Conférence TALN 1999, Cargèse, 12-17 juillet 1999

Acquisition automatique de connaissances morphologiques sur le vocabulaire médical

Natalia Grabar et Pierre Zweigenbaum

DIAM - SIM/AP-HP et Université Paris 6
{ngr,pz}@biomath.jussieu.fr
http://www.biomath.jussieu.fr/

Résumé

La morphologie médicale est riche et productive. À côté de la simple flexion, dérivation et composition sont d'autres moyens pour créer des mots nouveaux. La connaissance morphologique se révèle par conséquent très importante pour toute application dans le traitement automatique du langage médical. Nous proposons une méthode simple et puissante pour l'acquisition automatique d'une telle connaissance. Cette méthode tire avantage de listes de termes synonymes disponibles afin d'amorcer le processus d'acquisition. Nous l'avons expérimentée dans le domaine médical sur le Microglossaire de Pathologie SNOMED. Les familles de mots morphologiquement reliés que nous avons obtenues sont correctes à 95 %. Utilisées dans un outil d'aide au codage avec expansion de requête, elles permettent d'en améliorer les performances.

1. Introduction

Les mots médicaux présentent une morphologie riche et productive. En français, en anglais et dans d'autres langues européennes, ces mots sont souvent formés avec des morphèmes grecs ou latins. La décomposition d'un mot en morphèmes est utile pour obtenir ses unités de sens élémentaires. C'est une des clés de l'analyse sémantique des expressions médicales. En particulier, cela mène à une indexation plus fine des textes et des termes médicaux, et une meilleure exactitude dans l'extraction d'information et l'aide au codage (Wingert *et al.*, 1989).

La morphologie médicale a été étudiée par plusieurs chercheurs et dans plusieurs langues (Pacak *et al.*, 1980; Wingert *et al.*, 1989; Dujols *et al.*, 1991; McCray *et al.*, 1994; Spyns, 1994; Lovis *et al.*, 1995), mais il n'existe à ce jour aucune base de données morphologiques disponible publiquement sur la langue médicale française. La connaissance morphologique qui se trouve à la base de ces travaux a été recueillie grâce à l'analyse manuelle. Sa constitution étant une tâche laborieuse et difficile, des outils automatiques ont été utilisés dans certains (McCray *et al.*, 1994). Des modèles linguistiques informatiques comme la morphologie à deux niveaux (Koskenniemi, 1983; Antworth, 1990) ont par ailleurs servi de base à des descriptions générales de langues naturelles, également mises au point principalement manuellement.

D'autres chercheurs ont travaillé sur l'acquisition de données morphologiques à partir de cor-

pus. Ainsi, (Jacquemin, 1997) cherche à identifier dans un corpus des variantes morphologiques de mots présents dans un thésaurus. Il recherche les mots d'un même terme du thésaurus qui apparaissent, sous une forme approchée (avec une chaîne initiale commune), dans une même collocation dans le corpus (à la distance de 2 à 5 mots). Sa méthode retrouve, par exemple, les collocations *« gene expression »* et *« genic expression »*, où *« gene »* et *« genic »* sont effectivement de la même famille morpho-sémantique. Elle a été appliquée au corpus Medic et son thésaurus Pascal (créés à l'INIST). Une autre méthode a été proposée par (Xu & Croft, 1998) qui ont travaillé avec un corpus de presse. Ils examinent des collocations d'une taille plus grande : une fenêtre de 100 mots, ou bien un texte entier. Le résultat obtenu en sortie est assez fiable et peut être utilisé pour améliorer la performance des algorithmes de désuffixage. En reprenant et améliorant une méthode proposée initialement par Harris, (Déjean, 1998) apprend les morphèmes les plus fréquents dans les mots d'un corpus en observant la variation du nombre de lettres différentes qui peuvent suivre une séquence de lettres. Une autre méthode, fondée sur l'emploi du désuffixeur `findaffix` (disponible sous Unix), est proposée par (Dal *et al.*, 1999) pour acquérir automatiquement des liens morphologiques entre les mots. Elle cherche à identifier des liens morphologiques entre des formes attestées du lexique TLFnome (INaLF), et a été testée sur la décomposition des adjectifs en *« -able »* et en *« -ité »*. Cette méthode donne en sortie une liste de décompositions candidates qui doit être soumise à un filtrage manuel important. Enfin, (Theron & Cloete, 1997) ont montré comment acquérir des règles à deux niveaux à partir de paires de mots morphologiquement reliés.

