# Analyse morphosémantique des composés savants : transposition du français à l'anglais

Louise DELÉGER[1], Fiammetta NAMER[2], Pierre ZWEIGENBAUM[3,4]
[1] INSERM, UMR_S 872, Éq. 20, Les Cordeliers, 75006 Paris
Université Pierre et Marie Curie-Paris6, UMR_S 872, 75006 Paris
Université Paris Descartes, UMR_S 872, 75006 Paris
[2] ATILF et Université Nancy 2, CLSH, 54015 Nancy
[3] CNRS, UPR3251, LIMSI, 91403 Orsay
[4] INALCO, CRIM, 75343 Paris Cedex 07
louise.deleger@spim.jussieu.fr,
fiammetta.namer@univ-nancy2.fr, pz@limsi.fr

**Résumé.** La plupart des vocabulaires spécialisés comprennent une part importante de lexèmes morphologiquement complexes, construits à partir de racines grecques et latines, qu'on appelle « composés savants ». Une analyse morphosémantique permet de décomposer et de donner des définitions à ces lexèmes, et semble pouvoir être appliquée de façon similaire aux composés de plusieurs langues. Cet article présente l'adaptation d'un analyseur morphosémantique, initialement dédié au français (DériF), à l'analyse de composés savants médicaux anglais, illustrant ainsi la similarité de structure de ces composés dans des langues européennes proches. Nous exposons les principes de cette transposition et ses performances. L'analyseur a été testé sur un ensemble de 1299 lexèmes extraits de la terminologie médicale WHO-ART : 859 ont pu être décomposés et définis, dont 675 avec succès. Outre une simple transposition d'une langue à l'autre, la méthode montre la potentialité d'un système multilingue.

**Abstract.** Medical language, as many technical languages, is rich with morphologically complex words, many of which take their roots in Greek and Latin – in which case they are called neoclassical compounds. Morphosemantic analysis can help generate decompositions and definitions of such words, and is likely to be similarly applicable to compounds from different languages. This paper reports work on the adaptation of a morphosemantic analyzer dedicated to French (DériF) to analyze English medical neoclassical compounds, and shows the similarity in structure of compounds from related European languages. It presents the principles of this transposition and its current performance. The analyzer was tested on a set of 1,299 compounds extracted from the WHO-ART terminology: 859 could be decomposed and defined, 675 of which successfully. Aside from simple transposition from one language to another, the method also emphasizes the potentiality for a multilingual system.

**Mots-clés :** analyse morphosémantique, composition savante, terminologie médicale.

**Keywords:** morphosemantic analysis, neo-classical compounding, medical terminology.

# 1 Introduction

La plupart des vocabulaires spécialisés, et en particulier le vocabulaire médical, comprennent une part importante de lexèmes morphologiquement complexes, construits à partir de racines grecques et latines, qu'on appelle « composés savants ». Segmenter ces composés en lexèmes de base est la tâche de l'analyse morphologique. Lorsque celle-ci inclut à la fois une partie formelle et une partie sémantique, on parle d'analyse morphosémantique. Ce type d'analyse est particulièrement adapté aux composés savants, où le sens est souvent « compositionnel », c'est-à-dire qu'il est la combinaison au moins partielle du sens des composants du lexème complexe. L'analyse morphosémantique est donc utile pour les méthodes intéressées par la sémantique, comme la génération de définitions ou la détection de termes similaires.

Il a de plus été observé que la structure morphologique des lexèmes composés savants est similaire dans de nombreuses langues européennes (Iacobini, 2003). Il semble donc possible d'appliquer une analyse linguistique dédiée aux composés savants d'une langue à d'autres langues proches. (Namer, 2005a) l'a montré pour un certain type de composés médicaux en proposant une analyse des noms de pathologies (comme HYPERCALCIURIE) pouvant être appliquée au français, à l'allemand, à l'espagnol, à l'italien et à l'anglais. L'analyse morphosémantique de tels composés montre ainsi un potentiel multilingue.

