Morphosémantique pour l'appariement de termes dans le vocabulaire médical : approche multilingue

Fiammetta Namer

UMR "ATILF" CNRS & Université Nancy2
CLSH - BP3397 - 54015 Nancy Cedex - Fiammetta.Namer@univ-nancy2.fr

RESUME : *Cet article s'intéresse à la manière dont la morphosémantique peut contribuer à l'appariement multilingue de variantes terminologiques entre termes. L'approche décrite permet de relier automatiquement entre eux les noms et adjectifs composés savants d'un corpus spécialisé en médecine (synonymie, hyponymie, approximation). L'acquisition de relations lexicales est une question particulièrement cruciale lors de l'élaboration de bases de données et de systèmes de recherche d'information multilingues. La méthode est applicable à au moins cinq langues européennes dont elle exploite les caractéristiques morphologiques similaires des mots composés dans les langues de spécialité. Elle consiste en l'interaction de trois dispositifs : (1) un analyseur morphosémantique monolingue, (2) une table multilingue qui définit des relations de base entre les racines gréco-latines des lexèmes savants, (3) quatre règles indépendantes de la langue qui infèrent, à partir de ces relations de base, les relations lexicales entre les lexèmes contenant ces racines. L'approche est implémentée en français, où l'on dispose d'un analyseur morphologique capable de calculer la définition de mots construits inconnus à partir du sens de ses composants. Le corpus de travail est un lexique spécialisé médical d'environ 29000 lexèmes, que le calcul des relations de synonymie, hyponymie et approximation a permis de regrouper en plus de 3000 familles lexicales.*

ABSTRACT: *This paper addresses the issue of the interaction between morphosemantics and term variants extraction. The described method enables neoclassical compound nouns and adjectives of a biomedical specialized corpus to be automatically related by synonymy, hyponymy and approximation links. Acquiring lexical relations is a particularly crucial issue when elaborating multilingual databases and when developing cross-language information retrieval systems. This method can be applied at least to five European languages and exploits the similarity between the morphological characteristics of compound words in specialized domains. It requires the interaction of three techniques: (1) a language-specific morphosemantic parser, (2) a multilingual table defining basic relations between word roots, and (3) a set of language-independant rules to draw up the list of related terms. This approach has been fully implemented for French, on an about 29,000 terms biomedical lexicon, resulting to more than 3,000 lexical families.*

MOTS CLES : morphologie, sémantique, multilinguisme, composition savante, relation lexicale, terminologie médicale

KEYWORDS: morphology, semantics, multilingualism, neoclassical compounding, lexical relation, medical terminology

1. Introduction

1.1. *Variation terminologique et relations lexicales entre termes*

Dans le domaine bio-médical, comme dans toute langue de spécialité, l'extraction de variantes terminologiques constitue un enjeu important (Bourigault *et al.*, 2000b; Bourigault *et al.*, 2001; Bourigault *et al.*, 2004). L'objectif que la démarche présentée ici vise à atteindre, est de montrer en quoi la morpho-sémantique et la terminologie peuvent collaborer, dans le but d'optimiser l'appariement terminologique (bilingue, voire translinguistique) en corpus et la recherche et extraction de variantes de termes à partir des ressources disponibles dans les thésaurus. Le but est l'enrichissement des relations entre termes dans les bases multilingues de connaissances. L'interrogation de ressources hétérogènes (bases de données, notices bibliographiques etc.) dans plusieurs langues est une préoccupation constante dans les domaines de spécialité, qui a conduit au développement de plusieurs techniques pour l'établissement de terminologies multilingues. L'extraction terminologique et l'alignement de corpus parallèles (Bourigault *et al.*, 1999; Gaussier, 2001; Hull, 2001), sont deux étapes classiques dans la conception de tels systèmes (voir l'expérience de (Tran *et al.*, 2003)). En matière de synergie entre terminologie et morphologie, différentes études et applications existent.

Certaines se basent sur la reconnaissance de séquences au moyen de patrons (Jacquemin *et al.*, 1999; Bourigault *et al.*, 2000b; Daille, 2001) d'autres utilisent plutôt des systèmes statistiques fondés sur l'apprentissage de règles (Grabar *et al.*, 2000; Hathout, 2003). En particulier, (Zweigenbaum *et al.*, 2003 (à paraître)) et (Grabar, 2004) se sont intéressés à l'utilisation des règles morphologiques apprises pour enrichir le réseaux de liens que tissent entre eux les termes dans un domaine spécialisé, en l'occurrence, la biomédecine[1]. Avant eux, d'autres auteurs (à partir notamment de (Krovetz, 1993)) se sont servi de connaissances morphologiques apprises en corpus pour retrouver des relations entre termes, ou pour calculer des nouvelles variantes terminologiques ; d'autres enfin ont procédé de la façon inverse, à savoir acquérir des traits morphologiques à partir de données terminologiques structurées (Jacquemin, 1997; Xu *et al.*, 1998; Zweigenbaum *et al.*, 1999). Enfin, des travaux ont été menés dans le but de faire coopérer morphologie et terminologie bilingue, entre autre par (Chiao *et al.*, 2003).

[1] Parmi les relations lexicales acquises, citons par exemple (Grabar *et al.*, 2004), qui calculent automatiquement des relations dites transversales entre les termes, et (Grabar, 2005) qui réutilise des liens synonymiques pour les adapter en terminologie médicale.

1.2. *Morphologie constructionnelle au service de la terminologie biomédicale*

L'approche présentée se situe plutôt dans la lignée des systèmes basés sur l'application de contraintes. L'analyseur morphologique DériF[2] (Namer, 2003), qui constitue l'une des étapes du système, a été récemment adapté pour l'analyse du vocabulaire bio-médical, dans le cadre des projets UMLF et Vumef[3].

En guise de préambule au rôle original joué par DériF en matière de synergie entre morphologie et terminologie, voici les principaux aspects de l'appariement terminologique où l'analyse automatique en morphologie peut apporter une aide non négligeable. Il est possible d'attribuer à l'analyse morphologique quatre types de contributions ; et DériF est à même d'assumer ces quatre fonctions (cf. §3.1) :

1. Association de termes reliés morphologiquement : c'est évidemment ce vers quoi tend toute application de repérage de variantes terminologiques à base de connaissances morphologiques. Cette fonction consiste à établir le lien entre termes comme en (1), mais aussi comme en (2) : or pour ce dernier exemple, le lien n'est pas identifiable formellement.

(1) bactérienADJ / bactérieNOM
(2) hépatiqueADJ / foieNOM

2. Définition automatique des termes : dés lors qu'un terme est représenté par un lexème morphologiquement construit, il en existe une définition : (3), (4), (5) qui le relie à sa base (ou à ses constituants, si le terme est construit par composition). Le programme de calcul du rapport de sens doit être suffisamment souple pour rendre compte de situations apparemment irrégulières, ce que montre le rapport entre les exemples (3) et (4), et les exemples (5) et (6). Enfin, l'intérêt de cette fonction est de deviner automatiquement le sens des mots inconnus dés lors qu'ils sont conformes à une règle morphologique : (7) et (8).

(3) bactérien : « relatif au(x) bactérie(s) »
(4) antibactérien : « contre le(s) bactérie(s) »
(5) hépatique : « relatif au(x) foie(s) »
(6) antihépatique : « contre l'hépatite »
(7) schtroumpfien : « relatif au(x) schtroumpf(s) »
(8) antischtroumpfien : « contre le(s) schtroumpf(s) »

---

[2] DériF a été conçu et développé dans le cadre du projet ACI MorTAL (G. Dal, CNRS) ; voir (Dal *et al.*, 2004).

[3] Le projet UMLF « Lexique Médical Francophone Unifié », programme ACI du MENRT 2002-2004, a été coordonné par P. Zweigenbaum STIM/DPA/DSI/AP-HP INSERM, Paris), voir (Zweigenbaum *et al.*, 2005) ; le projet RNTS du MENRT VumeF « Vocabulaire Unifié Médical Français », 2003-2005, est coordonné par S. Darmoni (CHU Rouen, et équipe CISMeF, L@STICS) et J-F. Forget (Vidal SA), voir (Darmoni *et al.*, 2003).

3. Regroupement des termes d'un corpus reliés lexicalement par synonymie, hyponymie ou approximation. Ces relations lexicales intéressent tous les lexicologues, sémanticiens, logiciens et philosophes (parmi les auteurs en ayant proposé des critères formels définitoires, voir e.g. (Halliday *et al.*, 1976:278–282; Lyons, 1977; Cruse, 1986:chap7)). En dehors de ces relations directes, les bases de données lexicales (e.g. les versions plus récentes de WordNet, cf. (Miller, 1990), (Fellbaum, 1998)) ont éprouvé le besoin d'y adjoindre des liens indirects, e.g. co-hyponomie et co-méronymie : ce sont ces deux relations que subsume la notion que j'ai appelée voisinage ou approximation. La recherche d'informations est évidemment très utilisatrice de ces liens, qui fournissent des réponses indirectes (donc supplémentaires) aux requêtes. Un autre usage évident de ces regroupements de termes est bien sûr l'enrichissement de bases terminologiques et de thesaurus. Les exemples (9) à (11) illustrent tous des cas de noms ou adjectifs construits par composition néoclassique[4] (cf. §3.1.1 et (Warren, 1990; Fradin, 2000)). Les lexèmes dans (9a) et (9b) sont synonymes (symbole '=') ; dans (10a, b, c), les noms *entéralgie*, *métrorragie* et *albuminémie* sont, respectivement, des cas particuliers de *abdominodynie*, *hystorrhée* et *protéinémie* (symbole '<') ; enfin, (exemples 11a et 11b) il existe un rapport de voisinage ou d'approximation (symbole '~') entre les adjectifs *cancériforme* et *carcinoïde*, d'une part, et entre les noms *xérophtalmie* et *sclérophtalmie*, d'autre part.