Nous proposons une méthode qui extrait les données morphologiques sans information linguistique a priori. Cette méthode, comme `findaffix`, décompose les mots en racines et suffixes et, en même temps, produit les règles morphologiques et les familles de mots morphologiquement reliés. On induit en premier lieu, dans un contexte très restreint, les règles de dérivation potentielles. Ces règles sont ensuite appliquées à une plus grande liste de mots. Le niveau de bruit reste très bas grâce aux contraintes appliquées à deux moments clé du traitement : (i) les règles dérivationnelles sont apprises sur des termes synonymes; (ii) les familles de mots morphologiquement reliés n'incluent que des mots attestés du domaine.

On distingue classiquement trois types de dérivation (Moeschler & Auchlin, 1997) : (i) la variation flexionnelle (féminin, pluriel) (ii) la dérivation trans- et isocatégorielle (formation de mots d'une catégorie syntaxique différente ou identique) (iii) la composition. Les deux premiers types de dérivation ne causent pas de changement de sens majeur, tandis que le troisième est basé sur la modification du sens des mots dans la mesure où il opère sur des suffixes porteurs de sens. La méthode présentée traite indistinctement l'ensemble de ces types de dérivation.

Dans cet article, nous décrivons d'abord les données en français extraites des terminologies SNOMED Internationale (Nomenclature systématique de la médecine) et CIM-10 (Classification internationale des maladies) à partir desquelles l'étude a été menée. Nous détaillons ensuite les spécifications de la méthode d'acquisition et l'illustrons avec des exemples. Nous présentons les familles de mots obtenues en résultat, qui sont incluses dans un outil d'aide au codage. Nous analysons les avantages et les limites de cette méthode et réfléchissons sur ses perspectives.

2. Matériel de départ

Comme d'autres terminologies médicales, la SNOMED Internationale recense pour chaque concept le terme vedette (marqué '01') et des termes synonymes pour certains d'entre eux (marqués '02', '03' ou '05'). Le tableau 1 présente l'exemple de termes synonymes extraits de la

TAB. 1 – *Termes préférés et termes synonymes dans la SNOMED.*

Code	Classe	Terme
F-00470	01	symbiose
F-00470	02	commensalisme
F-00470	05	symbiotique
F-00470	05	commensal
T-51100	01	palais, SAI
T-51110	02	voûte palatine

version française du Microglossaire de Pathologie de la SNOMED Internationale (Côté, 1996). Le Microglossaire de Pathologie inclut dans sa version française 12 555 termes représentant 9 098 concepts différents. Parmi ceux-ci 2 344 concepts ont des termes synonymes. Le corpus sur lequel ont été apprises les règles dérivationnelles contient donc 2 344 séries de synonymes extraites du Microglossaire, ce qui fait 5 801 termes. La même méthode pourrait être appliquée à toute la nomenclature SNOMED Internationale (150 000 termes), mais seul le Microglossaire était disponible en français au moment de l'expérience.

Un autre type de données linguistiques utilisé est une liste de mots de référence. Cette liste a servi de base pour l'application des règles de dérivation apprises. Elle est représentative du domaine, mais peut être enrichie pour atteindre une meilleure couverture. Comme nous cherchons à enrichir un outil d'aide au codage dans la CIM-10, nous avons travaillé avec les mots des 10 797 principaux termes français de la Classification internationale des maladies (CIM-10) et les mots provenant du Microglossaire SNOMED. La liste de mots obtenue contient 8 150 formes distinctes.

Le processus d'acquisition peut être enrichi si d'autres sources de données sont disponibles. Nous avons pu disposer d'un lexique reliant les formes fléchies à leurs lemmes, couvrant le Microglossaire SNOMED et la CIM-10. Grâce à ces sources de données, nous avons étendu la couverture des familles morphologiques. Mais ceci n'est pas indispensable et les familles acquises sans connaissance externe sont directement utilisables.

