.

La suffixation constitue le principal procédé morphologique permettant de construire les adjectifs dénominaux [1]. Les règles morphologiques de construction de lexèmes formant des adjectifs dénominaux spécifient les différents plans du lexème base et du lexème dérivé. Pour le plan graphémique, la règle spécifie la forme graphique du suffixe qui doit être ajouté. Pour le plan syntactique, les règles indiquent que le lexème base est un nom et que le lexème dérivé est un adjectif. Enfin, l'instruction sémantique portée par le procédé dérivationnel dépend en partie du choix du suffixe. La ressource lexicale étant basée sur le code écrit, seul le plan graphémique va être détaillé dans ce qui suit et les bases et les suffixes vont être considérés comme des chaînes de caractères.

Du point de vue formel, les adjectifs dénominaux français sont construits principalement par les suffixes suivants : *-aire* (*alimentaire*), *-al* (*artisanal*), *-el* (*nutritionnel*), *-esque* (*livresque*), *-eux* (*mousseux*), *-ien* (*rotulien*), *-ier* (*sourcilier*), *-ique* (*folklorique*) et *-u* (*barbu*). Il ne s'agit pas d'une liste exhaustive, car d'autres suffixes peuvent dériver des adjectifs à partir de noms. Cependant, ils posent des contraintes spécifiques sur les noms bases (par exemple les suffixes *-ais*, *-ain*, *-an*, *-ois* apparaissent presque exclusivement avec les bases toponymes : *islandais, africain, castillan, lillois*) ou sont plus marginaux et ils n'apparaissent que dans quelques lexèmes (*-ard* : *campagnard*). Dans la situation idéalisée ou canonique, au sens de Corbett (2010), le suffixe serait simplement concatené à la base nominale. Néanmoins, en français, ce qui précède le suffixe ne correspond pas toujours à un nom base existant dans le lexique. Plusieurs cas de figure doivent être distingués.

Pour la construction de certains adjectifs, l'addition d'un suffixe peut entraîner des modifications (graphémiques) de la base. Il s'agit notamment du dédoublement de la consonne finale (*poisson* > *poissonn+eux*), de la suppression de la voyelle finale (*université* > *universit+aire*), de l'alternance de deux graphèmes (*adjectif* > *adjectiv+al)* ou de la combinaison de plusieurs changements (*muscle* > °*muscl+aire* > *muscul+aire*). Comme ces cas présentent une certaine régularité au sein du système de la dérivation adjectivale en français, il est possible de les décrire en spécifiant des règles allomorphiques [2] pour les différents types de bases qui subissent ces modifications.

Afin de pouvoir décrire tout le système dérivationnel des adjectifs dénominaux en français, il faut résoudre un autre problème qui se présente sous forme des adjectifs qui n'ont pas été construits en français, mais qui ont été soit hérités du latin, soit empruntés au latin, au grec ou à une autre langue. Dans ce cas, la base ressemble formellement à un nom du français, mais les deux formes sont trop éloignées pour qu'elles puissent être reliées par une règle (*moine* > *monacal, temps* > *temporel*). Le phénomène se limite à des lexèmes particuliers ou à des petites séries de lexèmes et il est donc difficile de le décrire par des règles générales. Enfin, il existe des adjectifs dont les bases ne sont pas corrélées formellement à un nom du français, même si une relation sémantique existe (*jeu - ludique, oiseau - aviaire*). Ces bases dites supplétives ou non autonomes (Corbin, 1985) sont assez nombreuses et le domaine de la dérivation adjectivale en abonde. Les deux derniers cas sont assez proches et il est souvent difficile de distinguer les exemples qui en relèvent. Dans les deux cas, on reconnaît les suffixes mais les bases n'apparaissent jamais comme des noms autonomes. Seule la similarité phonémique ou graphémique semble pouvoir les différencier, ainsi que l'origine étymologique commune entre la base utilisée pour la dérivation et le nom associé (*tempus/temporis* > *temps ; tempor-*).