(9) a. abdominoscopie = laparoscopie
b. thanatogène = mortifère
(10) a. entéralgie < abdominalgie
b. métrorragie < hystorrhée
c. albuminémie < protéinémie
(11) a. cancériforme ~ carcinoïde
b. xérophtalmie ~ sclérophtalmie

4. Constitution des termes d'un corpus en familles lexicales : Etant donné un corpus spécialisé de grande taille, les relations lexicales ci-dessus sont le moyen de constituer des ensembles (que je baptise « familles lexicales ») qui, pour un nom ou adjectif composé donné, réunit la totalité des autres lexèmes composés (et éventuellement affixés) qui entretiennent avec lui l'une de ces relations. La famille d'un lexème du lexique biomédical fournit d'autres renseignements : il rappelle la définition calculée par analyse, et indique la tête de chapitre à laquelle se rapporte ce lexème, dans les classifications standards reconnues (CCAM, MesH, SNOMED, CIM-10[5], etc.) :

[4] L'intérêt de ces composés néoclassiques ou savants est qu'ils constituent à eux seuls près de la moitié des néologismes recensés dans les textes médicaux (Lovis *et al.*, 1998).

[5] L'organisation de ces nomenclatures, ainsi que le sous-domaine médical dont elles constituent chacune le thésaurus de référence sont consultables, respectivement, aux adresses : http://www.caducee.net/DossierSpecialises/systeme-information-sante/ccam.asp,

(12)

```
        (4400) blastomycose/NOM (mycose)
"  (Partie de -- Type particulier de) mycose en rapport avec
le(s) cellule embryonnaire  "
      blastomycose/NOM: synonym of  (blastomycosique/ADJ)
      blastomycose/NOM: subtype of  cytomycose/NOM
      blastomycose/NOM: see also  (chromoblastomycose/NOM)

           (22284) phlébodynie/NOM (maladie)
"  douleur (du -- liée au) veine  "
      phlébodynie/NOM: synonym of  phlébalgie/NOM
      phlébodynie/NOM: subtype of  angialgie/NOM
      phlébodynie/NOM: see also  phlébite/NOM
```

La possibilité de bâtir une méthode translinguistique vient du fait que, contrairement à la langue générale, la morphologie des lexèmes spécialisés obéit à des règles constructionnelles extrêmement proches dans toutes les langues européennes (Iacobini, 2003). Pour les mêmes raisons, la démarche multilingue s'applique au calcul des liens lexicaux entre mots composés savants du vocabulaire médical[6] ; trois types de ressources intéragissent : un analyseur morphologique, une table établissant des relations de base entre les racines gréco-latines pouvant entrer dans la formation de mots, et un système de règles calculant les relations lexicales entre les termes. Alors que la conception d'un analyseur est une tâche qui doit être réitérée pour chaque nouvelle langue, nous allons voir que la table est une donnée unique multilingue et que le système de règles est indépendant de la langue choisie. L'approche a été implémentée en français, au moyen de DériF dont quelques aspects viennent d'être dévoilés, sur un lexique d'environ 29 000 termes ; elle donne lieu à l'émergence d'environ 3 000 familles lexicales. L'article s'organise comme suit. Je présente tout d'abord (§2) les connaissances et données sur lesquelles repose l'approche morphologique pour la définition multilingue de relations lexicales entre termes. Ensuite, (§3) je développe la méthode utilisée pour réaliser cet objectif, et j'expose (§4) les résultats obtenus en français. Ces résultats conduisent naturellement à une discussion et à des perspectives (§5) qui clôturent cette présentation.

2. Genèse

Comme annoncé *supra*, la méthode proposée s'appuie sur la synergie entre un analyseur morphologique basé sur règles, une table qui classe et annote les racines

---

http://disc.vjf.inserm.fr:2010/basismesh/mesh.html, http://www.snomed.org/, http://www.med.univ-rennes1.fr/noment/cim10/,

[6] Les exemples sont donnés en français, italien, espagnol, allemand anglais, notés respectivement : FR, IT, ES, DE,EN

gréco-latines utilisées dans les termes médicaux, et des règles de calcul de relations lexicales entre mots composés savants. Un certain nombre de constatations sont à l'origine de cette démarche qui est à la fois indépendante de la langue de travail, et spécifique aux domaines de spécialité proches du biomédical.

(1) Les principes théoriques en morphologie lexicale[7] permettent de déduire la définition d'un mot morphologiquement complexe en fonction de celui de ses constituants. Donc, un système implémentant une telle approche théorique (comme DériF, cf. §4) est à même de calculer la pseudo-définition de mots inconnus à partir des procédés morphologiques mis en œuvre.

(2) Quelle que soit la langue européenne considérée, les mots complexes en biomédecine contiennent dans leur grande majorité des racines gréco-latines (*gastr-*, *-phage*, *-hydr-*), qu'à la suite de (Haspelmath, 2002) entre autres, je nomme éléments de formation, notés désormais EFs. Un EF partage sa catégorie et son sens avec l'entrée lexicale contemporaine auquel il supplée (ainsi, *gastr-* signifie *estomac*$_{\text{FR}}$, et son type catégoriel est NOM). D'une langue à l'autre, la réalisation des EFs ne présente que de légères variations graphiques, et leur emploi dans la formation de termes de spécialité met en jeu des règles quasiment identiques (Iacobini, 2003). Il en résulte que les EFs et les structures de mots complexes peuvent avantageusement être représentés par des symboles abstraits, qui gomment les différences entre les langues. Ainsi, le terme abstrait VASCUL--ITE[8] correspond à *vascul--ite*$_{\text{FR}}$, *Vascul--itis*$_{\text{DE}}$, *vascol--ite*$_{\text{IT}}$ et *vascul--itis*$_{\text{ES/EN}}$. L'analyse de ce terme abstrait produit des EFs tout aussi abstraits : VASCUL et ITE. Chacun de ces EF abstraits possède une réalisation phonétique et graphique qui varie légèrement d'une langue à l'autre : *vascul*$_{FR/ES/EN}$, *vasco*$_{IT}$, *Vascul*$_{DE}$. Parallèlement à ces réalisations, les EFs abstraits donnent lieu à des traductions dans chaque langue, partageant la même catégorie lexicale : *vaisseau sanguin*$_{FR}$, *vaso sanguíneo*$_{ES}$, *vaso sanguigno*$_{IT}$, *blood vessel*$_{EN}$, *Blutader*$_{DE}$. D'autres exemples sont présentés dans la **Fig. 1,** §.3.1. Une description linguistique plus détaillée des EFs est, elle, donnée au §.3.1.2.

(3) La dernière observation qui sous-tend cette approche, peut-être la plus importante, est l'exploitabilité des systèmes internationaux de classification (SNOMED, CIM-10, MesH), qui organisent la terminologie médicale au moyen notamment de relations lexicales (synonymie, méronymie, (co)hyponymie…). L'identité entre un EF et sa traduction rend transposables ces systèmes classificatoires pour l'organisation hiérarchique des EFs : comme je le montre au §3.2., de la même façon que *estomac* est une partie du *ventre*, tous deux étant décrits dans le chapitre *ANATOMIE*, GASTR est une partie de ABDOMIN, les deux EFs se trouvant également sous le descripteur *ANATOMIE*.

---

[7] Nos travaux suivent des hypothèses liées à une morphologie de type lexématique, où sens et structure se calculent conjointement, et constituent une adaptation de la théorie élaborée à l'origine dans (Corbin, 1987).

[8] Les EFs abstraits sont écrits en petites majuscules, les frontières entre EFs sont représentées par '--'

3. Démarche

L'approche s'articule autour de trois types de données et techniques, qui répondent aux observations faites en §2 : un ensemble réduit de règles générales (§3.3) infèrent des relations lexicales entre les noms et adjectifs composés d'un corpus à partir de relations de base établies entre les EFs qui constituent ces termes, et réunis dans une table (§3.2) ; enfin, l'identification de ces EFs requiert l'intervention d'un analyseur morphologique (§3.1).

3.1. *Analyseur Morphologique monolingue*

Le processus de décomposition d'un lexème complexe en constituants est une tâche monolingue, dévolue à un analyseur morphologique qui peut fonctionner selon des approches diverses, allant de la simple segmentation (Lovis *et al.*, 1995)) à l'application de contraintes permettant d'annoter les résultats d'informations sémantiques (Namer, 2002). Les résultats des analyseurs basés sur contraintes ont l'avantage d'associer à une décomposition hiérarchique la définition du mot analysé en fonction du procédé morphologique identifié, ainsi que l'illustrent les exemples (3) à (8), §1.2 : le sens d'un lexème obtenu par affixation y est calculé à partir de celui de sa base, via la traduction de celle-ci, lorsqu'elle est réalisée sous forme d'EF : *hépatique*$_{\text{ADJ}}$ = "en relation avec le foie". Les exemples de la **Fig. 1** illustrent différentes situations (et différentes langues) dans lesquelles des EFs servent à construire des lexèmes spécialisés suffixés, préfixés ou composés.