Dans le domaine médical, plusieurs travaux se sont intéressés à l'analyse de ces lexèmes complexes. Les premiers se concentrent sur un type particulier de règles de formation des lexèmes, comme les règles de suffixation en -ITIS (Pacak *et al*., 1980) ou -OSIS (Dujols *et al*., 1991), puis élargissent leur champs d'analyse (Wolff, 1984). (Lovis *et al*., 1995) décomposent les termes médicaux en introduisant la notion de morphosémantèmes, unités ne pouvant être décomposées sans perdre leur sens original. On trouve une notion similaire dans le système Morphosaurus (Schulz *et al*., 1999; Markó *et al*., 2005). Cet outil, qui ne se limite pas aux composés savants, est l'un des rares fonctionnant en multilingue ; il ne va cependant pas jusqu'à l'interprétation sémantique. (Iavindrasana *et al*., 2006) utilisent un outil statistique de segmentation morphologique (Creutz *et al*., 2005) pour mettre en correspondance les termes similaires d'une terminologie médicale (WHO-ART). L'outil DériF (Namer & Zweigenbaum, 2004) effectue une analyse morphosémantique des lexèmes dérivés ou composés français et produit une décomposition hiérarchique (par opposition à (Markó *et al*., 2005) ou (Lovis *et al*., 1995) où la segmentation reste linéaire) ainsi qu'une définition sémantique des lexèmes et un ensemble de lexèmes sémantiquement apparentés (relations d'hyponymie ou d'équivalence, par exemple). Son potentiel pour une application multilingue a été souligné dans (Namer, 2005b).

Cet article présente nos travaux concernant l'adaptation de DériF aux composés savants médicaux anglais[1]. Notre but est de transposer l'analyse par DériF des composés français à l'anglais, afin d'illustrer la similarité des mécanismes des composés savants dans des langues européennes proches et d'obtenir un outil qui fait défaut sur l'anglais, que nous baptisons DériA. Ce travail peut en outre être vu comme une première étape vers un système multilingue.

Nous posons dans un premier temps les éléments théoriques sous-jacents à ce travail, puis décrivons l'analyseur morphosémantique et notre liste de lexèmes test. Nous expliquons ensuite les modifications effectuées sur le système et son mode d'évaluation. Nous exposons nos résultats et discutons la méthode, puis concluons avec quelques perspectives.

[1] Cet article actualise une précédente version de nos travaux soumise à la conférence MEDINFO 2007.

# 2 Eléments théoriques

La base de ce travail est l'analyse morphosémantique, c'est-à-dire une analyse morphologique associée à une interprétation sémantique. Nous cherchons à lier le lexème d'entrée à sa base (en cas de dérivation) ou à ses composants (en cas de composition). La décomposition est associée à une description du sens du lexème complexe basée sur le sens de ses composants. Un lexème complexe est formé à partir de n'importe quelle combinaison des deux règles suivantes :

– la **dérivation** qui se manifeste formellement par l'ajout d'un affixe (préfixe ou suffixe) à une base ; par exemple : $\text{BOSSE}_{\text{NOM}}$ / $\text{BOSSU}_{\text{ADJ}}$ ; $\text{GRIPPE}_{\text{NOM}}$ / $\text{ANTIGRIPPE}_{\text{ADJ}}$
– la **composition** consiste à former un lexème en combinant deux composants, qui peuvent être chacun soit des lexèmes de la langue moderne, soit des racines grecques et latines appelées éléments de formation (EF) ; par exemple THERMORÉGULATION, ARTHRALGIE.

Nous avons choisi ici de traiter les composés savants anglais (formés à partir d'EF). Cependant, comme nous l'avons expliqué, un lexème complexe peut avoir été formé à la fois par composition et par dérivation. Un composé peut donc être la base d'une règle de dérivation ; de même, un lexème dérivé peut être l'un des composants d'une règle de composition. C'est pourquoi nous devons prendre en compte les lexèmes composés et dérivés (comme HAEMORRHAGIC).