3 Extraction de liens de dérivation entre noms et adjectifs dénominaux

Notre objectif est de construire une base d'adjectifs dénominaux associés à leur base nominale, que ces adjectifs soient déjà connus des ressources lexicales existantes ou non — ce qui est également un moyen à la fois de construire de nouvelles entrées candidates à ajouter au lexique et/ou d'extraire des termes techniques à partir de corpus spécialisés. Nous sommes partis de deux ressources lexicales : (1) le Le*fff* (Sagot, 2010), qui repose sur le formalisme lexical Alexina, dont la couche morphologique permet de représenter les opérations de dérivation applicables au sein d'une classe flexionnelle donnée (cf. ci-dessous) ; (2) Morphalou (Romary *et al.*, 2004), que

1. Les adjectifs dénominaux peuvent aussi être construits par conversion, un procédé morphologique fondé sur l'identité de la forme du lexème base et du lexème converti. Toutefois, ce phénomène sort du cadre du travail décrit ici.

2. Cette étude se situant au plan graphémique, la description développée ici ne correspond pas à ce qui est souvent considéré comme allomorphie en linguistique où l'intérêt porte sur le plan phonologique. Il est également possible de parler de règles de *sandhi* ou de règles morpho-phono-graphémiques. Tous ces termes véhiculent l'idée de certains changements subis par le nom base lors de l'adjonction du suffixe.

```
<table name="nc-2m" rads="..*[^sxz]">
    <form suffix="" tag="ms"/>
    <form suffix="s" tag="mp"/>
    <derivation name="adj_dénominal_en_-iel" suffix="∂iel" table="adj-l4"/>
    <derivation name="adj_dénominal_en_-iel_variante" suffix="∂∂iel" table="adj-l4" rads=".*(ce|eur)"/>
    <derivation name="adj_dénominal_en_-uel" suffix="∂uel" table="adj-l4" except=".*u"/>
    <derivation name="adj_dénominal_en_-al" suffix="∂al" table="adj-al4"/>
    <derivation name="adj_dénominal_en_-al_variante_nn" suffix="nal" table="adj-al4" rads=".*n"/>
    <derivation name="adj_dénominal_en_-al_variante" suffix="∂∂al" table="adj-al4" rads=".*(ion|eur)"/>
    ...
</table>
<sandhi source="e_∂" target="_"/>
<sandhi source="ce_∂∂" target="t_"/>
```

FIGURE 1 – Extrait de la table `nc-2m` des noms masculins à pluriel en *-s* et exemples de règles de *sandhi*. On note les contraintes sur la base : ainsi, la variante *-nal* du suffixe *-al* n'est possible que sur des bases en *-n*, redoublant ainsi cette consonne. Le graphème factice *∂* est utilisé dans des règles de sandhi simulant les opérations à effectuer sur la base avant de lui ajouter l'affixe dérivationnel, comme illustré par la règle qui efface le *-e* final lorsqu'il n'y a qu'un symbole *∂* (*ferme* > *fermier*), ou la suivante, qui transforme le *-c* en fin de base (éventuellement après élimination d'un *-e* final) en *-t* (*séquence* > *séquentiel*) lorsque l'affixe de dérivation est à deux *∂*.

nous avons converti automatiquement dans le formalisme Alexina, en faisant en sorte que les classes flexionnelles utilisées par cette version convertie du lexique Morphalou soient les mêmes que celles du Le*fff*, ce qui permet de le fléchir avec la même description morphologique que le Le*fff*. À partir de ces ressources, l'idée générale est de générer un grand nombre de candidats dérivés à l'aide de règles de dérivation décrites manuellement, puis de chercher parmi ces candidats d'une part ceux qui sont connus des lexiques de départ et d'autre part ceux dont les formes sont bien attestées dans tel ou tel gros corpus tout en n'étant pas couvertes par les lexiques de départ. Nous avons fait le choix de développer manuellement les règles de dérivation. En effet, une approche automatique (non supervisée) ne garantit ni la couverture des cas rares ou des variantes spécifiques ni la qualité des règles extraites. En effet, ces règles peuvent être des règles de dérivation valides (dont l'orientation sera du reste délicate à déterminer), mais elles peuvent également relier différents dérivés d'une même base (*-ation* – *-able*). À l'inverse, une approche manuelle permet des descriptions fines, qui prennent en compte les procédés peu fréquents et leurs différentes variantes, bien que le travail doive être répété pour chaque nouvelle langue.