3. Méthodes

3.1. Présentation

Nous considérons comme *morphologiquement reliés* une paire de mots qui sont dérivés de la même racine commune et, donc, partagent plus ou moins un sens commun. Dans la plupart des cas, de tels mots ont une chaîne de caractères commune. Dans beaucoup de langues, dont le français, cette chaîne de caractères est continue, par exemple, **symbio** dans la paire de mots *«* **symbio***se »* / *«* **symbio***tique »* (mais ceci n'est pas le cas dans les langues sémitiques, par exemple). Cette chaîne se trouve souvent au début des mots, comme dans l'exemple donné. Le moyen le plus simple de trouver un indice indiquant que deux mots sont potentiellement reliés morphologiquement est d'examiner leur *préfixe commun le plus long*. Si ce préfixe est « suffisamment long », les mots ont de bonnes chances d'appartenir au même paradigme morphologique.

Une telle approche peut conduire à beaucoup de bruit. Considérons, par exemple, la paire suivante : *«* **admi***nistratif »* / *«* **admi***ssion »*, dont les mots se trouvent dans notre liste. Bien

qu'ils aient une chaîne initiale commune longue de quatre caractères, ils ne sont pas morphologiquement reliés, car ils n'ont pas été créés à partir de la même racine. Même avec une longueur minimale du préfixe commun augmentée, il est possible de trouver des paires de mots semblables mais non pertinentes. Ainsi dans « **antidia***betique* » / « **antidia***rrhée* », où le préfixe commun est composé de sept lettres, les mots ne dérivent pas de la même racine (mais ils partagent le même préfixe commun).

3.2. Amorçage : apprentissage de règles de dérivation dans un contexte favorable

L'idée forte de cette étude est d'appliquer la méthode *dans un contexte très spécifique et favorable*. Ce contexte doit être tel que les mots comparés doivent, par leurs occurrences d'apparition, partager, avec une forte probabilité, le même sens. Nous avons trouvé ce contexte favorable dans les séries de synonymes extraites de terminologies, comme par exemple la SNOMED. Le tableau 1 montre que dans un tel contexte on retrouve des mots morphologiquement reliés : par exemple, « **symbio***se* » / « **symbio***tique* », « **commensal** » / « **commensal***isme* », « **pala***is* » / « **pala***tine* ».

Le cas de « **pala***is* » / « **pala***tine* » est particulièrement intéressant car la correspondance entre ces deux mots n'est pas explicite : *« palais, SAI »* / *« voûte palatine »* sont des termes composés (*« SAI » = « sans autre indication »*). Leurs mots doivent être alignés pour être comparés afin de trouver leur préfixe commun le plus long. Nous avons expérimentalement fixé la longueur minimale du préfixe commun à trois caractères : les mots sont considérés morphologiquement reliés dans ce contexte s'ils partagent au moins leurs trois premiers caractères.

Le fait de travailler avec des séries de termes synonymes réduit le risque de trouver des mots reliés morphologiquement, mais pas sémantiquement. Les paires de mots trouvés avec ce principe sont exactes dans leur quasi-totalité (nous discutons de l'évaluation globale des familles morphologiques trouvées à la section 4).

Par contraste, un algorithme procédant hors contexte au rapprochement systématique des mots commençant par les trois mêmes premières lettres conduirait à énormément de bruit. On arriverait ainsi en partant des formes du Microglossaire SNOMED à 1246 groupes comportant jusqu'à 109 formes. Par exemple, dans le groupe des 17 formes commençant par *« tro- »*,

> *« trocart »*, *« trochléaire »*, *« trois »*, *« troisième »*, *« troisièmement »*, *« trompe »*, *« trompes »*, *« tronc »*, *« troncs »*, *« trop »*, *« trophique »*, *« trophoblaste »*, *« trophoblastique »*, *« tropicale »*, *« trou »*, *« trouble »*, *« trouvé »*

on peut distinguer au moins 11 familles morphologiques différentes. Et parmi les 130 groupes de trois formes que l'on trouverait ainsi, moins de la moitié (56) constitueraient des familles morphologiques correctes.

Revenons à notre méthode. Dans l'étape suivante, les paires contenant le même préfixe sont réunies dans une famille morphologique, par exemple :

> « **cardi***aque* » / « **cardi***o* » / « **cardi***omégalie* » / « **cardi***opathie* » / « **cardi***te* »
>
> « **oesophag***e* » / « **oesophag***ien* » / « **oesophag***ienne* » / « **oesophag***o* »

Les paires de mots identifiés comme morphologiquement reliés constituent des instances potentielles de *règles de dérivation* comme celles décrites dans (McCray *et al.*, 1994) ou (Jacquemin, 1997).

