Acquisition automatique de sens à partir d'opérations morphologiques en français : études de cas

Fiammetta NAMER

ATILF – Université Nancy 2
CLSH – BP3397 – 54015 Nancy Cedex
namer@univ-nancy2.fr

Résumé – Abstract [*]

Cet article propose une méthode de codage automatique de traits lexicaux sémantiques en français. Cette approche exploite les relations fixées par l'instruction sémantique d'un opérateur de construction morphologique entre la base et le mot construit. En cela, la réflexion s'inspire des travaux de Marc Light (Light 1996) tout en exploitant le fonctionnement d'un système d'analyse morphologique existant : l'analyseur DériF. A ce jour, l'analyse de 12 types morphologiques conduit à l'étiquetage d'environ 10 % d'un lexique composé de 99000 lemmes. L'article s'achève par la description de deux techniques utilisées pour valider les traits sémantiques.

This paper presents an approach which aims at automatically tagging a French lexicon with semantic features. This approach makes use of correspondences which are fixed by the semantic instruction of morphological operators, between the base and the derived word. It is partly inspired by Light (1996) work, and makes use of an existing morphological parser : the DeriF system. So far, 12 morphological types have been analysed, enabling the semantic tagging of circa 10 % of a lexicon made up of 99000 lemmas. The paper ends up with the description of two techniques that have been applied to validate the semantic features.

Keywords – Mots Clés

Morphologie dérivationnelle - affixation et conversion - acquisition de traits sémantiques

Derivational Morphology - affixation and conversion -Semantic feature acquisition

1 Introduction

L'information sémantique est cruciale pour l'ensemble des applications en TALN. Cette information, cependant, n'est pas toujours accessible : le codage sémantique est une tâche complexe, donc à coût élevé. Nous allons voir comment la morphologie peut être utilisée pour

[*] Je remercie les relecteurs anonymes pour leurs commentaires et suggestions, qui m'ont aidée dans la rédaction de la deuxième version de cet article

amorcer le codage d'un lexique d'environ 99000 lemmes catégorisés au moyen de traits sémantiques. L'expérience relatée se fonde sur les hypothèses linguistiques énoncées à l'origine dans (Corbin, 1987) et tire parti du fonctionnement d'un analyseur de mots construits existant : le système DériF (Namer, 1999). A la section 2, nous situons notre approche dans le cadre des travaux d'acquisition de sémantique lexicale. Ensuite, la section 3 illustre par des exemples le rôle des contraintes sémantiques dans des opérations de construction lexicale par suffixation, préfixation et conversion. Une partie de ces contraintes est mise en œuvre, au début de la section 4, dans le programme DériF pour annoter automatiquement certaines bases et dérivés au moyen des traits pertinents. La fin de la section 4 présente enfin une méthode de validation de ces traits, basée sur deux approches : filtre automatique, et recherche sur Internet.

2 Sens et ressources lexicales

Disposer de ressources lexicales munies d'informations sémantiques à la fois facilement exploitables et publiquement disponibles est un atout précieux dans l'élaboration de corpus annotés sémantiquement, tâche primordiale dans de nombreuses applications TALN : l'anglais dispose du réseau WordNet, développé à Princeton ((Miller, 1995), (Fellbaum, 1998)). Le consortium EuroWordNet (Vossen, 1998) a réalisé des réseaux sémantiques en néerlandais, italien, espagnol, français, allemand, tchèque et estonien, en lien avec WordNet. La base lexicale du français présente cependant des lacunes : seules les catégories noms et verbes y sont représentées, et les seules relations disponibles sont la synonymie, l'hyperonymie et l'hyponymie. De nombreux travaux récents portant sur l'étiquetage sémantique des corpus décrivent des méthodes d'acquisition de relations sémantiques complémentaires, comme par exemple la structure argumentale et les contraintes de sélection des verbes, ce que réalise pour l'anglais (MacCarthy et al. 2001). Parmi les travaux concernant le français, Fabre et Jacquemin (2000) proposent le codage manuel d'un ensemble minimal de traits verbaux binaires de manière à accroître la précision dans l'appariement de variantes teminologiques.