3.1 Description formalisée de la dérivation nom-adjectif

Nous avons donc tout d'abord ajouté manuellement à chaque classe de flexion nominale de la description morphologique du Le*fff* l'ensemble des opérations de dérivation listées à la section 2 (cf. figure 1). Pour cela, nous utilisons différentes caractéristiques du formalisme morphologique d'Alexina, et notamment les suivantes.

- La possibilité de contraindre une base à correspondre et/ou à ne pas correspondre à des motifs réguliers (par exemple, contraindre le rajout de *-naire* aux bases se terminant en *-n*) ;
- Les opérations morpho-phono-graphémiques (dites règles de *sandhi*) qui permettent de définir des transformations appliquées à la frontière entre base et affixe. Ces opérations sont de deux ordres :
  - d'une part des règles dédiées indiquant les opérations à effectuer sur la base avant de lui ajouter l'affixe dérivationnel, tel que la suppression du *e* final (*muscle* > *muscl*) ; dans certains cas, une opération de dérivation morphologique est dédoublée en différentes variantes, qui sélectionnent des affixes ou des règles de sandhi différentes : par exemple, il peut y avoir insertion d'un *-t-* au cours de la dérivation en *-ique* sur bases en *-s(e)* et *-ma* (*phrase* > *phrastique*) ou de *-at-* pour les bases en *-m(e)* (*programme* > *programmatique*).
  - d'autre part des « vraies » règles de sandhi, déjà présentes dans la description ou rajoutées pour l'occasion, telles que *-cl* + *aire$* > *-cul* + *aire* (*muscl(e)* + *-aire* > *musculaire*).

Au total, nous avons rajouté 596 règles de dérivation réparties dans 29 tables nominales, couvrant ainsi les adjectifs dérivés sur base nominale par différentes variantes de la suffixation en *-el/-iel/-uel*, *-al/-ial*, *-aire/-uaire*, *-ique/-tique/-atique*, *-esque*, *-u*, *-er/-ier* et *-eux/-ieux*. Chacun de ces affixes est donc associé à des contraintes sur les bases admissibles, et le résultat de cette affixation peut être lui-même modifié par les règles de sandhi.

3.2 Construction de relations dérivationnelles entre entrées du lexique

L'union des entrées nominales du Le*fff* et de Morphalou comporte 65 651 lemmes (forme citationnelle et classe flexionnelle), qui produisent par ces règles un total de 886 526 couples (nom base, adjectif dérivé) candidats. Parmi ces couples, 3 293 ont un adjectif dérivé qui est lui aussi du Le*fff* ou de Morphalou. Ils concernent 2 687 adjectifs distincts, un même adjectif pouvant être obtenu par dérivation à partir de plusieurs bases (correctes ou non). Ces couples sont donc des relations dérivationnelles candidates entre noms et adjectifs déjà connus du lexique.

Type de couple (nom base, adjectif dérivé)	couples (base, adj. dér.)	adj. distincts
Candidats produits	**886 526**	**844 519**
1. dont retenus car l'adjectif est connu du lexique (Le*fff* +Morphalou)	**3 293**	**2 687**
2. dont retenus après confrontation au corpus (adjectif soit inconnu du lexique)		
avant filtrage	12 140	10 064
après filtrage des candidats orthographiquement proches de mots connus	11 463	8 317
après filtrage additionnel des couples à nom ou adjectif trop court	**8 736**	**7 449**
Total des candidats retenus	***12 029***	***9 692***

TABLE 1 – Résultats quantitatifs de l'extraction de couples (base nominale, adjectif dérivé).