	Lang.	**Affixation/Composition**[9]	***traduction***	**EFs abstrait**
	IT	epatico	*hépatique*	HEPAT
	FR	buccal	*buccal*	BUCC
	EN	analges(ic)	*analgésique*	ALGES
	DE	Hypothermie	*hypothermie*	THERM
	ES	intracefal(ico)	*intracéphalique*	CEPHAL
(1)	IT	gastroectomia	*gastrectomie*	GASTR, ECTOMI
(2)	FR	buccodent(aire)	*buccodentaire*	BUCC, [DENT][10]
(3)	EN	thermoalges(ia)	*thermoalgésie*	THERM, ALGES
(4)	DE	Thermotaxis	*thermotaxie*	THERM, TAXI
(5)	ES	braquicefalo	*brachycéphale*	BRACHY,CEPHAL

Figure 1. *Eléments de Formation et procédés morphologiques*

[9] L'instance d'EF servant de base est soulignée. En cas de présence de ce que j'analyse comme un marqueur de classe suffixoïde, celui-ci est mis entre parenthèses.
[10] Le nom *dent* est un lexème autonome du français, et pas un EF : je l'indique entre crochets.

3.1.1. *Composition populaire et composition néoclassique*

Contrairement à la suffixation et à la préfixation, qui consistent en l'application d'un opérateur sur une base, la composition est un procédé de formation lexicale qui fait intervenir deux unités possédant un sens référentiel. Les composants appartiennent à l'une des catégories majeures : NOM, ADJ, VER (on se reportera à (Fradin, 2003: 199-206) pour une synthèse des critères – souvent avancés à l'origine par D. Corbin - permettant de distinguer composition morphologique et composition syntaxique). Le lexème construit est soit un nom, soit un adjectif. Le calcul de son sens est fonction des deux composants en présence. La distinction traditionnelle qui s'opère entre les différents types de composés emprunte la terminologie de la grammaire du sanskrit (cf. par exemple (Benveniste, 1974:chap.XI)). En français, les composés relèvent en général de l'une des trois classes suivantes :

- dvandva (coordinatifs) : le sens du composé est la coordination du sens de chaque composé. Ainsi, *buccodentaire*$_{ADJ}$ est ce « qui est relatif à la bouche et aux dents » (*-aire* fonctionne comme un marqueur de classe), et un *otorhinolaryngologiste*$_{NOM}$ s'occupe d'oreilles, nez et larynx ; (voir aussi **Fig. 1**, (2))

- bahuvrihi (appelé également exocentrique) : le composé désigne une entité qui n'appartient pas au domaine conceptuel auquel appartient son constituant **recteur** (cf. *infra*) : c'est ce qu'on observe avec *casse-pieds*$_{ADJ}$ (qui n'a à avoir ni avec 'casser', ni avec 'pied' mais désigne une propriété), *tourne-broche*$_{NOM}$ (qui dénote un instrument), *gastéropode*$_{NOM}$ (qui n'est ni un estomac (°*gast(é)r*)[11], ni un pied (°*pode*), mais nomme un mollusque), *microcéphale*$_{ADJ}$ (qui n'est pas une petite (°*micro*) tête (°*céphale*), mais caractérise un individu). (voir aussi **Fig. 1**, (5))

- tatpurusha (déterminant-déterminé, endocentrique) : contrairement au cas précédent, le sens du composé constitue un cas particulier du sens de son constituant **recteur** (cf. *infra*). L'autre constituant détermine (spécifie) le constituant recteur. Ainsi, un *homme-grenouille*$_{NOM}$ est un homme, *gris-bleu*$_{ADJ}$ est une sorte de gris, *thermomètre*$_{NOM}$ est un instrument de mesure (°*mètre*) destiné à la chaleur (°*thermo*), *électromagnétique*$_{ADJ}$ décrit un cas particulier de propriété *magnétique*, due à l'électricité. (voir aussi **Fig. 1**, (1), (3) et (4))

On oppose composition néoclassique et composition populaire. Alors que cette distinction est tributaire pour certains de la nature des composants, (lexèmes autonomes ou éléments de formation), ou, pour d'autres, de la position de la tête du composé, D. Corbin (Corbin, 1992, 2004) préfère opérer une bipartition en fonction de l'élément **sémantiquement recteur** : il se trouve en deuxième position, dans lexème composé néoclassique (*bronchopneumonie*$_{NOM}$, *anglophile*$_{ADJ}$), en première position, en composition populaire (*coupe-feu*$_{ADJ}$, *requin-marteau*$_{NOM}$). Quant à la prétendue homogénéité étymologique des EFs, elle n'est au plus que normative. En effet, la nature ainsi que l'origine des composants de nombreux noms et adjectifs composés est souvent hétérogène : *théatrolâtre*$_{ADJ}$ associe un lexème (*théâtre*$_{NOM}$) à un composant d'origine grecque (°*lâtre*), *larviphage*$_{ADJ}$ est composé d'un EF latin

[11] '°' marque un élément lexical non autonome en syntaxe.

(°*larvi*) et l'autre grec (°*phage*), *télévore*$_{ADJ}$ enfin est composé à partir de l'apocope de *télévision*$_{NOM}$, et de l'EF latin °*vore*. [12]

Dans ce qui suit, j'emploie la notation suivante. YX est un nom ou adjectif composé néoclassique, dont l'élément recteur est X et l'élément régi est Y.

3.1.2. *Eléments de Formation*

Quel est le statut de ces éléments de formation ? Tout d'abord, la notion elle-même d' « élément de formation » que j'ai décidé d'adopter, suivant (Haspelmath, 2002) et C. Iacobini (entre autres (Iacobini, 2003)) possède un sens équivalent de celui que D. Corbin et J. Paul attribuent aux archéo-constituants (Corbin *et al.*, 1999). La notion d'EF fait référence aux éléments non-autonomes du lexique qui sont munis d'un sens référentiel et dont la forme indique qu'ils sont d'origine grecque ou latine (pour l'essentiel). On les nommes également racine liée, base non autonome[13]. Donc, contrairement aux classifications de (Tournier, 1985; Warren, 1990) et (Fradin, 2000), pour ne citer qu'eux, j'exclus des « éléments de formation » ce que (Corbin *et al.*, 1999) nomment fracto-constituants (e.g. *pétro* de *pétrodollar*, ou *euro*, de *euromissile*).

Etant munis par définition d'un sens référentiel, pouvant par définition intervenir dans des procédés constructionnels (la composition néoclassique) les éléments de formation (EFs) sont des lexèmes à tous les titres, sauf celui de l'autonomie en syntaxe. Ayant le statut de lexème, ils doivent par conséquent en recevoir les propriétés : c'est-à-dire au moins une graphie, une prononciation, un sens, et une catégorie lexicale. Dans le cadre strict d'un de l'analyse morphologique des lexèmes construits du vocabulaire biomédical, les décisions suivantes ont été prises :

- Une séquence est identifiée comme étant un EF si elle constitue une variante d'un élément du vocabulaire grec ancien ou latin. Une telle appartenance est établie en consultant les dictionnaires, e.g. (Bailly, 2000; Gaffiot, 2000) ; et en croisant les renseignement étymologiques du Littré Médical (Littré, 1884), du Dorlands (Dorlands, 2002), et du Cottez (Cottez, 1988).

- Parce qu'ils sont spécifiques du langage biomédical, j'admets comme EFs ceux que (Cottez, 1988) (suivi en cela par (Corbin *et al.*, 1999)) appelle suffixes : *-ite, -matose, -ome, -ose*. Il s'agit des éléments qui n'ont pas d'origine gréco-latine, mais qui font spécifiquement référence à des symptômes pathologiques divers, révélés par : une inflammation (*-ite*), une manifestation tumorale (*-ome*) éventuellement

[12] On constate que la composition populaire ne manipule que des composants appartenant au lexique (français) contemporain, en contraste avec l'hétérogénéité de ceux-ci en composition néoclassique. On relève deux exceptions à cette organisation : il s'agit des séries formées par les adjectifs obéissant à la structure VN où V est l'un des EF °*philo* (*aimer*), °*miso* (*haïr*) et N est soit un EF (°*gyne* (*femme*) dans *mysogyne*$_{ADJ}$, °*sophe* (*sagesse*) dans *philosophe*$_{ADJ}$) soit un lexème autonome, e.g. dans *philorusse*$_{ADJ}$.

[13] Sur une énumération plus complète des termes par lesquels cet objet lexical est désigné, voir (Iacobini, 2003:70-71).

prolifératoire (*-matose*), ou encore une affection chronique (*-ose*). Les formes de maladies auxquelles renvoient ces séquences sont conceptuellement hiérarchisables, ce qui constitue une motivation supplémentaire de ma décision de les ranger parmi les EFs.