Cet article propose une méthode de codage automatique de traits lexicaux sémantiques en français qui se veut complémentaire à la fois d'EuroWordNet, pour la nature des traits codés, et de (Fabre & Jacquemin, 2000), pour la méthode d'acquisition de ces informations : en effet, les traits sémantiques qui codent les entrées lexicales sont déduites de contraintes sémantiques associées aux opérations de construction de mots, s'exerçant tant sur la base que sur le dérivé. Cette méthode s'inspire des travaux de Light (1996) ; cependant, ce système d'acquisition s'appuie sur un programme d'analyse automatique du lexique construit en français, effectuant à ce jour l'analyse complète de 18500 mots construits appartenant à divers types morphologiques. En cela, notre approche s'éloigne des préoccupations de M. Light, dont l'objectif est d'associer pour l'anglais à chaque opération morphologique la représentation sémantique *la plus probable* de la base et du dérivé. Son souci étant de préserver un équilibre entre précision et rappel, un taux d'erreur non nul est inévitable (il varie de 0% pour l'assignation de l'étiquette "changement d'état" aux dérivés en *–en*, à 80% pour l'assignation du trait "état final opposé à la base" aux dérivés en *de-*). A l'inverse, notre objectif est de privilégier les opérations morphologiques conduisant à un codage sémantique d'une précision proche de 100%. L'approche proposée ici est illustrée par l'étude de quelques opérations de construction de mots (suffixation, préfixation ou conversion) sélectionnées pour leur régularité et, si possible, leur productivité, et donc les plus pertinentes du point de vue

qualitatif (les opérateurs réguliers affichant un taux élevé de précision dans l'assignation des traits sémantiques à la base et au dérivé) et, le cas échéant, quantitatif (les opérateurs productifs garantissant un nombre élevé d'entrées codées dans la base).

3 Cadre théorique

Le cadre théorique dans lequel se situent les analyses linguistiques proposées dans cette section est décrit essentiellement dans (Corbin, 1987), où la morphologie dérivationnelle est conçue comme le calcul conjoint de la structure et du sens des unités lexicales. Cette perspective " associative " donne à analyser un mot comme construit s'il a une structure construite et un sens construit qui soit calculable à partir de celle-là.

3.1 Opérateurs constructionnels et contraintes sémantiques

Outre la composition, qui met en jeu deux unités lexicales à sens référentiel, les opérations de construction de mots relèvent de deux types possibles de procédés : l'affixation (application d'un suffixe ou un préfixe à une base) associe sens et forme, alors que la conversion (appelée aussi "dérivation impropre", comme le relèvent, entre autre Arrivé *et al.* (1986)) met en relation deux catégories grammaticales sans l'entremise de matériel lexical. Les opérations qui nous intéressent ici doivent avoir, comme propriété commune, la faculté d'exercer des contraintes sélectionnelles spécifiques et régulières sur la base et/ou le dérivé qu'elles relient. De cette manière, le calcul automatique des traits sémantiques sur les entrées de la base sera précis, le nombre d'exceptions limité, et l'amorçage du codage sémantique qui en résultera sera qualitativement pertinent. Un certain nombre d'opérations vérifient ces conditions, qu'il s'agisse d'affixation ou de conversion. Nous laissons volontairement de côté l'étude théorique des noms déverbaux de procès ; soit parce que le typage du dérivé en tant que nom abstrait est une propriété connue (cf. *-age* (*repassage*) ou *–ment* (*gonflement*)) et que toute spécification supplémentaire est impossible en l'absence de traits sémantiques sur le verbe ; soit parce que le typage du dérivé est non déterministe, et dépend de l'aspect télique du verbe de base (ainsi, les noms en *–tion* désignent à la fois un procès, et soit une manifestation de ce procès (*destruction*) soit l'entité concrète résultante (*construction*)) ; pour les mêmes raisons, nous ne dirons rien des noms abstraits de propriété construits par *-ité* [1] . En revanche, nous examinons quelques opérateurs peut-être moins connus, mais dont l'instruction sémantique permet d'étiqueter de façon fine des noms , des verbes, et, plus rarement, des adjectifs.