Notre hypothèse de travail pour la transposition de l'analyse du français vers l'anglais est que l'on peut appliquer une même analyse linguistique aux composés savants de plusieurs langues. En effet, nous supposons que ces composés sont formés de la même façon et que les composants mis en jeu sont similaires, les principales différences étant orthographiques (par exemple, ALGIE en français et ALGIA en anglais). Les difficultés éventuelles se situeraient :

– au niveau de l'ordre combinatoire des composants : l'analyse ne fonctionnera pas si l'ordre n'est pas le même dans les deux langues. Ce cas devrait être peu fréquent, car il s'agit ici de composition classique qui reprend l'ordre latin ou grec ;
– au niveau des composants eux-mêmes : la correspondance des analyses ne peut se faire que si les lexèmes anglais et français sont bien dans les deux cas formés à partir d'EF. C'est notre hypothèse et l'analyse pourra se faire car ces EF sont listés et en nombre limité.
– au niveau de la jointure des composants : la modification du premier composant lors de sa combinaison avec le deuxième (allomorphie), par exemple le rajout de l'élément de jointure *-o-*. Si ces phénomènes sont très différents, il peut y avoir des problèmes dans l'analyse. Nous supposons qu'ils sont similaires, mis à part quelques modifications orthographiques ;
– au niveau des processus morphologiques de suffixation et préfixation appliqués aux composés savants. En effet, les affixes ne sont pas les mêmes dans les deux langues. Cependant, nous supposons que, dans le cas des composés savants, il est suffisant de remplacer les affixes français par des affixes anglais de même « classe » (par exemple, les suffixes formateurs d'adjectifs relationnels français peuvent être remplacés par leurs homologues anglais).

# 3 Matériel de départ

## 3.1 L'analyseur morphosémantique DériF

Nous nous basons sur la version française de l'analyseur morphosémantique DériF (« Dérivation en Français »), destiné aussi bien à la langue générale qu'à des vocabulaires plus spécialisés

| | | |
|---|---|---|
| arthralgia/NOM | cardiomegaly/NOM | crystalluria/NOM |
| atelectasis/NOM | cerebellar/ADJ | dermatomyositis/NOM |
| blepharospasm/NOM | claustrophobia/NOM | dextrocardia/NOM |
| calcinosis/NOM | clostridial/ADJ | dorsal/ADJ |
| capillary/ADJ | cryptococcal/ADJ | dysmenorrhea/NOM |

TAB. 1 – Extrait de la liste de lexèmes construits contenant un élément de formation

tels que la langue médicale. L'analyse effectuée est purement linguistique et implémente un certain nombre de règles de décomposition ainsi que de schémas d'interprétation sémantique des lexèmes. Les ressources nécessaires à l'outil comprennent des lexiques de lexèmes lemmatisés et étiquetés morphosyntaxiquement et une table des éléments de formation. Lorsque DériF est appliqué à un vocabulaire biomédical, le système ne se limite pas à une simple décomposition et interprétation sémantique, mais produit également un ensemble de lexèmes lexicalement reliés.

Le système prend en entrée une liste de lexèmes étiquetés et lemmatisés et produit en sortie :

1. une représentation ordonnée des règles d'analyse qui se sont appliquées successivement ;
2. une définition (ou glose) du lexème en langue naturelle ;
3. une catégorie conceptuelle, inspirée des descripteurs principaux du thésaurus MeSH (anatomie, organisme, maladie, etc.) ;
4. un ensemble de lexèmes qui sont potentiellement liés lexicalement au lexème d'entrée. Les différentes relations identifiées sont l'équivalence (*eql*), l'hyponymie (*isa*) et la relation de proximité sémantique « voir-aussi » (*see*).

Un exemple d'analyse morphosémantique pour le lexème ACRODYNIE est donné ci-dessous (*N* signifie *nom* et *N** est associé à un EF nominal).