3.3 Identification en corpus d'adjectifs inconnus du lexique dérivés de noms connus

Un certain nombre des 883 233 couples (nom base, adjectif dérivé) produits précédemment et dont l'adjectif est inconnu du Le*fff* comme de Morphalou sont corrects. Pour les identifier automatiquement, nous avons confronté les formes fléchies des adjectifs concernés l'union de trois corpus tous volumineux mais de natures différentes :

– **ER** : le corpus de l'Est Républicain (presse quotidienne régionale, tout ou partie des années 1999, 2002 et 2003), librement disponible sur le site du CNRTL (37,5 millions d'occurrences, 330 000 tokens distincts) ;
– **frwiki** : l'ensemble de la version française de l'encyclopédie libre Wikipedia, que nous avons transformé en un format texte utilisable dans nos outils (232 millions d'occurrences, 2,6 millions de tokens distincts) ;
– **g1g** : la collection de 1-grammes (mots associés à leur fréquence) distribuée par Google à partir des ouvrages scannés puis numérisés (reconnaissance optique des caractères) par l'entreprise, en se restreignant aux occurrences issues de livres datés postérieurs à 2000 (4,6 milliards d'occurrences, 540 000 tokens distincts environ).

Pour choisir les plus plausibles des adjectifs dérivés candidats restant, nous avons appliqué un processus itératif qui ajoute à chaque itération un nouvel adjectif au lexique, initialement l'union du Le*fff* et de Morphalou. À chaque itération i du processus, chaque adjectif dérivé a candidat reçoit un score $s_i(a)$ définit comme suit. Soient $a_1, \ldots, a_{n(a)}$ les formes fléchies de a, et $occ(a_1), \ldots, occ(a_n)$ leur nombre d'occurrences en corpus. On définit par ailleurs une fonction L_i d'appartenance au lexique tel que pour une forme f on a $L_i(f) = 1$ si f est dans notre lexique après la $i-1$-ième itération, et $L_i(f) = -1$ sinon. On pose alors $s_i(a) = -\sum_{k=1}^{n(a)} L_i(a_k) occ(a_k)$. Ceci défavorise les candidats dont un maximum de formes fléchies sont fréquentes en corpus mais inconnues du lexique, et défavorise ceux dont certaines formes fléchies sont déjà connues du lexique, surtout si elles sont fréquentes. Le mieux classé des adjectifs dérivés candidats est alors ajouté au lexique : ses formes fléchies sont désormais considérées comme connues. Et on passe à l'itération suivante, sauf si plus aucun candidat n'obtient de score strictement positif. Nous obtenons ainsi 10 064 adjectifs dérivés, qui sont tous par construction des adjectifs inconnus et du Le*fff* et de Morphalou. Nous avons conservé la ou les bases nominales des candidats dont ils proviennent, soit un total de 12 140 couples (nom base, adjectif dérivé) sur les 883 233 concernés.

À ce stade, nous avons constaté qu'une proportion importante de ces 12 140 couples étaient proposés en raison d'erreurs orthographiques dans le corpus, et notamment de problèmes d'accentuation (p.ex. *extremal* pour *extrémal*, *douannier* pour *douanier*). Nous avons donc utilisé le correcteur orthographique `sxspell`, intégré à la chaîne de traitements de surface SXPipe (Sagot & Boullier, 2008), pour proposer une correction orthographique pondérée pour chacun de ces adjectifs. Nous avons fixé empiriquement un seuil sur ce poids (50), et éliminé tous les couples dont l'adjectif peut être corrigé à un coût qui lui est inférieur en un mot connu du Le*fff*, lexique sur lequel repose `sxspell`, ne conservant ainsi que 11 463 couples. Enfin, comme mentionné par (Hathout & Tanguy, 2005), les dérivés proposés à partir de bases très courtes sont souvent erronés. Nous avons ainsi éliminé les couples dont la base fait 3 lettres ou moins (ainsi (*bru*, *bruel*) proposé en raison d'occurrences erronnées de *bruel* sans majuscule) [3]. Nous avons également éliminé les couples dont l'adjectif dérivé fait 6 lettres ou moins. Au final, nous avons ainsi retenu 8 736 couples candidats couvrant 7 449 adjectifs dérivés distincts inconnus du lexique.