3.1.1 Typage des noms

Chacun à sa manière, les suffixes *–aille, -aie,* et *-oir(e)* entraînent un typage sémantique homogène du nom en position de base et/ou du nom construit.

Il existe deux suffixes **–aille** (Aliquot-Suengas, 1999), qui sélectionnent des bases concrètes, mais qui construisent des noms (également concrets) ayant des caractéristiques différentes.

[1] Pour le typage des nominalisation en *–tion*, on pourra se référer à (Jacquey 2002). En ce qui concerne *–ment* et *–age*, cf. (Kelling 2001) ; enfin, voir (Dal *et al*, 1999) au sujet de *–ité*.

On distingue en effet une majorité de noms en *–aille* collectifs à valeur évaluative (*flicaille* ou *ferraille* sont marqués négativement), qui désignent des agglomérats dont il est impossible de distinguer les morceaux, qu'ils soient comptables (*flic*) ou massifs (*fer*), et les noms en *–aille* à valeur énonciative argotique non péjorative, qui ne sont pas des collectifs (*jupaille*). Seuls les premiers sont pertinents dans un lexique de la langue générale. Comme *-aille*, le suffixe **-aie**, (cf. Aliquot-Suengas, 1996), est un constructeur de noms concrets collectifs qui sélectionne exclusivement des bases désignant des espèces végétales, et programme le nom qu'il construit à décrire un tout (*ormaie, bananeraie*), où chaque entité qui le constitue (*orme, bananier*) est indépendante des autres, contrairement, par exemple, à ce que l'on a observé avec *-aille* formateur d'évaluatifs. Contrairement aux deux premiers, enfin, le suffixe **-oir(e)** sélectionne des bases verbales et construit des noms concrets désignant le lieu (*fumoir*) ou l'instrument (*hachoir*) participant au déroulement du procès que décrit la base. Souvent, le nom peut désigner une entité qui sert à la fois de lieu et d'instrument (*abreuvoir*).

3.1.2 Typage croisé de verbes et adjectifs

L'ensemble des opérations de construction de verbes désadjectivaux (affixation et conversion) produisent par essence des prédicats de changement d'état, l'état final étant décrit par l'adjectif en position de base, qui de ce fait exprime une propriété que l'on peut qualifier d'extrinsèque [2]. Quand **a-** fabrique un verbe à partir d'une base adjectivale (*aplatir*, *allonger*), celui-ci est, causatif, transitif et le référent de son objet direct se retrouve dans l'état décrit par l'adjectif à la fin du déroulement du procès. A quelques exceptions près (*brutaliser, bêtifier*), il en est de même pour **-is(er) et –ifi(er).** Quant aux verbes désadjectivaux construits par conversion, ils sont soit causatifs, transitifs (*vider*) soit résultatifs intransitifs (*blêmir*). Certains admettent les deux constructions : c'est notamment le cas des verbes à base adjectivale chromatique (*rougir*, *brunir*).

Un autre cas de typage verbal homogène s'observe par exemple avec les bases sélectionnées par le suffixe **-able** formateur d'adjectifs (cf. (Plénat, 1988), repris et modélisé dans (Dal *et al.,* 1999)). En effet, quand ce suffixe construit un adjectif déverbal, celui-ci désigne une aptitude, une propriété intrinsèque, et la base verbale est généralement transitive. Quand elle ne l'est pas, elle désigne un prédicat intransitif de mouvement se construisant avec modifieur exprimant la surface sur laquelle s'effectue le mouvement : *circulable, skiable, navigable, roulable, surfable.*

On remarquera (Dal et Namer, 2000) que la propriété extrinsèque observée sur les adjectifs sélectionnés par les opérateurs formateurs de verbes est incompatible avec la propriété intrinsèque des adjectifs en *–able*, ce qui explique que ces derniers sont des bases très improbables pour des verbes désadjectivaux de changement d'état, comme l'illustrent les exemples agrammaticaux suivants : **lavabiliser, *imperméabl(ir|er), *aportabl(ir|er), *croyabilifier*.

[2] A l'exclusion de *dé-,* qui marque la suppression, l'annulation pour le thème du verbe, de l'état décrit par l'adjectif, à la fin du déroulement du procès (*déniaiser, délasser*).