ACRODYNIE/N ==> (1) [ [ acr N* ] [ odyn N* ] ie N ]
(2) douleur (de–lié(e) à) articulation
(3) maladie
(4) eql : acr/algie, eql : apex/algie, see : acr/ite, see : apex/ite

## 3.2 Le corpus de test : la terminologie WHO-ART

Pour tester DériA, notre transposition de DériF à l'anglais, une liste de lexèmes a été sélectionnée. Ceux-ci ont été extraits de la terminologie WHO-ART, qui décrit les effets secondaires des médicaments, car une des applications visées par ce travail est d'apporter une contribution au domaine de la pharmacovigilance. Nous avons segmenté les termes anglais en lexèmes, car DériF traite des lexèmes simples et non des unités polylexicales. Parmi ces lexèmes, nous n'avons retenu que les composés savants (nous cherchons en effet à adapter DériF pour ce type de lexèmes). Cette sélection a été effectuée à la fois automatiquement, en rejetant tous les lexèmes de 4 caractères ou moins (ils ne sont quasiment jamais construits), et manuellement en parcourant la liste pour extraire les composés savants. Nous avons ainsi obtenu une liste de 1299 lexèmes sur un total de 3476. Comme expliqué précédemment, nous avons sélectionné à la fois les composés « purs » et les lexèmes dérivés contenant un EF. Un extrait de la liste est donné dans le tableau 1. Afin d'avoir une idée de la proportion des différents types de composés, nous avons calculé leur pourcentage sur un échantillon de 200 lexèmes. Nous avons 45 % de composés purs et 55 % de lexèmes dérivés.

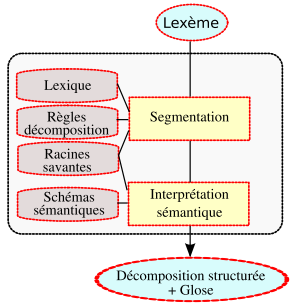


FIG. 1 – Analyse morphosémantique

Les lexèmes ont été préalablement étiquetés morphosyntaxiquement et lemmatisés avec l'étiqueteur Treetagger[2]. Un lexique de lexèmes étiquetés, extrait du Specialist Lexicon de l'UMLS[3] a été utilisé pour aider Treetagger à traiter les lexèmes inconnus.

# 4 Transposition de DériF en anglais

## 4.1 Modifications effectuées

Le mécanisme de l'analyse morphosémantique de DériF est schématisé sur la figure 1. Nous avons identifié plusieurs niveaux où l'analyse fait appel à des éléments spécifiques à la langue :

1. *Le lexique* de lexèmes étiquetés qui est utilisé par le système pour tester l'existence d'un composant et pour obtenir sa catégorie grammaticale.
2. *La table des EF*, où chaque élément est associé à un lexème en français moderne décrivant son sens, à un type conceptuel, à une catégorie grammaticale et à un ensemble d'éléments liés par des relations d'équivalence, d'hyponymie ou de proximité sémantique.
3. *Les règles de décomposition*, qui sont déclenchées dans un certain ordre selon la catégorie grammaticale du lexème et l'affixe identifié (s'il y en a un). Ces règles identifient la tête et les autres composants, en se basant sur la table des EF et le lexique. Chaque règle inclut des exceptions et des normalisations orthographiques sur les composants.
4. *Les schémas d'interprétation sémantique*, qui produisent des gloses à partir de la relation entre la tête du lexème et ses autres composants.