Nous supposons l'exactitude de telles règles et enregistrons chacune d'elles. Une règle morphologique est composée d'une paire de suffixes, comme « *se/tique* ». Elle permet de transformer un mot terminé par *« –se »* en un mot où le suffixe *« -se »* est remplacé par le suffixe *« -tique »* (*e.g.*, *« sténose/sténotique »*). Les règles sont symétriques, et cette même règle permet de transformer les mots en *« -tique »* en mots en *« -se »*.

Puisque ces règles ont été apprises sur des paires de mots dans un contexte restreint, elles sont attestées au moins dans un exemple. Le pas suivant applique ces règles à d'autres mots.

3.3. Expansion des familles morphologiques avec des mots attestés

En principe, chaque règle morphologique peut être appliquée à un mot si ce mot est terminé par l'un des deux suffixes de la règle. Même si une telle dérivation est possible, il existe un risque important de génération de mots inexistants. C'est d'autant plus évident que certaines règles ont un suffixe nul. Par exemple, la règle *« /al »*, qui provient de l'appariement de *« ombilic »* et *« ombilical »*, peut ajouter le suffixe *« -al »* à n'importe quel mot.

C'est pourquoi il est important d'appliquer des contraintes ici aussi. Nous ne retenons que des *mots attestés*, et nous ne nous aventurons pas à supposer l'existence de mots nouveaux. Nous essayons d'appliquer chaque règle à chaque mot d'une liste de référence et acceptons la dérivation (i) si le mot est terminé par l'un des suffixes de la règle et (ii) si le mot dérivé se trouve également dans la liste de référence. Par exemple, la règle *« /al »* a marché pour 20 mots seulement (10 paires de mots) : *« médiastin »* / *« médiastinal »*, *« vagin »* / *« vaginal »*, etc., et n'a pas été appliquée sur les 245 autres mots de la liste qui se terminent par le suffixe *« -al »*, bloquant ainsi des dérivations incorrectes dans 98,7 % des cas.

Les paires de mots obtenues dans cette deuxième étape sont ajoutées aux familles initiales trouvées lors de la première étape. Par exemple, la famille *«* **alvéolaire** *»* / *«* **alvéolaire***s »* a été étendue et affinée en *«* **alvéol***aire »* / *«* **alvéol***aires »* / *«* **alvéol***e »*.

Les familles qui partagent un mot en commun sont réunies. Dans l'exemple ci-dessous, nous avons quatre familles différentes avec la racine *« adéno »* :

> *« adénomateuse »* / *« adénomatose »*
>
> *« adénomyomatose »* / *« adénomyome »*
>
> *« adénomateux »* / *« adénomatose »* / *« adénome »*
>
> *« adéno »* / *« adénoacanthome »* / *« adénocarcinome »* / *« adénomateuse »* / *« adénomateux »* / *« adénomatose »* / *« adénomatoïde »* / *« adénome »* / *« adénose »* / *« adénoïde »*

Le programme de regroupement de familles produit les familles suivantes :

> *« adéno »* / *« adénoacanthome »* / *« adénocarcinome »* / *« adénomateuse »* / *« adénomateux »* / *« adénomatose »* / *« adénomatoïde »* / *« adénome »* / *« adénose »* / *« adénoïde »* / *« adénoïdes »*
>
> *« adénomyomatose »* / *« adénomyome »*

La deuxième famille, n'ayant pas de mots communs avec les trois autres, est restée séparée. Parfois, la portée de regroupement peut être augmentée grâce à l'ajout de nouvelles dérivations. Nous avons effectué un seul regroupement, avant l'ajout de données externes.

Des données externes peuvent donc être ajoutées. Nous avons complété les familles avec des lexiques contenant les mots du Microglossaire SNOMED et les termes principaux de la CIM-10, qui relient chaque forme à son lemme. Ceci nous a permis d'étendre certaines familles morphologiques, ce qui est utile dans une optique d'expansion de requête. Par exemple, la famille *« abrasion »* / *« abrasé »* a été étendue en *« abrasion »* / *« abrasions »* / *« abrasé »* / *« abrasée »* / *« abrasées »* / *« abrasés »*.