3.1.3 Typage croisé de verbes et noms

Le dernier exemple concerne le préfixe **é-,** illustrant le cas des verbes dénominaux construits [3]. Aurnague et Plénat (1997) décrivent *é-* comme un marqueur de dissociation entre une partie concrète le plus souvent inanimée, désignée par ce à quoi réfère la base, et le tout (l'objet direct du verbe construit, qui par conséquent est un causatif transitif). Ils proposent une tripartition de ces verbes selon que la partie est (1) un constituant naturel du tout (*effeuiller une plante, épépiner un raisin*), (2) produite à partir du tout à l'issue du procès, et ne préexistant pas à celui-ci *(effranger un tapis)*, (3) un parasite du tout (*épouiller un chien, épousseter un meuble).*

3.2 Traits codés

Il est possible de déduire un ensemble de traits à partir des descriptions présentées en 3.1. Bien sur, le nombre de ces traits ainsi que leur structuration sera vraisemblablement remis en cause avec l'analyse de nouveaux types morphologiques. Nous avons alors préféré dans un premier temps adopter une typologie allégée, parfois hétérogène, de manière à simplifier les méthodes de traitement. C'est ainsi que le codage des adjectifs se limite à l'opposition extrinsèque/intrinsèque. En ce qui concerne le codage des noms et des verbes, il a été procédé à chaque fois à une structuration minimale sous forme de listes de trois champs. Tout champ non pertinent est codé xxx. Ainsi, un nom est soit concret, soit abstrait, trait qui constitue le premier champ de la liste. Le second spécifie les noms concrets : inanimé, que l'on oppose à animé. Cette position reste sous-spécifiée pour les noms abstraits. La troisième position est occupée par la caractérisation ultérieure du nom : collectif, lieu, espèce-végétale, instrument, pour les noms concrets, et propriété ou procès pour les abstraits. En ce qui concerne les verbes, on oppose d'abord causatif et résultatif, puis transitif et intransitif. Le troisième champ décrit la structure argumentale correspondante, étiquetée thématiquement. Le Tableau 1 à la section 4.1.2. illustre la structuration choisie.

4 Mise en oeuvre : programme et résultats

Le calcul automatique de traits sémantiques sur les entrées d'un lexique de lemmes étiquetés, inspiré des hypothèses théoriques exposées en 3, tire profit de l'existence d'un programme d'analyse automatique de mots construits, baptisé DériF ("Dérivation en Français"). Dans

[3] Je ne parlerai par contre ni de suffixation (les bases des verbes dénominaux construits par *–is(er)* et *–ifi(er)* appartiennent à des types sémantiques variés (*vampiriser* vs *révolveriser*, *momifier* vs *baronifier*) reflétant les variantes de l'instruction sémantique du suffixe) ni de conversion N->V. En effet, cette opération sélectionne aussi bien des noms d'instrument ou lieu concrets (*hache/hacher, coffre/coffrer*) que abstraits (*force/forcer*) (Corbin, à paraître). La base peut également faire référence à un humain, jouant alors le rôle d'agent typique (*singe/singer*), alors même que l'opération de conversion inverse, produisant essentiellement des noms de (résultat ou manifestation de) procès (*voler/vol, chuter/chute*), peut conduire aussi à la construction de noms d'agents (*garder/garde*) (Corbin, 1987). On le voit, le choix des critères à prendre en compte pour déterminer les contraintes sous-jacentes à l'opération de conversion N/V et à son orientation nécessite une réflexion qu'abordent entre autre Kerleroux (1996, 1997) et qui dépasse le cadre de cet exposé. Il est néanmoins clair que la conversion N->V n'est pas homogène du point de vue de l'étiquetage sémantique du nom ou du verbe.

cette dernière section, et à la suite d'une brève description du fonctionnement du programme, ainsi que de l'extension permettant le codage sémantique, nous examinerons les résultats auxquels nous parvenons au moyen de ce système.