- Cette identification s'accompagne du (ou des) traductions que cet EF possède en français. Par exemple, la séquence initiale °*auri-* est soit reliée à lat :*aurum* (or_{NOM}) ou lat :*auris* ($oreille_{NOM}$). Ainsi, $auriforme_{ADJ}$ décrit ce qui est en forme d'oreille, alors que $aurithérapie_{NOM}$ désigne la thérapie qui utilise l'or.

- la traduction sert à décider de la catégorie lexicale à attribuer à l'EF.

Comme on le verra au §.4, et comme le suggère l'exemple (5), §.1.2, ces triplets (EF, traduction, catégorie de la traduction) suffisent à un analyseur morphologique comme DériF pour réaliser les analyses des lexèmes construits contenant un EF. La traduction sert à bâtir la définition du lexème, et la catégorie, à vérifier la compatibilité de l'EF avec les contraintes imposées par les procédés[14].

3.2. *Table multilingue des Eléments de Formation*

Les observations (2) et (3) du §2 conduisent tout naturellement à la conception d'une table réunissant l'ensemble des quelques 900 EFs utilisés dans le vocabulaire biomédical, et dont la **Fig. 2** donne un échantillon. Chaque entrée de la Table est caractérisée par deux dimensions : l'abstraction des EFs réalisés dans chaque langue de travail, et le codage des relations lexicales de base sur chaque EF abstrait.

Abstraction des EFs : Chaque EF est codé sous la forme d'un symbole abstrait, qui apparaît dans la colonne numérotée (1) de la **Fig. 2**. Pour chaque langue de travail que l'on veut pouvoir utiliser, la sous-colonne appropriée de la colonne (2) code la réalisation de cet EF abstrait, ainsi que sa traduction dans la langue : l'ajout d'une nouvelle langue dans le système suppose donc uniquement l'insertion d'une nouvelle sous-colonne dans la colonne (2).

La colonne (3) elle, enregistre la catégorie grammaticale que cette traduction possède quelle que soit la langue. C'est principalement grâce aux entrées des dictionnaires spécialisés francophone (Synapse (Synapse, 1997), Larousse médical (Levallois, 2000), les ressources en ligne Biotop (Dolisi) et AltMédica (Altmedica)) et anglophone (Dorlands, 2002) que les réalisations et les valeurs des EFs sont codées en français et en anglais, puis dans les autres langues par recoupements successifs, grâce au lexique bilingue anglais-allemand (Nowak, 2005) aux bases lexicales en ligne multilingues (EuroDicAutom, Multilingual Glossary of medical terms[15]) aux portails de dictionnaires bilingues spécialisés, etc.

---

[14] le traitement des quelques EFs ambigus (e.g. °*auri*) fait intervenir des listes d'exception.

[15] cf. URLs europa.eu.int/eurodicautom/Controller, users.ugent.be/~rvdstich/eugloss/welcome.html, www.atoute.org/dictionnaire_medical.htm

Relations lexicales de base entre EFs : Cette dimension du codage est la plus importante, puisque c'est à partir de ces données que vont être projetées les relations lexicales entre termes. Elle tire profit des facilités classificatoires offertes par les structures hiérarchiques des nomenclatures médicales (cf. note 5, §.1.2). Parce qu'elle couvre un domaine de description plus large que les autres, et parce que son système hiérarchique est considéré le plus homogène par les experts, c'est la classification de la SNOMED qui m'a servi essentiellement de source. Le principe est le suivant.

La SNOMED organise la terminologie médicale au moyen, notamment, de relations lexicales (synonymie, méronymie, (co)hyponymie…). Il y est par exemple établi que le concept **ESTOMAC** dépend du chapitre **ANATOMIE**, et entretient une relation de **méronymie** avec le concept **ABDOMEN**. Or, tant *estomac* que *abdomen* apparaissent dans la table des EFs sous la forme des traductions, respectivement, des EFs abstraits GASTR et ABDOMIN. Il est par conséquent légitime d'attribuer à ces EFs abstraits la même relation lexicale de **méronymie** que celle observée entre **ESTOMAC** et **ABDOMEN**. En second lieu, le chapitre SNOMED sous lequel est enregistré chaque concept, devient par là-même la tête de chapitre de l'EF abstrait correspondant. Cette technique aboutit au codage de quatre types de relations lexicales de base entre les EFs :

(1) **synonymie, notée =.** Elle relie par exemple GASTR et STOMAC, que le français traduit par *estomac*, ALGIE et ODYNIE (*douleur*), OPT et OPHTALM, (*vision*) … Cette relation est transitive et symétrique.

(2) **hyponymie, notée <.** Elle associe les EFs comme BLAST *(cellule embryonnaire*) et CYT *(cellule*), PHLEB (*veine*) et VASCUL (*vaisseau sanguin*). Cette relation est anti-symétrique ; comme la précédente, elle est transitive.

(3) **méronymie, notée ←.** Les « parties » de systèmes ou d'appareils sont ainsi reliés au « tout » qui les contient directement, comme CORO (*pupille)* et OCUL (*œil*), HEPAT (*foie*) et ABDOMIN (*abdomen*), ODONT (*dent*) et BUCC (*cavité buccale*). Cette relation est anti-symétrique et arbitrairement intransitive (voir note 16).

(4) **approximation, notée ~.** Cette relation regroupe les co-méronymes : RHIN (*nez) et* OTO (*oreille*), HEPAT (*foie*) et GASTR (*estomac*), PHLEB et ANGI (*vaisseau sanguin*) et les co-hyponymes ALGIE et ITE (*inflammation*), ECTOMIE (*ablation*) et TOMI (*incision*). Cette relation est intransitive[16].

Dans la **Fig. 2**, la colonne (5) inscrit, suivant les définitions ci-dessus, l'ensemble des relations lexicales de base que l'EF courant entretient avec les EFs appropriés et présents dans la Table. La colonne (4) code elle la tête de chapitre SNOMED à laquelle l'EF est indirectement relié. L'échantillon proposé dans la **Fig. 2** se compose de deux parties :

[16] Le blocage de la transitivité permet d'éviter des enchaînements qui entraîneraient des appariements incontrôlables : ainsi, du point de vue anatomique : BUCC (*bouche*) ~ NAS (*nez*) ; or selon une perspective fonctionnelle, NAS ~ OLFACT (*odorat*). Imposer l'intransitivité évite d'associer BUCC à OLFACT.

(a) sur fond gris, les informations abstraites, indépendantes de la langue, qui vont être mobilisées pour le calcul des relations entre termes ; Par exemple, l'EF abstrait nominal GASTR, terme d'*ANATOMIE*, est synonyme de STOMAC, appartient à ABDOMIN, et a à voir avec HEPAT (*foie*), ENTER (*intestin*) et PANCREAT (*pancréas*).

(b) sur fond blanc, les valeurs de chaque EF abstrait en fonction de la langue de travail : chaque instance couple la réalisation du symbole abstrait, avec sa traduction : en français, ces traits sont utilisés lors du calcul de la définition littérale du lexème contenant ces EFs, et lors de la tâche finale du calcul des traits lexicaux qui relient ce lexème (s'il s'agit d'un composé néoclassique) à d'autres noms ou adjectifs composés du corpus de travail. Ainsi, ALGI est instancié par *algia/algy*$_{EN}$: 'pain', *algie*$_{DE}$: 'Schmerz', *algie*$_{FR}$: 'douleur', *algia*$_{IT}$: 'dolore', *algia*$_{ES}$: 'dolor'.

EF		**Instanciation (2)**					CAT	**Chapitre SNOMED**	**Relation lexicale**
(1)		**Anglais**	**Allemand**	**Français**[17]	**Italien**	**Espagnol**	(3)	**(4)**	**(5)**
GASTR	réal trad	gastr stomach	Gastr Magen	gastr estomac	gastr stomaco	gastr estomago	N	*ANATOMIE*	=STOMAC, ←ABDOMIN, ~HEPAT, ~ENTER , ~PANCREAT
ALGI	réal trad	algia/algy pain	algie Schmerz	algie douleur	algia dolore	algia dolor	N	*SYMPTOME*	=ODYN, ~ITE
ITE	réal trad	itis inflammation	ite Inflammation	ite inflammation	ite infiammazione	itis inflamación	N	*SYMPTOME*	~ALGI, ~ODYN
PHLEB	réal trad	phleb vein	Phleb Vene	phléb veine	fleb vena	fleb vena	N	*ANATOMIE*	=VEN, <ANGI, <VASCUL
ANGI	réal trad	angio blood vessel	Angio Blutader	angio vaisseau sanguin	angio vaso sanguigno	angio vaso sanguíneo	N	*ANATOMIE*	=VASCUL, ~VAS
ECTOMI	réal trad	ectomy ablation	ektomie Ablation	ectomie ablation	ectomia ablazione	ectomía ablación	N	*ACTE MEDICAL*	~TOMI, ~STOMI

Figure 2. *Table multilingue des Eléments de Formation (échantillon)*

3.3. *Règles indépendantes de la langue pour le calcul des relations lexicales*

La projection des relations lexicales entre EFs (**Fig. 2**), sur les noms et adjectifs composés dont l'analyse morphologique fait apparaître ces EFs, requiert l'activation de l'une des quatre règles indiquées dans la **Fig. 3**. Ces règles sont totalement indépendantes de la langue. Chacune est décrite formellement dans la col. 1, et exemplifiée dans les colonnes suivantes. La règle **R2**, par exemple, établit que tout couple de composés A et B dont les constituants X_A et X_B sont synonymes,