4. Résultats

4.1. Un premier résultat

Nous avons implémenté cette méthode en utilisant des programmes `perl`, des scripts `sed`, `awk` et le filtre `sort`. Nous l'avons appliquée à 2 344 séries de synonymes provenant du Microglossaire de Pathologie de SNOMED et une liste de 8 150 mots provenant du Microglossaire et de la CIM-10. Le nombre de familles a été réduit pendant l'étape d'expansion, tandis que le nombre de mots a été augmenté. Le tableau 2 dresse le bilan du travail.

TAB. 2 – *La base de données morphologiques obtenue.*

Type de données	Nombre
Règles morphologiques induites	566
Nombre initial de familles	755
Mots par famille	3.39
Dérivations générées (paires de mots)	4 396
Nombre de familles après l'expansion	1 304
Mots par famille, avec les dérivations générées	3.67
Connaissance externe (paires lemme / mot fléchi)	8 280
Nombre de familles avec la connaissance externe	4 498
Mots par famille, avec la connaissance externe	4.27

4.2. Analyse manuelle des types de règles

Nous avons effectué une analyse manuelle des règles morphologiques. On y trouve les types présentés en introduction :

Les règles de flexion relient les formes différentes d'un mot (dans un dictionnaire ordinaire il n'y aura qu'une seule entrée). Par exemple, l'ajout de suffixe *« -s »* pour la formation du pluriel, ou de suffixe *« -e »* du féminin (règle« */s* » et « */e* »); on peut trouver aussi des règles flexionnelles plus complexes, comme la règle de formation du pluriel « *l/ux* » dans *« abdomina***l** *»* / *« abdomina***ux** *»*.

Les règles de dérivation transcatégorielles relient un mot d'une catégorie syntaxique à un mot d'une autre catégorie syntaxique. Par exemple, un nom à un adjectif, comme *« oesophag***e** *»* / *« oesophag***ien** *»* ou *« abdom***en** *»* / *« abdom***inal** *»* (règles « *e/ien* » et « *en/inal* »); ou un nom à un participe, comme *« abras***ion** *»* / *« abras***é** *»* (règle « *ion/é* »).

Les règles de dérivation isocatégorielles relient deux mots différents de la même catégorie, par exemple deux noms, comme *« hématom***ètre** *»* / *« hématom***étrie** *»*. Ce cas est rare.

Les règles de composition ajoutent un morphème lexical à la base du mot. Certains de ces morphèmes sont liés et ne peuvent pas apparaître isolément : *« -ite »* dans *« bronch***e** *»* / *« bronch***ite** *»* (règle « *e/ite* »). D'autres correspondent à des mots : *« -blastome »* dans *« angio »* / *« angio***blastome** *»*.

Il peut être utile de faire la différence entre les règles qui préservent le sens (ne causant qu'une légère variation : règles flexionnelles et dérivationnelles) et les règles qui modifient le sens (généralement par l'ajout d'information : règles de composition). Le premier type correspond à des variations morphosyntaxiques du lemme (suffixes grammaticaux). 404 familles morphologiques obtenues (65 %) contiennent des mots reliés par ce type de règles.

Le deuxième type de règles utilise des suffixes lexicaux qui ajoutent une information spécifique. Cette information relève généralement de la classe des affections, à travers, par exemple, le suffixe *« -ite »* (il est difficile cependant de s'accorder sur la frontière entre dérivation et composition, et on pourrait aussi considérer que *« -ite »* est un suffixe dérivationnel). Les suffixes peuvent aussi être plus spécifiques et correspondre à un mot existant par ailleurs, par exemple *« -blastome »*. 188 familles (30 %) se composent de mots reliés par ce type de règles, souvent en combinaison avec des règles du premier type.

4.3. Analyse des erreurs

Les 30 familles qui restent (5 %) contiennent chacune au moins une paire de mots que nous considérons comme non reliés. Quelques paires présentent des cas d'erreurs évidentes. Par exemple, *« **chro**me »* / *« **chro**nique »* produit par la règle « *me/nique* » apprise sur la paire de synonymes *« polyembryo**me** »* / *« dysembryome malin polyembryo**nique** »*; *« **pla**ce »* / *« **pla**t »* a été créé par la règle « *t/ce* » apprise sur *« absen**t** »* / *« absen**ce** »*. Notons que les erreurs pourraient être bloquées a posteriori par l'emploi de listes d'exception, comme dans l'analyseur morphologique de Fiametta Namer (Dal *et al.*, 1999) ou des contraintes syntaxiques (Corbin, 1987; Dal *et al.*, 1999).