4.1 Etiquetage sémantique

4.1.1 DériF

Le programme d'analyse morphologique DériF est développé dans le cadre du projet MorTAL (cf. (Namer, 1999) et (Hathout *et al,* à paraître)). A ce jour, DériF effectue l'analyse morphologique complète des mots construits au moyen des suffixes *–able, –ité, –et(te), –is(er)* et *–ifi(er*), ainsi que des préfixés en *dé–, in–*. L'étude d'autres affixes est partiellement réalisée : les noms dénominaux en *–aie* et *–aille*, les verbes déverbaux préfixés par *re–,* les noms déverbaux suffixés par *–age*, *-oir*, *–tion, -ment* et *–eur*, les verbes désadjectivaux et dénominaux construits au moyen des préfixes *é-* (*ébarber, éclairer)* et *a-* (*aplatir, allonger, atterrir*). De même, les opérations de conversion N->V et A->V sont en cours de réalisation. La conception de DériF obéit aux hypothèses linguistiques qui sous-tendent le modèle morphologique présenté à la section 3 : l'objectif visé étant la réalisation d'une base lexicale dont chaque entrée est munie d'informations morpho-sémantiques avec un taux d'erreur nul, des validations humaines sont nécessaires, et entraînent, le cas échéant, des versions successives du système. DériF est appliqué à un lexique de lemmes étiquetés extraits de la compilation effectuée à partir des nomenclatures du TLF et du Robert électronique, et totalisant quelque 99000 entrées : 18500 lemmes de ce lexique correspondent aux séquences ci-dessus, et sont analysés par DériF.

DériF calcule l'arbre d'analyse de chaque entrée : cette représentation crochetée est reprise sous forme de famille ordonnant l'ensemble des bases successives reconnues par l'analyseur ; le troisième composant du résultat est une glose en langue naturelle exprimant l'opération sémantique induite par l'affixe s'étant appliqué en dernier (ex. (1)). DériF est récursif et calcule donc toute la famille d'un mot construit (ex (2)). Quand un mot est le produit de plusieurs opérations de construction, l'analyseur procède à un traitement hiérarchisé correspondant aux portées respectives de ces opérations : (3a) vs (3b) illustrent l'ordre entre conversion et affixation, (3c) vs (3d), la portée relative du suffixe et du préfixe. DériF manipule les données et les résultats sous forme de listes, et est donc capable de générer autant de solutions qu'il y a d'ambiguïtés éventuelles dans l'analyse d'un mot construit (ex. (4)). Enfin, de par sa conception, DériF est capable d'analyser des mots inconnus (possibles et non attestées), comme le montre l'exemple (5).

1) continuité, N => [[continu A] ité N] (continuité/N, continu/A) :: "faculté d'être ***continu"***
2) explicabilité, N => [[[expliquer V] able A] ité N] (explicabilité/N, explicable/A, expliquer/V) :: "faculté d'être ***explicable"***
3a) ragrafer, V => [re [[agrafe N] (er) V] V] (ragrafer/V, agrafer/V, agrafe/N) :: ***agrafer*** *une nouvelle fois*
3b) portable, N => [[[porter V] able A] N] (portable/N, portable/A, porter/V) :: "entité ayant pour propriété principale d'être ***portable"***
3c) introuvable, A =>[in[[trouver V] able A]A] (introuvable/A, trouvable/A, trouver/V) :: "non ***trouvable"***

3d) désossable, A => [[dé [os N] V] able A] (désossable/A, désosser/V, os/N) :: "que l'on peut ***désosser"***
4) décoller,V => [dé1 [cou N] V] (décoller/V, cou/N) ::(Enlever | (Faire) sortir de) ***cou***
décoller,V => [dé1 [colle N] V] (décoller/V, colle/N) ::(Enlever | (Faire) sortir de) ***colle***
5) adhérette, N => [[adhérer V] ette N] (adhérette/N, adhérer/V) :: "objet de ou instrument pour ***adhérer"***