[17] Les réalisations indiquées possèdent des variantes allomorphiques codées également dans la Table quand elles reflètent des situations morphologiquement pertinentes. Ainsi, *algo* et *algés* sont des variantes de *algie* ne pouvant occuper que la position Y dans un composé.

entretiennent la même relation R que celle établie entre Y_A et Y_B, sauf si R est la relation de méronymie : si Y_A est une partie de Y_B en effet, A est hyponyme de B. A titre d'exemples, la synonymie entre MORT et THANAT (*mort*) se propage entre les adjectifs *mortifero*$_{IT}$ et *tanatogeno*$_{IT}$, la relation d'hyponymie entre *apivore*$_{FR}$ et *entomophage*$_{FR}$ provient de celle entre API (*abeille*) et ENTOMO (*insecte*), alors que celle entre *Enterodynie*$_{DE}$ et *Abdominalgie*$_{DE}$ résulte de la méronymie entre ENTER (*intestin*) et ABDOMIN (*abdomen*). Enfin, l'approximation entre BACTERI et BACILL entraîne celle entre *bacilliform*$_{EN}$ et *bacterioid*$_{EN}$. La règle **R4** est symétrique à **R2**, en ce que Y_A et Y_B y sont synonymes, et la relation entre A et B dépend alors de celle qu'entretiennent X_A et X_B. Enfin **R1** (resp. **R3**) est la version simplifiée de **R2** (resp. **R4**), où A et B partagent le constituant X (resp. Y) [18].

Règle	Exemple		
	Y	X	$[Y_AX_A]$ R $[Y_BX_B]$
R1 A = $[Y_AX]$ et B=$[Y_BX]$ Si $Y_A \leftarrow Y_B$ alors A < B sinon si Y_A R Y_B et si R est { =, <, ~} alors A R B	PROCTO ← COLO LEUCO ← HEMATO ABDOMIN=LAPAR ALBUMIN<PROTEIN XER ~SCLER	RRAGIE GRAMME SCOPIE EMIE OPHTALMIE	EN: proctorrhagia < colorrhagia DE: Leukogramm < Hämatogramm FR: abdominoscopie = laparoscopie IT: albuminemia < proteinemia ES: xerophthalmia ~sclerophthalmia
R2 A = $[Y_AX_A]$ et B=$[Y_BX_B]$ et $X_A = X_B$ si $Y_A \leftarrow Y_B$ alors A < B sinon si Y_A R Y_B et si R est { =, <, ~} alors A R B	ENTER←ABDOMIN MORT = THANAT API < ENTOMO BACILL ~BACTERI	$X_A = X_B$ ALGIE = ODYNIE FERE = GENE VORE = PHAGE FORME = OÏDE	DE: Enterodynie < Abdominalgie IT: mortifero = tanatogeno FR: apivore < entomophage EN: bacilliform ~bacterioid
R3 A = $[YX_A]$ et B=$[YX_B]$ Si X_A R X_B et si R est {=, <, ~} alors A R B	BACTER OTO ARTHR	OÏDE = FORME RRAGIE < RRHEE ALGIE ~ITE	FR: bactérioïde = bactériforme DE: Otorrhagie < Otorrhö ES: artralgia ~artritis
R4 A = $[Y_AX_A]$ et B=$[Y_BX_B]$ et $Y_A = Y_B$ si X_A R X_B et si R est {=, <, ~} alors A R B	$Y_A = Y_B$ ORTHO = RECTI METR = HYSTER LIP = ADIP	DONTE = DENT RRAGIE < RRHEE MATOSE ~OME	FR: orthodonte = rectident FR: métrorragie < hystérorrée EN: lipomatosis ~adipoma

Figure 3. *Règles de Calcul des Relations Lexicales*

L'intéraction entre analyseur morpho-sémantique, table multilingue des EFs et règles de calcul des relations lexicales résulte en une chaîne de traitement qui conduit à l'appariement des mots composés savants au moyen des relations lexicales

[18] Une version monolingue des règles et de la table des EFs est présentée dans (Namer *et al.*, 2004).

de synonymie =, hyponymie < et approximation ~. L'analyseur décompose le lexème d'entrée, en identifiant s'il y a lieu, les EFs qui le constituent[19]. Ces EFs servent à alimenter le système des règles de calcul des relations lexicales : pour chaque EF, rapporté à sa structure abstraite, l'ensemble des relations lexicales de base définies dans la table est collecté. Les règles **R1** à **R4** sont activées, et prédisent toutes les relations potentielles abstraites avec l'input. La dernière tâche à effectuer consiste alors à filtrer les relations correspondant à des mots inexistants dans le corpus dans lequel les appariements sont calculés. C'est cet enchaînement, réalisé en français sur un lexique de grande taille, qui fait l'objet du prochain paragraphe.

4. Résultats pour le français

L'approche décrite ci-dessus a été implémentée en français. Les résultats ont été obtenus à partir d'un lexique totalisant 29 000 noms, adjectifs et verbes du vocabulaire spécialisé, collectés à partir de diverses sources, librement accessibles en ligne, ou mises à la disposition des projets UMLF et VumeF [20]. Comme l'annonçait le §.1.2., la réalisation de la chaîne de traitement en français est rendue possible avant tout par l'existence de l'analyseur morpho-sémantique DériF.

L'analyse par DériF d'un lexème catégorisé adapte les hypothèses théoriques avancées à l'origine dans (Corbin, 1987). Basé sur l'application d'un système ordonné de règles, le mécanisme est récursif et permet la gestion des ambiguïtés, se réappliquant sur chaque (liste de) résultat obtenu précédemment. L'analyse morphologique d'un lexème construit sur une base elle-même construite est donc hiérarchisée. Le résultat est un triplet, la première partie retrace sous forme crochetée l'historique des étapes d'analyse, la seconde réunit les lexèmes résultats obtenus à chaque étape, et la troisième est constituée d'une formulation en langue naturelle de la relation morphologique liant l'input à son (ses) constituant(s) immédiat(s). Les néologismes sont analysés et pseudo-définis comme des mots régulièrement construits (ce qui est généralement le cas). Quand il analyse un mot composé, enfin, DériF fournit une représentation linéaire Y/X de la décomposition de celui-ci en constituants. Le fonctionnement ainsi résumé de DériF est illustré par l'analyse de *gastralgie*$_{\text{NOM}}$, dans les 3 premières lignes de la **Fig. 4**. On note que la définition calculée pour *gastralgie* mobilise la table des EFs qui fournit la traduction, respectivement de *gastr* (estomac) et *algie* (douleur).

Les représentations abstraites de Y et X (**'Constituants'**, **Fig. 4**, ligne 4) sont transmises au système de calcul des relations lexicales. Comme cela a été mentionné en §3.3, les quatre règles **R1** à **R4** sont activées pour produire les relations lexicales

[19] Selon le type d'analyseur, l'analyse morphologique de l'input fournit éventuellement aussi une pseudo-définition, sous-forme de relation entre l'input et ses composants.

[20] Pour ne mentionner que quelques sources : les versions françaises de la CIM-10, du MesH et le dictionnaire en ligne BIOTOP (URL : http://georges.dolisi.free.fr/)

candidates de l'input (i.e. dans l'exemple, *gastralgie*), que, par commodité, je nomme A (et dont la structure supposée est donc Y_AX_A) : chacune de ces règles va « fabriquer » l'ensemble des structures de lexème composé B potentiels reconstituables à partir des éléments Y_B et X_B compatibles avec les indications dictées par X_A et Y_A dans la règle en question.

Tout d'abord, **R1** s'applique. X_B est une copie de X_A ; Y_B est instancié par tous les EFs abstraits trouvés dans la Table avec lesquels Y_A est en relation : STOMAC, HEPAT, ABDOMIN, ENTER, PANCREAT ; toutes les combinaisons Y_BX_B sont conservées ; elles sont restitués sous leur forme de réalisation en français, et la relation potentielle entre A et B est calculée par **R1** pour chaque Y_B à partir de la relation de base qui lui est affecté dans la Table (cf. **Fig. 2**), e.g. `eql:stomach/algie`[21] (**Fig. 4**). C'est ensuite, **R2** qui est activé : X_B reçoit successivement l'ensemble des synonymes de X_A dans la Table des EFs ; et l'opération d'instanciation de Y_B est identique à ce qui se passe avec **R1 ;** toutes les combinaisons X_BY_B sont réunies ; (e.g. `isa:abdomin/odyn`) ;

Ensuite, les rôles de Y_B et X_B sont inversés, lors de l'activation de **R3** (e.g. `see:gastr/ite`) et de **R4** (e.g. `see:stomach/ose`).

```
gastralgie/NOM => [ [ gastr N* ] [ algie N* ] NOM ]
(gastralgie/NOM, algie/N*)
" douleur (du -- liée au) estomac "
Constituants = /gastr/algie/
Type = maladie
Relations possibles = (eql:gastr/algo, eql:gastr/algés,
eql:gastr/odyn, eql:stomac/algie, eql:stomac/algo,
eql:stomac/algés, eql:stomac/odyn, eql:stomach/algie,
eql:stomach/algo, eql:stomach/algés, eql:stomach/odyn,
isa:abdomin/algie, isa:abdomin/algo, isa:abdomin/algés,
isa:abdomin/odyn, see:entéro/algie, see:entéro/algo,
see:entéro/algés, see:entéro/odyn, see:gastr/ite,
see:gastr/ose, see:hépat/algie, see:hépat/algo,
see:stomach/ite, see:hépat/algés, see:hépat/odyn,
see:pancréat/algie, see:pancréat/algo, see:pancréat/algés,
see:pancréat/odyn, see:stomac/ite, see:stomac/ose,
see:stomach/ose )
```