D'autres cas concernent les mots qui ont en commun un *préfixe* plutôt qu'une racine. Ils sont tous formés sur des racines différentes auxquelles un même préfixe a été ajouté. Par exemple, le préfixe *« auto- »* dans *« **auto**greffe »* / *« **auto**logue »* / *« **auto**plastique »*, ou bien *« homo- »* dans *« **homo**greffe »* / *« **homo**logue »*. Dans la méthode rien ne permet actuellement de faire la différence entre de tels préfixes et les racines des mots.

4.4. Utilisation de l'outil d'aide au codage

Les familles morphologiques obtenues sont utilisées par un outil d'aide au codage dans le but d'augmenter son rappel. Le fonctionnement schématique de l'outil est le suivant. L'utilisateur saisit une expression et l'outil présente les codes de la CIM-10 qui contiennent le plus grand nombre de mots présents dans la requête. Avec le recours aux familles morphologiques utilisées dans une phase d'expansion de la requête, des mots qui ne figuraient pas dans sa requête sont retrouvés (par exemple, *« sténose aortique »* à la place de *« sténose de l'aorte »*). Le mécanisme de base remplace chaque mot de la requête qui appartient à une famille morphologique par la disjonction de tous les membres de cette famille. L'expansion améliore le rappel. Le risque réside dans la substitution de mots sémantiquement éloignés, ce qui diminue la précision. Cet outil de codage est un prototype, et nous avons encore peu de données sur son utilisation. Néanmoins, sur un jeu de 220 requêtes saisies par des utilisateurs variés, nous avons constaté une augmentation du rappel de 12 % et une diminution de la précision de 2,5 % lorsque

cette expansion de requête est mise en fonction.

5. Discussion et perspectives

Cette méthode demande très peu de connaissances linguistiques a priori. Les seules hypothèses faites sont que (i) la segmentation du mot en racine (appelée préfixe commun) et suffixe est pertinente; (ii) la définition d'une longueur minimale de la racine réduit suffisamment le bruit et ne cause pas de silence (la longueur a été fixée à trois caractères, elle peut être modifiée), et (iii) les bases nulles sont éliminées lors de l'étape de génération de mots avec les règles trouvées. Ces hypothèses sont sans doute vraies pour d'autres langues ; nous comptons les tester sur la version anglaise et la version russe du Microglossaire SNOMED.

On aurait également pu prendre en compte, comme un relecteur nous l'a fait remarquer, une contrainte supplémentaire : que dans chaque décomposition « préfixe + suffixe » obtenue, le suffixe lui aussi puisse se combiner avec d'autres préfixes. Dans l'expérience décrite ici, c'était le cas pour 50 % des suffixes (91 % des occurrences). Il reste à évaluer de quelle façon une prise en compte de cette contrainte influencerait les résultats, en termes de bruit et de silence.

L'évaluation du bruit dépend beaucoup de la tâche visée. Si une équivalence sémantique stricte est requise, il peut être utile de séparer les suffixes grammaticaux des suffixes sémantiques, ce que notre méthode ne fait pas. La distinction des suffixes sémantiques pourrait s'appuyer sur l'identification de mots connus utilisés comme suffixes (par exemple, la racine « *-blastome* » est un mot en soi). Si, par contre, la proximité sémantique n'est pas une nécessité, comme par exemple dans la recherche d'information, notre méthode, avec ses 95 % de familles correctes, apporte une ressource utilisable. Dans tous les cas, il serait intéressant de comparer le bruit généré par notre méthode avec celui obtenu sans contrainte de départ avec une liste de mots (contexte de séries de synonymes dans notre cas). Ceci pourrait être testé avec l'utilitaire `findaffix` d'Unix . Nous pensons que le niveau de bruit très bas produit par notre méthode est lié au contexte de départ (séries de synonymes).