4.1.2 Annotation de la base au moyen de traits sémantiques

Les types morphologiques décrits en 3.1. auxquels s'ajoutent les noms déverbaux en *–age* et *–ment*, et les noms désadjectivaux en *–ité* donnent lieu au codage automatique d'environ 9000 entrées, soit 9% du lexique de départ. Quand les opérations morphologiques décrites à la section 3 sont activées lors de l'analyse d'un lemme L du corpus, la tâche de codage sémantique est déclenchée en parallèle sur une copie du corpus, où la description de L est enrichie au moyen du ou des traits dépendants de l'instruction sémantique de l'opérateur, ainsi qu'en témoigne l'échantillon codé dans le Tableau 1, où sont indiqués (en colonne 5) les traits calculés pour L (apparaissant dans la colonne 4) en tant que base ou dérivé (indiqué par « B » ou « D », colonne 3) de l'opération morphologique qui s'est appliquée (reportée dans la colonne 2). Les traits codés sont parfois sous-déterminés (ligne 2, 4), parfois erronés (ligne 3). Le fait que L accumule plusieurs séries de traits est dû soit au codage en soi (ligne 4), soit à l'implication de L dans plusieurs opérations morphologiques (ligne 7).

	Op.	**(D)érivé / (B)ase**	**Lemme Etiqueté**	**Traits Codés**
1	a-	D	`aplatir/VERBE`	`(causatif, transitif, [cause, thème])`
2	A->V	D	`blêmir/VERBE`	`(causatif, transitif, [cause, thème]), (résultatif, intransitif, [thème])`
3	-ifi(er)	D	`bêtifier/VERBE`	`(causatif,transitif, [cause, thème])`
4	-able	B	`skier/VERBE`	`(xxx, transitif, [agent, thème]), (xxx, intransitif, [agent, (lieu(sur))])`
5	-oir	D	`abreuvoir/NOM`	`(concret, inanimé, lieu), (concret, inanimé, instrument)`
6	-ité	D	`loyauté/NOM`	`(abstrait, xxx, propriété)`
7	é- -aille	B D	`tripaille/NOM` `tripaille/NOM`	`(concret, inanimé, xxx)` `(concret, xxx, collectif)`

Tableau 1

4.2 Validation des traits

Comme l'illustre le Tableau 1, une partie des traits codés automatiquement va nécessiter une vérification, et, le cas échéant, une correction. En effet 1547 entrées sont munies d'un codage ambigu (c'est le cas des bases verbales des adjectifs en *-able*, des noms déverbaux en *–oir*, et des verbes désadjectivaux obtenus par conversion) ; on s'attend à ce que la phase de validation permette de lever certaines ambiguïtés. A l'inverse, il est vraisemblable que certains codages soient erronés (ainsi, parmi les 425 verbes en *–iser*, les 58 verbes en *–ifier* et

les 90 verbes en *a-* analysés par défaut comme étant `causatifs transitifs`, une partie d'entre eux possède une construction exclusivement intransitive) ; la validation a ici pour objectif de guider vers le repérage de ces entrées mal codées, dont la correction *in fine* sera manuelle. D'autres traits erronés sont par contre indétectables automatiquement : c'est le cas de quelques noms de type **`animé`** désignant des parasites (*chenille, pou, puce, taupe*) en position de base des verbes en é- et codés par défaut - à tort - **`inanimé`** (cf. section 3.1.3). Ces vérifications vont être réalisées automatiquement au moyen de deux techniques.

4.2.1 Validation par filtrage :

C'est tout d'abord au moyen de filtres automatiques que sont repérés les codages des lemmes qui sont le fait d'opérateurs différents ; ces filtres comparent les informations co-occurrentes, de manière à relever les traits complémentaires ou unifiables. A l'heure actuelle, la mise en œuvre de cette méthode n'apporte pas de résultats spectaculaires, en raison du nombre encore faible de types morphologiques traités. Seuls 83 lemmes reçoivent leur codage sémantique à partir de deux opérations morphologiques distinctes : c'est le cas des bases verbales construites dans les adjectifs en *–able* qui se retrouvent avec une double série de traits, dont le filtre vérifie la compatibilité. Par exemple, sur les verbes convertis (ex. 6), le filtre élimine la dernière série de traits, incompatible avec le type de base verbale, et unifie les 1ère et 3ème séries (la 1ère étant un cas particulier de la 3ème).