Les résultats obtenus automatiquement aux deux étapes décrites dans cette section sont indiqués à la table 1.

4 Évaluation des résultats

Nous avons effectué une double évaluation de ces résultats. Pour en évaluer la précision, nous en avons validé manuellement une sous-partie choisie arbitrairement (section 4.1). Pour en évaluer la couverture, nous avons calculé une estimation à partir de l'analyse d'articles du Wiktionnaire, dictionnaire libre du français (section 4.2).

3. Notons que cette heuristique nous fait perdre un nombre très faible mais non nul de couples valides, comme (*pouillu*, *pou*) ou (*sonique*, *son*), mais élimine avantageusement (*pubien*, *pub*) ou encore (*motu*, *mot*), rendu plausible par son hypothétique pluriel *motus*...

Étiquette d'évaluation	**Couples dont l'adjectif est : connu du lexique**	**inconnu**
total (adj. en g-*)*	*156*	*265*
OK	121 (78 %)	136 (51 %)
OtherDer	35 (22 %)	38 (14 %)
NoDer	0	0 (0 %)
NO	–	91 (34 %)

TABLE 2 – Évaluation des couples retenus commençant par *g*. La précision de la seule suggestion de nouveaux adjectifs à ajouter au lexique, en ignorant toute information dérivationnelle, est de 70 %.

Étiquette d'évaluation	**Couples extraits du Wiktionnaire**
total (adjectifs en g- *ou* h-*)*	*86*
Reg	31 (36 %)
Suppl	20 (23 %)
OtherDer	31 (36 %)
Err	4

TABLE 3 – Répartition des couples d'AdjNWiktionnaire, extraits à des fins d'évaluation. Les deux premières étiquettes recouvrent les deux types d'adjectifs dérivés sur base nominale (dérivation régulière et sur base supplétive).

4.1 Évaluation de la précision

Nous avons choisi arbitrairement d'évaluer l'ensemble des couples candidats dont l'adjectif dérivé (et donc, par construction, la base nominale) commence par la lettre *g*. Sont concernés 156 couples dont l'adjectif est connu du lexique, pour un total de 116 adjectifs distincts, et 265 couples dont l'adjectif est inconnu du lexique, pour un total de 231 adjectifs distincts. Nous les avons tous passés en revue manuellement, et les avons classés au moyen des étiquettes suivantes, dont seules les trois premières concernent les 156 premiers couples :

– **OK** : le lemme adjectival est valide, et le lien dérivationnel également (ex. : *gestionnel/gestion*) ;
– **OtherDer** : le lemme adjectival est valide, il s'agit bien d'un adjectif dérivé mais à partir d'une base qui n'est pas le nom du couple (ex. : *gazeux/gaze*) ;
– **NoDer** : le lemme adjectival est valide, il ne s'agit pas d'un adjectif dérivé (ex. : *gotique/go*) ;
– **NO** : le lemme adjectival proposé est incorrect (ex. : *gabarru/gabarre*).