La segmentation de mots en racine et suffixe, avec une longueur minimale fixée pour la racine, limite les possibilités de génération de règles morphologiques. Premièrement, une décomposition en préfixe et radical pourrait aussi être pertinente. La même méthode peut être utilisée dans ce but, en l'appliquant sur des mots inversés. De premières expériences montrent que la plupart des préfixes sont détectés. La prise en compte de ces nouvelles données peut aider à l'élimination des erreurs liées à l'ajout de préfixes (voir section 4.3). Deuxièmement, des altérations plus complexes comme *« détruire »* / *« destruction »* n'ont pas été trouvées conformément à notre hypothèse de segmentation de base : mot = base + suffixe. Et, enfin, pour la même raison, les cas d'ajout de préfixe et de variation du suffixe (ex : *« strangulation »* / *« étranglement »*) ne peuvent pas être reconnus. Une approche récursive pourrait constituer une piste pour combler ces limites, en effectuant une décomposition en morphèmes minimaux.

Nous nous sommes concentrés sur la construction de familles de mots morphologiquement reliés. L'expansion de requête peut aussi se faire à l'aide de mots reliés sémantiquement qui n'ont pas (ou très peu) de ressemblance morphologique : les synonymes SNOMED comme *« coeur »* / *« cardiaque »* ou *« noyau »* / *« nucléaire »* sont pertinents, mais ne sont pas relevés par notre méthode. Ils peuvent aussi être extraits directement de la SNOMED.

La même méthode, lancée dans un contexte bilingue approprié, pourrait permettre de construire un dictionnaire de primitives morphosémantiques (« cognates ») communes aux langues étu-

diées. Les données pour le faire sont directement disponibles, dans la mesure où une traduction terme à terme du Microglossaire SNOMED existe dans plusieurs langues.

Nous avons utilisé les séries de synonymes trouvées dans un thésaurus : le Microglossaire de Pathologie SNOMED. D'autres séries de synonymes peuvent être extraites d'autres terminologies médicales, en français ou dans d'autres langues (par exemple, l'UMLS, la plus grande union de terminologies biomédicales, avec un million de termes, essentiellement anglais). Nous avons vu que d'autres contraintes de départ peuvent être utilisées pour amorcer l'acquisition : corpus uniquement (Xu & Croft, 1998), liste de mots étiquetés (Dal *et al.*, 1999), à la fois corpus et thésaurus (Jacquemin, 1997). Le travail sur corpus seul pourrait aussi se fonder sur l'identification d'expressions paraphrastiques : en recherchant dans un corpus étiqueté des paires de contextes comme « N_1 *prep* N_2 » / « A N_1 » où N_2 et A sont morphologiquement proches, on aurait une autre source de mots reliés morphologiquement.

Les méthodes d'acquisition de familles morphologiques examinées ont toutes besoin pour fonctionner d'exemples attestés de formes variantes. Certaines les cherchent dans des corpus (Xu & Croft, 1998), d'autres dans des thésaurus (ce travail), d'autres encore en contrastant les deux (Jacquemin, 1997). Un avantage potentiel de l'emploi de corpus est le volume de données examiné ; mais l'important est en réalité la variété des formes attestées dans le corpus considéré. Ainsi, dans un domaine spécialisé, on aura sans doute intérêt à constituer un corpus formé de textes de genres variés, plutôt qu'un corpus homogène sur ce plan. Un avantage que l'on peut attendre d'un thésaurus est de concentrer la plupart des termes pertinents d'un domaine. Mais tout dépend alors de la couverture du thésaurus, et de la systématicité de l'inclusion de synonymes en plus des termes préférentiels. La nomenclature SNOMED est considérée comme proposant la meilleure couverture du domaine clinique (Chute *et al.*, 1996), et est riche en synonymes, y compris adjectivaux; mais sa version française, restreinte au Microglossaire (soit un dixième de la SNOMED), est nettement plus limitée. Une approche contrastant thésaurus et corpus est une solution intéressante pour collecter une variation importante, dans la mesure où les formes attestées dans le corpus peuvent compléter et s'opposer à celles du thésaurus.

6. Conclusion

Nous avons présenté une méthode automatique de constitution de familles de mots morphologiquement reliés. Elle dépend de la présence de termes synonymes dans des thésaurus existants. L'expérience réalisée avec la version française du Microglossaire de Pathologie SNOMED a montré une précision de 95 %. Comme des connaissances linguistiques a priori ne sont pas nécessaires, la méthode et les outils peuvent être appliqués à d'autres thésaurus (nous les avons testés sur le thésaurus Agrovoc de la FAO) ainsi qu'à d'autres langues (nous les avons testés sur les versions anglaise et russe de la SNOMED). Cette méthode peut encore être améliorée pour prendre en compte les préfixes et pousser plus loin la décomposition en morphèmes, et en employant un formalisme linguistique plus élaboré (Theron & Cloete, 1997). Des méthodes automatiques comme celle-ci peuvent aussi être considérées comme des outils d'aide à la modélisation manuelle : les familles obtenues automatiquement peuvent constituer une information très utile pour un linguiste désirant modéliser plus finement les phénomènes en jeu.