Figure 4. *Relations lexicales candidates pour gastralgie$_{NOM}$*

A partir de cet ensemble de relations candidates, le système ne garde que celles qui relient les termes attestés. Pour *gastralgie*, et étant donné le contenu du lexique de 29 000 entrées du français, on s'attend à ce que seuls les éléments soulignés dans

[21] L'affichage par DériF des relations lexicales possibles est de la forme `'R:Y/X'`; R symbolise la synonymie = par `'eql'`, l'hyponymie < par `'isa'` et l'approximation ~ par `'see'`.

la **Fig. 4** correspondent à des lexèmes "réels". Les autres sont soient morphologiquement impossibles (gastr/algés, par exemple ne peut pas se réaliser, car *algés* est une forme que l'on ne trouve qu'en position Y), soit non attestés (dans le corpus du moins) : c'est par exemple le cas de pancréat/odyn, car *pancréatodynie* n'est pas dans le lexique. Etant donné un composé A (ex. *gastralgie*) l'identification de ses relations lexicales 'réelles' s'effectue au moyen du couple Y/X, calculé par DériF pour chaque entrée B du lexique et consigné en valeur du trait **'Constituants'**. Quand pour un input B donné, Y/X s'identifie à l'un des candidats de la listes des relations *possibles* de A, B est ajouté à la liste des relations *attestées* de A. A la fin de cette étape, chaque input A du lexique de travail se voit associer sa famille lexicale, regroupant l'ensemble des composés du corpus avec lesquels A entretient l'une des relations de synonymie, hyponymie et approximation. Clairement la technique mise en œuvre est tributaire de la taille du corpus de travail : plus celui-ci possède une couverture lexicale importante, plus il y aura de relations candidates (e.g. celles de la **Fig. 4** pour *gastralgie*) qui se réaliseront. La **Fig. 5** reproduit la famille lexicale de *gastralgie* collectée dans le lexique de termes spécialisés à partir des candidats calculés. Une règle secondaire insère à ce niveau d'autres termes qui, de par leurs propriétés morphologiques, vérifient l'une des trois relations que la **Fig. 5** note comme : 'synonym of', 'subtype of' et 'see also'. Ces termes sont indiqués entre parenthèses. Il s'agit par exemple des adjectifs dénominaux (e.g. *gastralgique*$_{ADJ}$), qui, du point de vue de leur usage terminologique, sont considérés comme 'synonymes' de leur nom base. Les lexèmes préfixés sont également ajoutés à la famille de leur base ; ils y sont reliés par relation d'approximation (e.g. *antigastralgique*$_{ADJ}$).

gastralgie/NOM (maladie)	« douleur (du – liée au) estomac »
synonym of	gastrodynie/NOM, stomacalgie/NOM, stomacodynie/NOM, stomachodynie/NOM, (gastralgique/ADJ)
subtype of	abdominalgie/NOM
see also	entéralgie/NOM, entérodynie/NOM, gastrite/ NOM, hépatalgie/NOM, gastrose/NOM, hépatodynie/NOM, pancréatalgie/NOM, (antigastralgique/ADJ)

Figure 5. *Famille lexicale de gastralgie*

Les modules d'analyse de DériF implémentent à ce jour divers procédés morphologiques, que ce soit la suffixation, la préfixation, la conversion ou la composition savante. DériF est actuellement à même d'analyser comme complexes 17 240 des 29 000 lexèmes du corpus de travail[22]. La chaîne de traitement enfin

[22] Les lexèmes complexes non analysés sont ceux formés suivant des patrons constructionnels non encore (complètement) intégrés dans DériF.

produit plus de 3 000 familles lexicales à partir des lexèmes composés du corpus, générant au total des liens entre 7 438 ADJs et/ou NOMs distincts.

5. Bilan : Discussion, perspectives

L'utilisation de la morphologie des lexèmes composés dans le but d'optimiser la recherche d'information en biomédecine a déjà fait l'objet d'expérimentations, entre autre par (Schulz *et al.*, 1999) et (Hahn *et al.*, 2001), qui se servent également d'EFs (qu'ils appellent 'subwords'). Cependant, contrairement à ce qui est présenté ici, ils n'exploitent pas les relations lexicales de base entre les EFs (et donc ne calculent pas de relations lexicales entre lexèmes construits), et leur analyse morphologique est réduite à un simple découpage linéaire, qui ne permet pas d'associer une définition à l'input.

En contrepartie, bien entendu, l'approche que je défends présente un inconvénient majeur, qui est celui de tout système basé sur l'utilisation de contraintes linguistiques, et demandant la gestion des exceptions. Il nécessite une validation humaine à trois niveaux au moins : pour vérifier la pertinence des analyses, pour valider les pseudo-définitions et surtout pour contrôler la pertinence des relations lexicales de base dans la Table des EFs. Notamment, il faut éviter des annotations trop spécifiques sur des EFs polyréférentiels en médecine. Ainsi, étiqueter LABI (*lèvre*) comme partie-de BUCC (*cavité buccale*) entraînerait un codage pour le moins curieux d'adjectifs comme *inguino-labial*, relatif à la gynécologie. La validation de ces trois dimensions a mobilisé la compétence d'experts biomédicaux membres des projets UMLF et VumeF (principalement S. Darmoni, A. Burgun et R. Baud), dont l'intervention a indubitablement contribué à une rapide amélioration des résultats. Dans le cadre de VumeF, la validation des appariements terminologiques, en contexte syntaxique (grâce notamment à SynteX, cf. (Bourigault *et al.*, 2000a)), est également un moyen indirect de corriger les éventuelles erreurs d'étiquetage des EF ambigus.

A l'avenir, les améliorations prioritaires de la démarche présentée (en dehors de l'évolution de DériF, qui ne concerne que le français) passent par l'ajout :

1. de nouvelles règles : (1) calcul des relations lexicales qui s'appliquent aux termes préfixés et/ou suffixés. Elles permettront par exemple d'identifier *gastrique* comme une propriété synonyme de *stomacal*, et hyponyme d'*abdominal*. **(2)** prise en compte de la symétrie des composés dvandvas dans l'identification de nouvelles relations sémantiques : *orbito-zygomatique*$_{\text{ADJ}}$ est par exemple synonyme de *zygomatico-orbitaire*$_{\text{ADJ}}$, et voisin de *zygomatico-malaire*$_{\text{ADJ}}$. Un adjectif ou nom composé A=YX ne peut pas avoir de symétrique B=XY si X et Y relèvent de têtes de chapitres SNOMED différentes : **ANATOMIE** et **SYMPTOME**, par exemple. Les nouvelles règles de définition de relations lexicales seront semblables aux règles **R1** à **R4** de la **Fig. 3**, la condition supplémentaire portant sur l'inversion des

constituants entre A et B, et, éventuellement, la présence d'un marqueur de classe identifiant un adjectif relationnel dont la valeur peut varier entre A et B (comparer par exemple *placento-fœtal* et *foeto-placentaire*).

2. de nouvelles relations lexicales de base : Dans la table des EFs, certaines relations d'approximation pourraient se spécialiser. Certains EFs constituent en effet des pôles opposés d'une même propriété : e.g. BRACHY *court*, versus DOLICHO *long*. D'autres relations d'antonymie peuvent ainsi être dévoilées, qu'elles soient, comme l'exemple ci-dessus, considérées comme ***scalaires***, ou au contraire qu'elles soient qualifiées traditionnellement de ***polaires*** : DEXTR *(droit*) et SINISTR *(gauche*) par exemple, sont mutuellement incompatibles. De nombreux auteurs ont proposé un recensement des aspects logiques, sémantiques, lexicologiques, etc. de la notion d'antonymie : je renvoie le lecteur, par exemple à (Amsili, 2003).

3. d'un nouveau module :

Enfin, l'évolution du système d'annotation passe par l'ajout d'un nouveau module monolingue au système. Le constat suivant est à l'origine de cette idée : les termes abstraits candidats reliés à un lexème donné, et linguistiquement plausibles ramenés par les règles **R1** à **R4** ne sont validés que s'ils sont attestés quelque part dans le corpus de travail. En d'autres termes, des structures de lexèmes bien formés mais non attestés sont rejetées. Cette situation, rappelons-le (cf. **Fig. 4**), s'observe en français avec `eql :abdomin / odynie`.

Plutôt que de dépendre de la couverture lexicale offerte par un corpus, pour optimiser le nombre de liens lexicaux entre lexèmes construits[23], on pourrait générer automatiquement (dans la langue de son choix) les termes ramenés par les règles sous forme abstraites, absents du corpus, et morphologiquement plausibles. Pour le français, cela reviendrait, par exemple (**Fig. 4**), à générer non seulement *abdominodynie*, mais également *pancréatodynie* et *stomac(h)ite.* Actuellement, ces trois termes sont introuvables sur Internet.