L'évaluation a été réalisée par deux validateurs qui ont chacun validé deux tiers des couples candidats de chaque type. L'accord inter-validateurs sur les couples dont l'adjectif est connu du lexique est de 88 %. Pour les couples dont l'adjectif est inconnu du lexique, l'accord est légèrement moins bon (85 %). Une annotation consensuelle a alors été attribuée à tous les couples pour lesquels il y avait initialement désaccord, afin de produire une annotation unique sur l'ensemble des couples évalués. Les résultats sont regroupés à la table 2. Concernant les couples dont l'adjectif est connu du lexique, les 22 % d'erreurs sont presque tous issus de couples dont l'adjectif est dérivé d'une autre base que le nom proposé (ex. : *galeux/gala*). De tels couples ne pourraient être écartés qu'au moyen d'outils complémentaires, tels que des modèles distributionnels qui garantiraient une certaine proximité sémantique entre base et dérivé. Par ailleurs, le taux de 51 % obtenu pour les couples dont l'adjectif est seulement trouvé en corpus reste tout à fait satisfaisant au regard du taux de 78 % obtenu pour ceux dont l'adjectif est connu du lexique, surtout si l'on prend en compte le fait que 14 % des couples proposés grâce au corpus (38 des 91 étiquetés "NO") proviennent d'erreurs orthographiques dans le corpus. Enfin, si l'on met de côté les liens de dérivation, plus de 70 % des adjectifs appartenant à des couples proposés mais inconnus du lexique sont valides.

4.2 Évaluation de la couverture

Il n'est jamais facile d'évaluer *in abstracto* la couverture d'une ressource lexicale. C'est d'autant plus vrai dans notre cas qu'il n'est pas toujours facile, comme expliqué à la section 2, de délimiter de façon nette les adjectifs dérivés d'une base nominale (fût-elle supplétive) et ceux qui sont en relation uniquement sémantique avec un nom. De plus, la couverture de notre ressource ne peut être estimée directement par le biais d'un corpus, puisqu'il n'existe pas de corpus indiquant pour chaque occurrence d'un lexème dérivé la base dont il est issu.

Nous avons donc estimé la couverture de notre ressource en recourant à une ressource externe incomplète composée de couples (adjectif, nom) que nous avons extraits à partir du Wiktionnaire comme suit : nous avons simplement construit un couple (adjectif, nom) pour toute entrée adjectivale du Wiktionnaire dont la définition suit un certain nombre de motifs caractéristiques tels que *Qui appartient à NOM*, *Relatif à/aux NOM*, etc. Cette ressource, que nous nommerons AdjNWiktionnaire, contient 1 433 couples (associés à leur définition). L'idée sous-jacente à l'utilisation d'AdjNWiktionnaire pour évaluer la couverture de notre ressource est que la forte incomplète d'AdjNWiktionnaire est distribuée aléatoirement. Autrement dit, si l'on ne conserve d'AdjNWiktionnaire que les couples qui sont dans le champ de notre étude, c'est-à-dire que l'adjectif est bien un dérivé du nom (hors bases supplétives), on peut évaluer par rapport à eux la couverture de notre ressource. Les autres couples d'AdjNWiktionnaire se répartissent en différentes catégories, dont on peut également tirer une estimation de la fréquence relative entre dérivation régulière (celle traitée ici) et dérivation sur base supplétive, pour peu que l'on ait annoté manuelle-

ment un nombre significatif d'entrées d'AdjNWiktionnaire. C'est ce que nous avons fait pour les 36 entrées dont l'adjectif commence par *g* (36 couples) ou *h* (50 couples), en utilisant les étiquettes suivantes :
– **Reg** : l'adjectif est effectivement dérivé du nom, et notre méthode devrait avoir produit ce couple ;
– **Suppl** : l'adjectif est effectivement dérivé du nom, mais à partir d'une base supplétive (*oreille* > *auriculaire*), et notre méthode ne peut donc pas avoir produit ce couple ;
– **OtherDer** : il y a une relation de dérivation entre l'adjectif et le nom, mais d'un autre type ;
– **Err** : le couple est issu d'une erreur dans la construction de AdjNWiktionnaire.
Les résultats de cette petite annotation sont donnés à la table 3.