Nous avons apporté des modifications pour chacun de ces niveaux :

1. Nous avons remplacé le lexique par un lexique anglais du Specialist Lexicon de l'UMLS.
2. Nous partons de l'hypothèse que les EF sont les mêmes en français et en anglais, avec seulement de légères différences orthographiques. Nous avons donc effectué de petites

[2] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html`
[3] `http://www.nlm.nih.gov/pubs/factsheets/umlslex.html`

| EF | Anglais | Type | Relations lexicales |
|---|---|---|---|
| algia | pain | disease | odyn, algo, ~itis |
| blephar | eyelid | anatomy | palpebr, <-ocul, ~coro |
| ectomy | surgical excision | act | ~tomy, ~stomy |
| gastr | stomach | anatomy | stomac, gaster, ~hepat, ~entero, <-abdomino, ~pancreat |

TAB. 2 – Table des éléments de formation (extrait)

modifications sur les EF français (par exemple, retirer les accents, comme pour BLÉPHARO / BLEPHARO) pour obtenir les équivalents anglais. Les types conceptuels attribués à chaque EF auraient pu être conservés en français, mais nous avons décidé que les traduire en anglais serait plus adapté, ce qui a été fait aisément car ces types sont peu nombreux. Nous avons également associé des lexèmes en anglais moderne à chaque EF, que nous avons obtenus à partir de deux listes de racines savantes, une extraite du Specialist Lexicon de l'UMLS, l'autre du dictionnaire médical Dorland[4], les racines étant appariées avec leur équivalent en anglais moderne dans ces deux listes. Nous avons mis en correspondance automatiquement les EF et validé les résultats. Les EF pour lesquels aucun équivalent n'a pu être obtenu ont été traités manuellement. Enfin, l'ensemble des EF lexicalement liés ont été remplacés par leurs équivalents anglais (suivant les modifications orthographiques effectuées pour obtenir les EF anglais). Les catégories grammaticales ont été conservées. Un extrait de cette table est fourni dans le tableau 2. Les relations lexicales entre les éléments sont étiquetées comme suit : <- pour une relation d'hyponymie, ~ pour une relation voir-aussi, et aucun signe pour une relation d'équivalence.

3. Nous sommes très peu intervenus au niveau des règles de décomposition puisque notre hypothèse est que les composés savants sont formés de la même manière. Nous avons donc simplement adapté les exceptions, les normalisations orthographiques et les affixes (par exemple, le suffixe français -IQUE est remplacé par le suffixe anglais -IC).
4. Enfin, nous avons traduit les schémas d'interprétation sémantique afin que ceux-ci génèrent des gloses anglaises. Par exemple, le schéma suivant est associé au suffixe -IA (-IE en français) et au préfixe HYPER- avec les composants Y/X comme noms : *Affection of **X** linked to the excess of **Y*** (où X et Y sont remplacés par les équivalents modernes des EF ou par des lexèmes simples du lexique, comme par exemple HYPERCALCEMIA dont l'analyse produit *« Affection of blood linked to the excess of calcium »)*.

## 4.2 Évaluation

Nous avons testé DériA sur les 1299 lexèmes extraits de la terminologie WHO-ART. La sortie attendue pour chaque lexème est donc une décomposition hiérarchique des lexèmes, une définition en langage naturel, un type conceptuel et un ensemble de lexèmes liés. Nous avons évalué la couverture, la précision et le rappel de ces résultats. Nous avons défini la couverture comme la proportion de lexèmes que le système analyse, la précision comme le rapport entre le nombre d'analyses correctes et le nombre total d'analyses, et le rappel comme le rapport entre le nombre d'analyses correctes et le nombre d'analyses attendues (i.e. le nombre de lexèmes). Une analyse

[4] http://www.merckmedicus.com/pp/us/hcp/thcp_dorlands_content.jsp?pg=/ppdocs/us/common/dorlands/dorland/dmd-a-b-000.htm

est considérée valide si sa décomposition et sa définition sont correctes ; nous n'avons pas pris en compte les lexèmes lexicalement liés ni le type conceptuel dans cette évaluation.

La méthode de (Iavindrasana *et al*., 2006) est appliquée, comme la nôtre, à la terminologie WHO-ART, de sorte que nous mis en place une comparaison. Nous avons soumis la même liste de lexèmes à Morfessor (qui est l'outil utilisé dans ces travaux) dans les mêmes conditions que (Iavindrasana *et al*., 2006) et avons évalué couverture, précision et rappel.