La réalisation dans d'autres langues que le français[24] de l'approche présentée ne nécessite pratiquement que de disposer d'un **analyseur morphologique** pour chaque nouvelle langue. En un premier temps, celui-ci peut être extrêmement rudimentaire (e.g. un simple raciniseur) dans la mesure où sa tâche fondamentale est avant tout d'identifier les composants X et Y d'un lexème composé néoclassique. Une fois cet analyseur disponible, par exemple pour les cinq langues qui m'ont servi à illustrer ma démarche, la chaîne de traitement (à l'exception de l'analyseur) ne manipule plus que des données abstraites.

---

[23] Cette dépendance est d'autant plus hasardeuse que certains lexèmes construits bien formés ne sont attestés nulle part. En français, je n'ai relevé aucune occurrence de *abdominodynie*$_{\text{NOM}}$, même en cherchant sur la Toile. Ce terme par contre est attesté en anglais.

[24] L'ébauche de ce travail d'extension est actuellement en cours pour l'anglais.

C'est ainsi que l'on pourra obtenir un ensemble de familles lexicales abstraites à l'image de ce qu'illustre la **Fig. 5**. Une famille lexicale abstraite peut se voir comme la superposition, puis l'abstraction des familles lexicales observables dans chacune des langues participant à l'expérience. Dans la **Fig. 5**, l'identifiant de la famille abstraite est en petites majuscules soulignées : GASTRALGI. Les autres exposants ('lexèmes abstraits') de la famille, tous en petites majuscules, entretiennent avec GASTRALGI l'une des trois relations : synonymie (codée *equals*) hyponymie (*is a*) et approximation (*see also*). Les liens représentés en pointillés visualisent les relations secondaires dans la famille examinée, c'est à dire externes à GASTRALGI : ces relations (synonymie entre ABDOMINALGI et ABDOMINODYN, hyponymie de HEPATALGI vis-à-vis de ABDOMINODYN, approximation entre GASTROSE et GASTRODYN) constituent autant de liens principaux dans les familles que ces lexèmes abstraits identifient. La dernière dimension présente dans cette figure consiste en la projection de ces lexèmes abstraits dans les différentes langues, chaque projection est représentée en encadré. Le contenu de chaque cadre est le fruit de ce qu'offrent les corpus auxquels j'ai accès.

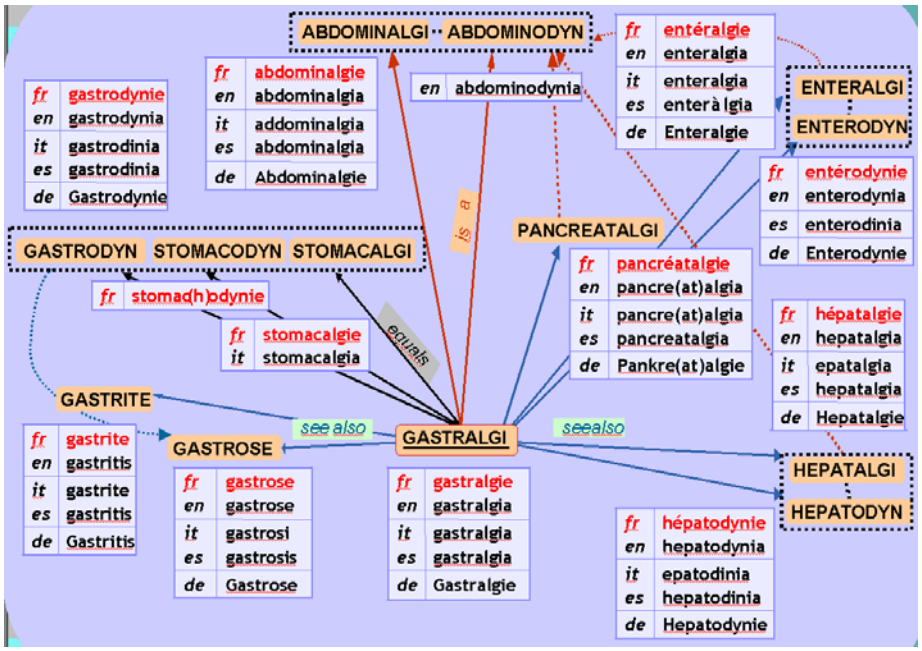

Figure 5. *Une Famille Lexicale Abstraite: celle de* GASTRALGI

A supposer que l'on puisse bénéficier de corpus de couverture comparable dans chaque langue de travail, une famille lexicale abstraite est alors assimilable à une sorte de photographie multilingue permettant de comparer, pour un lexème abstrait donné, les termes que chaque langue choisit ou pas d'utiliser. On y remarque que la

représentativité des lexèmes abstraits dans les cinq langues est variable. les synonymes HEPATALGI et HEPATODYN, par exemple, se réalisent chacun dans les cinq langues. On est alors en droit d'attendre qu'il en va de même pour les trois synonymes abstraits GASTRODYN, STOMACODYN et STOMACALGI. Mais seul GASTRODYN est instancié unanimement. Le français et l'italien attestent STOMACALGI et le français accepte aussi STOMACODYN (que l'on retrouve, d'ailleurs, avec deux variantes orthographiques : *stomachodynie* et *stomacodynie*). Il serait intéressant d'examiner pourquoi le français utilise 4 synonymes pour exprimer ce pour quoi toutes les autres langues (excepté l'italien) ne nécessite qu'un terme. A l'inverse, l'attestation d'un terme dans une langue peut servir de justification voire de prédiction à l'émergence dans un futur proche de ce même terme dans les langues où il constitue encore un trou lexical.

6. Conclusion

J'ai voulu montrer comment morphologie constructionnelle, sémantique lexicale et acquisition de terminologie peuvent coopérer. A cet effet, j'ai décrit une méthode permettant de regrouper les noms et adjectifs composés savants du langage biomédical selon des liens sémantico-lexicaux, grâce à une classification multilingue de base (la table des EFs) établie à partir des terminologies internationales du domaine médical. Quelques règles indépendantes de la langue servent à propager ces relations de base sur les composés qui contiennent ces EFs, pour calculer les relations lexicales qu'entretiennent les composés entre eux. Réalisée en français, cette méthode est facilement transposable à d'autres langues, dès lors qu'elles disposent d'un système d'analyse morphologique permettant au moins d'identifier les constituants des lexèmes composés néoclassiques.

Les résultats obtenus en français sont utilisés dans le cadre des projets UMLF et VumeF. Les définitions littérales des termes morphologiquement construits sont en cours d'intégration dans CiSMeF, portail abritant le catalogue des sites spécialisés médicaux francophones. CiSMeF[25]. En matière de recherche d'information et de synergie entre morphologie et terminologie, les derniers résultats produits par DériF fournissent les avancées suivantes dans les projets UMLF et VumeF :

- de nouveaux liens servent immédiatement à l'extension de systèmes d'extraction de variantes terminologiques : l'analyse par DériF de $hépatique_{ADJ}$ sur $foie_{NOM}$ conduit à relier *maladie du foie* et *maladie hépatique*. Ce lien n'est pas envisageable à partir des méthodes d'acquisition terminologiques à base de règles de

[25] CiSMeF, (cf. URL : http://www.cismef.org/) est un projet coordonné au CHU de Rouen depuis plus de dix par S. Darmoni (Douyère *et al.*, 2003) ; il répertorie l'ensemble des URLs indexées par le MesH, et répond à différents types de requêtes tout en donnant à l'utilisateur la possibilité d'accéder aux bibliothèques, bases de données, bibliographies, journaux électroniques, adresses d'établissements de soins et universités pertinents.

racinisation. Les relations de synonymie, hyponymie et approximation déduites des règles de la **Fig. 3** constituent autant de liens directs supplémentaires ;

- la définition attribuée par DériF aux lexèmes spécialisés constitue un second moyen, indirect celui-ci, de tisser des relations entre termes. comme *traitement contre la douleur à l'estomac, traitement contre la gastralgie* et *traitement antigastralgique* ;

- nous croisons également la définition littérale d'un lexème composé avec les relations lexicales afin de tester la réutilisabilité des liens de synonymie, hyponymie et approximation entre les termes polylexématiques. L'idée est de vérifier la validité des appariements du type : « *gastralgie* (et donc *douleur à l'estomac*) est un type particulier de *douleur au ventre* ».

La relation d'hyponymie sous-jacente aux compositions néoclassiques de constructions tatpurushas est également un dispositif exploitable en analyse du discours, pour la recherche de liens anaphoriques. Ainsi, *abdominalgie* peut être repris dans la suite d'un texte par son hyperonyme X= °*algie* sous la forme de sa traduction *douleur*. En combinant cette technique avec les informations de la Table des EFs, on multiplie les anaphores candidates : *hysterectomie* peut être repris par *ablation*, traduction de °*ectomie*, ou par la valeur de la tête de Chapitre SNOMED de cet EF : *acte chirurgical* (voir **Fig. 2**).