6) préciser/VERBE (causatif, transitif, [cause, thème]), (résultatif, intransitif, [thème]), (xxx, transitif, [agent, thème]), (xxx, intransitif, [agent, (lieu(sur)))

L'autre type de codage validé par filtre est celui de lemmes comme le nom *tripaille*, (cf. Tableau 1), qui possède deux séries de traits complémentaires : cette entrée réunit à la fois les contraintes des bases nominales sélectionnées par *é-* (cf. 3.1.3) et les propriétés des noms construits au moyen de *–aille* (cf. 3.1.1). Le filtre remplace automatiquement dans l'une des listes de traits, la valeur vide ("`xxx`") par l'instance occupant la même place dans l'autre liste, aboutissant à la fin au codage : `(concret, inanimé, collectif)`. On peut prédire qu'avec l'élargissement de la couverture en termes d'opérations morphologiques, la juxtaposition des traits repérée automatiquement au moyen de filtres pourra mettre en lumière des phénomènes intéressants : si les traits se contredisent, cela pourra signifier la présence d'exceptions à des règles de formation de mots bien sûr, mais aussi l'existence d'acceptions multiples pour une entrée lexicale ; ainsi, quand l'analyseur traitera les noms suffixés en *–ure* et en *–ier* on peut supposer que l'entrée nominale *avocat* recevra à la fois le trait `(concret, inanimé, xxx)` quand il est analysé à partir de *avocatier* et `(concret, humain, profession)` quand il est relié à *avocature* [4]. L'incompatibilité entre `inanimé` et `humain` témoigne ici de l'existence de deux homonymes : le fruit et l'individu. A l'inverse, on peut aussi imaginer que les traits calculés sur une entrée spécifieront l'instruction sémantique d'un opérateur accédant à cette base. C'est ce qu'on est en droit de s'attendre avec le verbe *accessoiriser*, actuellement analysé au moyen d'une disjonction de gloses (cf. note 3), que l'étiquetage instrumental du nom *accessoire*, permettra de réduire à : "`munir de` **`accessoire`**`".`

[4] Les contraintes sémantiques des suffixes –ier et –ure sont décrites, respectivement, dans (Corbin & Corbin 1991) et (Lecomte 1997).

4.2.2 Validation par Internet:

La spécification des informations sémantiques peut tirer profit des résultats d'une autre approche de validation. Cette approche, déjà mise en œuvre dans (Dal & Namer 2000), consiste à tester l'existence sur Internet de termes générés morphologiquement et présentant *a priori* les propriétés sémantiques dont on veut tester la validité. La démarche est la suivante : 1800 verbes obtenus par conversion A->V, suffixation (*-is(er), -ifi(er)*) ou préfixation (*a-, é-*) sont utilisés comme base pour générer des adjectifs en *–able* absents du lexique d'origine ; ces adjectifs sont ensuite employés comme requête par un robot interrogeant trois moteurs de recherche (Yahoo, Francité et Altavista) : les résultats (absence/présence de documents répondant à la requête, nombre de documents retournés, similarité des réponses obtenues par les trois moteurs) ont une double utilité : tout d'abord, ils permettent d'enrichir automatiquement le lexique d'origine au moyen de l'ensemble des adjectifs en *–able* retrouvés par les trois moteurs, soit près de 500 "néologismes" (e.g. *trivialisable, rigidifiable, amochable*). Ensuite et surtout, ils servent de guide pour la post-édition des traits codés automatiquement sur les verbes-bases. Ainsi, le fait de ne pas retrouver sur Internet un adjectif construit en *–able* (e.g. *blêmissable, bêtifiable, fraternisable*) peut signifier que la construction de celui-ci est illégitime, et donc que sa base est non transitive [5]. De telles bases verbales sont donc revues manuellement, ce qui peut conduire à désambiguïser le codage de verbes construits par conversion, qui ne sont que résultatifs (e.g. *blêmir*), ou à corriger le trait transitif des autres verbes désadjectivaux (e.g. *bêtifier, fraterniser*).