# 5 Résultats

Au stade actuel de DériA (DériF en anglais), nous avons pu obtenir 859 analyses sur les 1299 lexèmes de notre liste (voir tableau 3), ce qui donne une couverture de 66 %. Des exemples d'analyses sont donnés dans le tableau 4. On observe que le système a pu analyser aussi bien de purs composés savants (*[ [arthr N*] [algia N*] N]*) que des lexèmes dérivés construits à partir d'une base savante (*[ a+ [ dactyl N*] +y N]*).

Nous avons identifié plusieurs causes pour lesquelles un lexème n'a pas pu être décomposé :

– Certaines règles de suffixation ne sont pas actuellement implémentées dans DériF et DériA : c'est le cas pour les suffixes -ATION et -ISM, les lexèmes tels que LACRIMATION et HERMAPHRODITISM ne sont donc pas décomposés ;
– Certains éléments ne sont pas présents dans la table des EF ni dans le lexique de lexèmes : par exemple CAMPT- n'est pas dans la table des EF et le lexème CAMPTODACTYLY n'a donc pas été décomposé ;
– Des erreurs sont survenues pendant le pré-traitement (c'est-à-dire des lexèmes mals étiquetés) : par exemple, CORPORAL a été étiqueté comme nom au lieu d'adjectif.

Nous avons mesuré une précision de 78,5 % (voir tableau 3) ce qui est assez satisfaisant. Etant donnée une couverture modérée, cela donne un rappel de 52 %. Les erreurs sont dues à :

– Une mauvaise structuration de la décomposition. On en voit un exemple dans le tableau 4 avec le lexème MENINGOENCEPHALITIS, dont la bonne décomposition devrait être comme suit : [ **[** [mening N*] [encephal N*] **]** [itis N*] N]. L'élément -ITIS devrait être la tête de la combinaison des deux EF MENING- et ENCEPHAL-, ce qui donnerait une définition telle que *« inflammation related to head and meninges. »*
– Une définition insatisfaisante, ce qui est souvent dû au fait que le sens du lexème n'est pas suffisamment reflété par celui de ses parties. Le lexème ACANTHOSIS (voir tableau 4) en est une illustration : sa décomposition est correcte mais son sens s'est opacifié et ne peut être dérivé de celui de ses composants. Il devrait être de nos jours analysé comme non construit.
– Une erreur d'étiquetage : les lexèmes mal étiquetés n'ont pas pu être correctement analysés. C'est le cas de ALVEOLAR (dernière ligne du tableau 4) qui a été traité comme un nom dérivé d'un adjectif (ALVEOLAR en tant qu'adjectif est, lui, correctement analysé par le système).

Aucun cas de bruit ni de silence ne semble être dû à une spécificité des composés anglais par

| Nb total de lexèmes | Lexèmes décomposés (couverture) | Nombre d'analyses correctes | Précision | Rappel |
|---|---|---|---|---|
| 1299 | 859 (66 %) | 675 | 78,5 % | 52 % |

TAB. 3 – Résultats de l'évaluation de DériA

| Lexème/Cat | Décomposition | Définition | Type | Lexèmes proches |
|---|---|---|---|---|
| arthralgia/N | [ [arthr N*] [algia N*] N] | pain (of–linked to) joint | disease | eql : arthr/algesia see : arthr/itis |
| adactyly/N | [ a+ [ dactyl N*] +y N] | Affection characterized by the absence of digit | disease | |
| gastroeso-phageal/ADJ | [ [gastr N*] [oesophag N*] al ADJ] | Related to oesophagus and stomach | anatomy | eql : stomac/oesophag isa : abdo-min/oesophag |
| meningoence-phalitis/ADJ | [ [mening N*] [ [encephal N*] [itis N*] N] N] | (Part of – Specific type of) encephalitis related to meninges | | |
| acanthosis/N | [ [acanth N*] [osis N*] N] | (Part of – Specific type of) disease related to prickle | disease | |
| alveolar/N | [ [alveolar A] N] | Entity being alveolar | | |

TAB. 4 – Exemples d'analyses de lexèmes effectuées par DériA (correctes puis incorrectes)

rapport aux composés français.