Remerciements

Nous remercions le Dr. Roger A. Côté de nous avoir gracieusement prêté une copie précommerciale de la version française du Microglossaire de Pathologie SNOMED, Jean Royauté

(URI / INIST) pour le thésaurus Agrovoc, Christian Jacquemin pour ses remarques concernant une version précédente de cet article, et les relecteurs anonymes de la conférence.

Références

ANTWORTH E. L. (1990). *PC-KIMMO: a two-level processor for morphological analysis*. Number 16 in Occasional Publications in Academic Computing. Dallas, TX: Summer Institute of Linguistics.

CHUTE C. G., COHN S. P., CAMPBELL K. E., OLIVER D. E. & CAMPBELL J. R. (1996). The content coverage of clinical classifications. *J Am Med Inform Assoc*, **3**(3), 224–233. for the Computer-Based Patient Record Institute's Work Group on Codes and Structures.

CORBIN D. (1987). *Morphologie dérivationnelle et structuration du lexique*. Lille: Presse universitaire de Lille.

CÔTÉ R. A. (1996). *Répertoire d'anatomopathologie de la SNOMED internationale, v3.4*. Université de Sherbrooke, Sherbrooke, Québec.

DAL G., HATHOUT N. & NAMER F. (1999). Peut-on construire un lexique dérivationnel : théorie et réalisations. In P. AMSILI, Ed., *Actes de TALN 1999*, Cargèse. Ce volume.

DÉJEAN H. (1998). Morphemes as necessary concept for structures discovery from untagged corpora. In *Workshop on Paradigms and Grounding in Natural Language Learning*, p. 295–299, Adelaide.

DUJOLS P., AUBAS P., BAYLON C. & GRÉMY F. (1991). Morphosemantic analysis and translation of medical compound terms. *Methods Inf Med*, **30**, 30–35.

JACQUEMIN C. (1997). Guessing morphology from terms and corpora. In *Actes, 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97)*, p. 156–167, Philadelphia, PA.

KOSKENNIEMI K. (1983). *Two-level morphology: a general computational model for word-form recognition and production*. PhD thesis, University of Helsinki Department of General Linguistics, Helsinki.

LOVIS C., MICHEL P.-A., BAUD R. & SCHERRER J.-R. (1995). Word segmentation processing: a way to exponentially extend medical dictionaries. In R. A. GREENES, H. E. PETERSON & D. J. PROTTI, Eds., *Proc 8 th World Congress on Medical Informatics*, p. 28–32.

MCCRAY A. T., SRINIVASAN S. & BROWNE A. C. (1994). Lexical methods for managing variation in biomedical terminologies. In *Proc Eighteenth Annu Symp Comput Appl Med Care*, p. 235–239, Washington: Mc Graw Hill.

MOESCHLER J. & AUCHLIN A. (1997). *Introduction à la linguistique contemporaine*. Paris: Armand Colin, Masson.

PACAK M. G., NORTON L. M. & DUNHAM G. S. (1980). Morphosemantic analysis of -ITIS forms in medical language. *Methods Inf Med*, **19**, 99–105.

SPYNS P. (1994). A robust category guesser for Dutch medical language. In *Proceedings of ANLP 94 (ACL)*, p. 150–155.

THERON P. & CLOETE I. (1997). Automatic acquisition of two-level morphological rules. In *ANLP97*, p. 103–110, Washington, DC.

WINGERT F., ROTHWELL D. & CÔTÉ R. A. (1989). Automated indexing into SNOMED and ICD. In J. R. SCHERRER, R. A. CÔTÉ & S. H. MANDIL, Eds., *Computerised Natural Medical Language Processing for Knowledge Engineering*, p. 201–239. Amsterdam: North-Holland.

XU J. & CROFT B. W. (1998). Corpus-based stemming using co-occurrence of word variants. *ACM Transactions on Information Systems*, **16**(1), 61–81.