5 Conclusion

Le travail présenté propose une méthode pour coder automatiquement des traits sémantiques lexicaux en se servant des contraintes imposées par les opérateurs de construction de mots. Seuls quelques traits sémantiques sont accessibles par cette méthode, et seule une partie du lexique est codée à ce jour. Il est prévisible que l'évolution de l'analyseur entraîne un enrichissement du codage pour les noms et les verbes. Mais même partiels, ces résultats sont néanmoins utilisables d'ores et déjà en analyse morphologique pour expliciter des exceptions ou spécifier des gloses. Ils sont également exploitables pour sélectionner les bases candidates à la génération automatique de mots construits : ainsi, les traits codant les verbes en *é-* ou en *a-* en font des bases potentielles pour des adjectifs en *–able*. Enfin, l'approche présentée peut servir, dans un système d'analyse automatique, à la caractérisation sémantique des mots construits qui sont absents des dictionnaires de langue parce qu'ils appartiennent à un type morphologique trop productif, ou parce qu'il s'agit de néologismes : la nature régulière des néologismes construits laisse en effet présager une bonne capacité prédictive du système d'assignation des traits.

Références

Arrivé, M, Gadet F. et Galmiche M., 1986 , *La Grammaire d'Aujourd'hui,* Paris, Flammarion.

[5] La prudence face aux résultats, et contrôle manuel de ceux-ci est rendue nécessaire du fait de l'instabilité bien connue des informations sur Internet.

Aliquot-Suengas S., 1996, *Référence collective / sens collectif. La catégorie du collectif dans les noms suffixés du lexique français*, Thèse de doctorat, Université de Lille III.

Aliquot-Suengas S., 1999, De la poiscaille dans la piscaille. Les noms dénominaux évaluatifs construits avec une forme suffixale -ail(le), *Silexicales, n°2*, Lille.

Aurnague, M. et Plénat, M., 1997, Manifestations morphologiques de la relation d'attachement habituel, *Silexicales N°1*, Lille

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

Corbin D., à paraître, French (Indo-european : Romance), *in* G. Booij, C. Lehmann & J. Mugdan eds, *Morphology. A Handbook on Inflection and Word Formation*, Berlin/New York, Walter de Gruyter [article n° 121].

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

Dal, G., Hathout, N. et Namer, F., 1999, Construire un lexique dérivationnel : théorie et réalisations**,** *Actes de TALN 1999***,** Cargèse**.**

Dal, G., et Namer, F., 2000, Génération et analyse automatiques de ressources lexicales construites utilisables en recherche d'informations, *TAL* 41-2, pp. 423-446.

Fabre, C. et Jacquemin, Ch., 2000, Boosting Variant Recognition with Light Semantics, in *Proceedings of COLING2000*, Saarbrücken.

Fellbaum, C. 1998, *WordNet, An Electronic Lexical Database*, MIT Press.

Hathout, N., Dal, G. et Namer, F., à paraître, Une base de données constructionnelles expérimentale : le projet MorTAL, in P. Boucher ed., *Many Morphologies*, Cambridge Mass., Cascadilla Press.

Jacquey E., 2002, Les déverbaux d'action en français : quel type d'ambiguïté et quelle catégorie conceptuelle, à paraître dans les Actes de la conférence *Représentations du sens linguistique*, Lincom Europa.

Kelling C., 2001, Agentivity and Suffix Selection, in M. Butt,T. H. King (eds): *Proceedings LFG 2001*, Stanford: CSLI Publications.

Kerleroux, F., 1996, *La coupure invisible*, Presses Universitaires du Septentrion, Lille.

Kerleroux, F., 1997, De la limitation de l'homonymie entre noms déverbaux convertis et apocopes de noms déverbaux suffixés, *Lexicales N°1*, Lille p. 163-172

Lecomte E., 1997 : Tous les mots possibles en –ure existent-ils?", *Silexicales N°1*, Lille.

Light, M., 1996, Morphological Cues for Lexical Semantics, *Proceedings of the 34th ACL.*

McCarthy, D., J. Carroll and J. Preiss (2001) Disambiguating noun and verb senses using automatically acquired selectional preferences, In *Proceedings of the SENSEVAL-2 Workshop at ACL/EACL'01* , Toulouse, France.

Miller, G.A. 1995, WordNet: A Lexical Database, Communication of the *ACM*, vol 38: N°11.

Namer, F., 1999, Traitement automatique de la dérivation: le cas des noms et adjectifs en –et(te), *Silexicales n°2*, Lille.

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

Vossen P. (Ed.) (1998), *EuroWordNet A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic publishers.