L'évaluation des résultats de l'analyse des lexèmes par Morfessor (méthode de (Iavindrasana *et al*., 2006)) donne une couverture de 93,7 %, une précision de 53,2 % et un rappel de 49,9 % (voir tableau 5). Nous avons également calculé la précision des deux outils sur leur intersection de couverture, soit 830 lexèmes, et avons obtenu 58,7% pour Morfessor et 78,2% pour DériA.

# 6 Discussion

La précision de notre système adapté (DériA) est encourageante, et pourrait être améliorée en identifiant les mots composés dont le sens s'est figé. Le rappel est plus bas mais devrait augmen-

| Nb total de lexèmes | Lexèmes décomposés (couverture) | Nombre d'analyses correctes | Précision | Rappel |
|---|---|---|---|---|
| 1299 | 1217 (93,7 %) | 648 | 53,2 % | 49,9 % |

TAB. 5 – Résultats de l'évaluation de Morfessor

ter rapidement en implémentant des règles supplémentaires. Les résultats sont similaires à ceux obtenus sur le français par (Namer & Baud, 2006). Le système ne génère pas plus d'analyses erronées que dans la langue d'origine. De plus, l'analyse des résultats (section 5) n'a pas relevé d'erreurs dues à une différence de mécanisme dans la formation des composés. Les problèmes potentiels énumérés (section 2) ne semblent pas s'être réalisés. Bien que notre corpus de test ne soit pas exhaustif, nous en concluons que ces difficultés sont rares et pourraient, le cas échéant, être traitées grâce à des règles d'exception. Ce système spécifique à une langue a donc pu être adapté avec succès à une autre langue pour l'analyse de composés médicaux.

L'utilisation d'un analyseur morphosémantique comme DériF basé sur des méthodes linguistiques présente en outre un certain nombre d'avantages. Le système effectue à la fois une décomposition morphologique et une interprétation sémantique, tandis que d'autres méthodes restent au niveau de la segmentation morphologique ou ajoutent de la sémantique après l'application d'un outil de décomposition. Nous produisons également une décomposition hiérarchique et non simplement linéaire. Les résultats obtenus en suivant la méthode de (Iavindrasana *et al*., 2006) ont donné une couverture bien plus grande qu'avec DériA, mais une précision très inférieure (20-25 % de moins) et un rappel légèrement moindre. Ceci montre qu'un nombre presque identique de lexèmes sont correctement analysés (rappel), mais que DériA est bien plus exact car il propose beaucoup moins d'analyses incorrectes. En se basant sur un analyseur statistique, la méthode de (Iavindrasana *et al*., 2006) a l'avantage d'être indépendante de la langue ; cependant, l'implémentation utilisée fait également usage d'une table de morphèmes, de sorte que la transposition vers une autre langue n'est pas immédiate non plus. De plus, la segmentation sur des bases statistiques cause également certains types d'erreurs qui peuvent être facilement évitées avec des règles linguistiques comme celles de DériA (par exemple, CHEMOSIS est segmenté en C+HEM+OSIS par Morfessor). De plus, comme nous l'avons souligné ci-dessus, DériA propose en sortie une information plus riche que Morfessor (simple segmentation linéaire).

Ce travail offre aussi la perspective d'un système capable de fonctionner avec plusieurs langues. Nous avons transposé le système du français à l'anglais, mais une prochaine étape pourrait être de faire fonctionner le système sur les deux langues (utiliser le même système aussi bien pour analyser des composés anglais que français) ou de l'utiliser pour la traduction : produire une définition anglaise d'un lexème français (et vice-versa). Il faudrait pour cela une table multilingue de racines (comme proposé dans (Namer, 2005b) et préparé ici pour l'anglais et le français), des schémas d'interprétation sémantique multilingues (comme obtenus ici pour ces deux langues) et une traduction des lexèmes du lexique. Un tel système pourrait, par exemple, contribuer à la recherche d'information translingue, avec les mêmes principes que dans (Markó *et al*., 2005).