Comme expliqué plus haut, la couverture de notre ressource peut être estimée en calculant sa couverture sur les cas étiquetés "Reg", seuls cas que nous cherchons à traiter dans ce travail. Sur les 31 couples concernés, 24 ont été retenus parmi les 12 029 que contient notre ressource, soit un peu plus des trois quarts. Les 7 couples restant ont une base nominale inconnue du L*efff* comme de Morphalou, et ne pouvaient donc pas être construits. Mais parmi ceux dont le nom est connu, tous ont donc été proposés par notre méthode et sont présents dans la ressource finale. C'est donc une indication satisfaisante du taux de couverture de notre approche.

5 Conclusion et perspectives

La première étape de notre projet de développement d'une ressource lexicale dérivationnelle sur les adjectifs dénominaux du français nous a permis de construire un lexique préliminaire comportant plus de 12 000 couples (adjectif dérivé, base nominale). Dans un premier temps, nous comptons valider semi-automatiquement l'ensemble de ces couples, qui relèvent tous de la dérivation régulière. Dans un second temps, et grâce à des moyens complémentaires comme des lexiques bilingues français-latin, une modélisation partielle de l'évolution diachronique du lexique du français ou encore des modèles sémantiques distributionnels, nous complèterons notre ressource par des couples (adjectif dérivé, base nominale) relevant de la dérivation à base supplétive, dont nous avons pu estimer qu'elle concernait une petite moitié des adjectifs dénominaux (20 sur 51). Enfin, nous espérons montrer à travers des applications concrètes, tant en linguistique qu'en TAL, qu'une telle ressource peut être très utile.

Références

BERNHARD D., CARTONI B. & TRIBOUT D. (2011). Évaluer la pertinence de la morphologie constructionnelle dans les systèmes de question-réponse. In *Actes de la Conférence TALN*. À paraître.

BOURIGAULT D. & FRÉROT C. (2004). Ambiguïté de rattachement prépositionnel : introduction de ressources exogènes de sous-catégorisation dans un analyseur syntaxique de corpus endogène. In *Actes de la Conférence TALN*, p. 81–90, Fès, Maroc.

CARTONI B. (2009). Lexical morphology in machine translation : a feasibility study. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, p. 130–138, Stroudsburg, PA, USA : Association for Computational Linguistics.

CORBETT G. G. (2010). Canonical derivational morphology. *Word Structure*, **2**(3), 141–155.

CORBIN D. (1985). Les bases non-autonomes en français ou comment intégrer l'exception dans le modèle lexical. *Langue française*, **66**, 54–76.

DAL G., NAMER F., HATHOUT N. & AMSILI P. (1999). Construire un lexique dérivationnel : théorie et réalisations. In *Actes de la Conférence TALN 1999*, p. 115–124, Cargèse, France.

FRADIN B. (2003). *Nouvelles approches en morphologie.* Paris, France : PUF.

HATHOUT N. & TANGUY L. (2005). WEBAFFIX : une boîte à outils d'acquisition lexicale à partir du Web. *Revue Québécoise de Linguistique*, **32**(1), 61–84.

MATTHEWS P. H. (1974). *Morphology*. Cambridge, Royaume-Uni : Cambridge University Press.

ROMARY L., SALMON-ALT S. & FRANCOPOULO G. (2004). Standards going concrete : from lmf to morphalou. In *Actes du Workshop on Electronic Dictionaries de Coling 2004*, Genève, Suisse.

SAGOT B. (2010). The L*efff*, a freely available, accurate and large-coverage lexicon for French. In *Proceedings of the 7th Language Resource and Evaluation Conference*, La Valette, Malte.

SAGOT B. & BOULLIER P. (2008). SXPipe 2 : architecture pour le traitement présyntaxique de corpus bruts. *Traitement Automatique des Langues (T.A.L.)*, **49**(2), 155–188.

TANGUY L. & HATHOUT N. (2002). Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du web. In *Actes de la Conférence TALN*, p. 245–254, Nancy, France.