# 7 Conclusion

Dans cet article, nous avons montré comment transposer un analyseur morphosémantique basé sur des règles linguistiques depuis le français vers l'anglais afin d'analyser des composés savants médicaux. Ceci vérifie l'hypothèse que les composés savants de différentes langues peuvent être analysés de la même façon sur le cas du voculaire médical de deux langues de type romane (le français) et germanique (l'anglais). La méthode peut être appliquée sur les composés savants d'autres vocabulaires spécialisés et devrait donner des résultats similaires. Ce travail constitue également un premier pas vers la création d'un système multilingue, qu'on obtiendrait en appliquant la méthode à d'autres langues, tâche plus ou moins aisée selon la proximité des langues (le français et l'anglais étant proches, on peut supposer que de nouvelles difficultés se poseront

avec d'autres langues plus éloignées).

D'un point de vue directement applicatif, nous disposons désormais d'un système fonctionnant sur l'anglais que nous pouvons utiliser dans le domaine de la pharmacovigilance sur les termes de la terminologie WHO-ART afin de mesurer leur proximité, ce qui constituerait une alternative à la méthode proposée par (Iavindrasana *et al*., 2006).

## Références


CREUTZ M., LAGUS K., LINDEN K. & VIRPIOJA S. (2005). *Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0*. Rapport interne, Computer and Information Science, Helsinki University of Technology.

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

IACOBINI C. (2003). Composizione con elementi neoclassici. In M. GROSSMAN & F. RAINER, Eds., *La formazione delle parole in italiano*, p. 69–96. Tübingen : Niemeyer.

IAVINDRASANA J., BOUSQUET C. & JAULENT M.-C. (2006). Knowledge acquisition for computation of semantic distance between WHO-ART terms. In *Stud Health Technol Inform*, p. 839–44.

LOVIS C., MICHEL P., BAUD R. & SCHERRER J. (1995). Word segmentation processing : a way to exponentially extend medical dictionaries. *Methods of Information in Medicine*, p. 28–32.

MARKÓ K., SCHULZ S. & HAHN U. (2005). Morphosaurus - design and evaluation of an interlingua-based, cross-language document retrieval engine for the medical domain. *Methods of Information in Medicine*, p. 537–545.

NAMER F. (2005a). Guessing the meaning of neoclassical compounds within LG : the case of pathology nouns. In *Proceedings of Generative Approaches to the Lexicon*, p. 175–84, Geneva.

NAMER F. (2005b). Morphosémantique pour l'appariement de termes dans le vocabulaire médical : approche multilingue. In M. JARDINO, Ed., *Actes de TALN 2005 (Traitement automatique des langues naturelles)*, p. 63–72, Dourdan : ATALA LIMSI.

NAMER F. & BAUD R. (2006). Defining and relating biomedical terms : towards a cross-language morphosemantics-based system. *Int J Med Inform*.

NAMER F. & ZWEIGENBAUM P. (2004). Acquiring meaning for French medical terminology : contribution of morphosemantics. In M. FIESCHI, E. COIERA & Y.-C. J. LI, Eds., *MEDINFO*, p. 535–539, San Francisco.

PACAK M., NORTON L. & DUNHAM G. (1980). Morphosemantic analysis of -itis forms in medical language. *Methods of Information in Medicine*, **19**, 99–105.

SCHULZ S., ROMACKER M., FRANZ P., ZAISS A., KLAR R. & HAHN U. (1999). Towards a multilingual morpheme thesaurus for medical free-text retrieval. In *Proceedings of MIE'99*, Ljubliana, Slovenia : IOS Press.

WOLFF S. (1984). The use of morphosemantic regularities in the medical vocabulary for automatic lexical coding. *Methods of Information in Medicine*, **4**(23), 195–203.