Les applications multilingues de la démarche présentée (selon la **Fig. 5**) sont pour la plupart immédiatement concevables : question-réponse multilingue, recherche d'information, enrichissement de bases de connaissances translinguistiques… Une autre utilisation est la traduction par voisinage, que j'illustre par le biais de la **Fig. 5.** Chaque étiquette abstraite y regroupe les noms qui ont été effectivement rencontrés dans les corpus spécialisés de chaque langue. A ce sujet, *abdominodynia*$_{EN}$ tout comme *stomachodynie*$_{FR}$ sont des structures qui ne se rencontrent que dans une langue. Cependant, leur traduction est calculable immédiatement via le lien de synonymie qui affecte chaque lexème abstrait ; les traductions indirectes de e.g. *stomachodynie*$_{FR}$ sont donc : *stomacalgia*$_{IT}$, *Gastrodynie*$_{DE}$, *gastrodinìa*$_{ES}$, *gastrodynia*$_{EN}$. Enfin, on voit comment les relations d'hyponymie peuvent être exploitées de manière similaire, pour concevoir des classes lexicales translinguistiques.

7. Références

Altmedica, "Dictionnaire Altmedica, URL:http://www.atmedica.com/ ", Masson.

Amsili, P., "L'antonymie en terminologie: quelques remarques", *Actes de Terminologie et Intelligence Artificielle (TIA)*, 2003, Strasbourg.

Bailly, A., *Dictionnaire Grec-Français, Le Grand Bailly*. Paris, Hachette Education,2000.

Benveniste, E., *Problèmes de linguistique générale II*. Paris, Gallimard,1974.

Bourigault, D., N. Aussenac-Gilles, *et al*., "Construction de ressources terminologiques ou ontologiques à partir de textes: un cadre unificateur pour trois études de cas." *Revue d'Intelligence Artificielle (RIA)*, **18,**(1),2004, p. 87-110.

Bourigault, D., C. Chodkiewic, *et al*., "Construction d'un lexique bilingue des droits de l'homme à partir de l'analyse automatique d'un corpus aligné." *Terminologies nouvelles*, **19**, 1999, p. 70-77.

Bourigault, D. and C. Fabre, "Approche linguistique pour l'analyse syntaxique de corpus." *Cahiers de Grammaire*, **25**, 2000a, p. 131-151.

Bourigault, D. and C. Jacquemin, "Construction de Ressources Terminologiques", *Industrie des Langues*, J.-M. Pierrel, Paris, Hermès, 2000b, p. 215-233.

Bourigault, D., C. Jacquemin, *et al*., *Recent Advances in Computational Terminologies*. Amsterdam/Philadelphia, John Benjamins,2001.

Chiao, Y.-C. and P. Zweigenbaum, "The effect of a general lexicon in corpus-based identification of French-English medical word translations." *Actes MIE*, 2003.

Corbin, D., *Morphologie dérivationnelle et structuration du lexique.* Lille, Presses Universitaires de Lille,1987.

Corbin, D., "Hypothèses sur les frontières de la composition nominale", *Cahiers de grammaire*, Toulouse, ERSS, **17**, 1992, p. 25-55.

Corbin, D., "French (Indo-European: Romance). " *An International Handbook on Inflection and Word Formation*, G. Booij *et al*, New York, Mouton - Walter de Gruyter, **1**, 2004, p.1285-1299.

Corbin, D. and J. Paul, "Aperçus sur la créativité morphologique dans la terminologie de la chimie." *La Banque des mots*, **60**, 1999, p. 51-68.

Cottez, H., *Dictionnaire des Structures du vocabulaire savant. Eléments et modèles de Formation, 4ème édition.* Paris, Dictionnaires Le Robert,1988.

Cruse, D. A., *Lexical Semantics*. London, Cambridge University Press,1986.

Daille, B., "L'identification en corpus d'adjectifs relationnels: une piste linguistique pour l'extraction automatique de terminologie." *TAL*, **42,**(3),2001, p. 815-832.

Dal, G., N. Hathout, *et al*., "Morphologie constructionnelle et traitement automatique des langues: le projet MORTAL", *Lexique 16*, P. Corbin, Villeneuve d'Ascq, Presses Universitaires du Septentrion, **16**, 2004, p. 199-229.

Darmoni, S. J., E. Jarrousse, *et al*., "VumeF: Extending the French part of the UMLS*", The American Medical Informatics Association (AMIA) Symposium*, 2003, Washington, DC, AMIA, p. 824 (poster).

Dolisi, G., "BioTop: Terminologie Médicale: URL http://georges.dolisi.free.fr/Terminologie/"

Dorlands, W. A. N., "Online Dorlands Illustrated Medical Dictionary, 30th edition: URL http://www.mercksource.com/pp/us/cns/cns_home.jsp". Saunders. London, 2002.

Douyère, M., B. Thirion, *et al*., "Doc'CISMEF: un outil de recherche Internet dirigé vers l'enseignement de la médecine." *Document Numérique.*, **7,**(1-2),2003, p. 129-140.

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

Fradin, B., "Combining forms, blends and related phenomena." *Extragrammatical and Marginal Morphology*, U. Doleschal *et al*, München, Lincom, 2000, p. 11-59.

Fradin, B., *Nouvelles approches en morphologie*. Paris, PUF, 2003.

Gaffiot, F., *Le Grand Gaffiot - Dictionnaire Latin-Français*. Paris, Hachette,2000.

Gaussier, E., "General Considerations on Bilingual Terminology Extraction." *Recent Advances in Computational Terminology*, 2001, p. 167-184.

Grabar, N., "Terminologie médicale et morphologie. Acquisition de ressources morphologiques et leur utilisation pour le traitement de la variation terminologique". Paris, Université Paris 6. **Thèse de Doctorat,** 2004.

Grabar, N., "Adaptation de synonymes de la langue générale sans le traitement automatique de termes médicaux", *JFIM*, 2005, Lille

Grabar, N., V. Malaisé, *et al*., *Repérage de relations terminologiques transversales en corpus*, TALN2004, 2004, Fez, Maroc, ATALA.

Grabar, N. and P. Zweigenbaum, "A general method for sifting linguistic knowledge from structured terminologies." *Journal of American Medical Informatics Association*, **7(suppl)**,2000, p. 310-314.

Hahn, U., M. Honeck, *et al*., "Subword segmentation: Leveling out morphological variations for medical document retrieval." *Journal of American Medical Informatics Association*, **8(suppl)**,2001, p. 229-233.

Halliday, M. A. K. and R. Hasan, *Cohesion in English*. London, Longman, 1976.

Haspelmath, M., *Understanding Morphology*. London, Arnold, 2002.

Hathout, N., *L'analogie, un moyen de croiser les contraintes et les paradigmes*, Revue d'Intelligence Artificielle, 2003, p. p.923-934.

Hull, A. D., "Software tools to support the construction of bilingual terminology lexicons." *Recent Advances in Computational Terminology*, 2001, p. 225-244.

Iacobini, C., "Composizione con elementi neoclassici", *La formazione delle parole in italiano*, M. Grossmann *et al*, Tübingen, Niemeyer, 2003, p. 69-96.

Jacquemin, C., "Guessing morphology from terms and corpora", *Proceedings of the 20th Annual International ACM SIGIR*, 1997, Philadelphia, PA, p. 156-167.

Jacquemin, C. and E. Tzoukermann, "NLP for term variant extraction: A synergy of morphology, lexicon, and syntax." *NLP and Information Retrieval*, 1999, p. 25-74.

Krovetz, R., "Viewing morphology as an inference process", *Proceedings of the 16th Annual International ACM-SIGIR*, 1993, p. 191-202.

Levallois, M.-P., Ed. *Larousse Médical (CD-ROM)*. Paris, Larousse, 2000.

Littré, E., *Dictionnaire de Médecine - de chirurgie, de pharmacie de l'art vétérinaire et des sciences qui s'y rapportent*. Paris, J-B Baillère et Fils,1884.

Lovis, C., R. Baud, *et al*., "Medical dictionaries for patient encoding systems: a methodology." *Artificial Intelligence in Medecine*, **14**,1998, p. 201-214.

Lovis, C., P. A. Michel, *et al*., "Word segmentation processing: a way to exponentially extend medical dictionaries", *8th World Congress on Medical Informatics*, 1995, p. 28-32.

Lyons, J., *Semantics*. Cambridge, Cambridge University Press,1977.

Miller, G. A., "Nouns in WordNet: A Lexical Inheritance System." *International Journal of Lexicography*, **3,**(4),1990, p. 245-264.

Namer, F., "Acquisition automatique de sens à partir d'opérations morphologiques en français: études de cas", *TALN*, 2002, Nancy, ATALA-ATILF, p. 235-244.

Namer, F., "Automatiser l'analyse morpho-sémantique non affixale: le système DériF ", *Cahiers de Grammaire*, N. Hathout *et al*, Toulouse, ERSS, **28**, 2003, p. 31-48.

Namer, F. and P. Zweigenbaum, "Acquiring meaning for French Medical Terminology: contribution of Morphosemantics", *Eleventh MEDINFO International Conference*, 2004, San Francisco, CA, p. 535-539.

Nowak, M., "HEXAL Englischwörterbuch Medizin (http://www.hexal.de/subdomains/englisch-woerterbuch-medizin/index.php)", Urban & Fischer Verlag GmbH & Co. KG**,** 2005..

Schulz, S., M. Romacker, *et al*., "Towards a multilingual morpheme thesaurus for medical free-text retrieval", *Proceedings of MIE'99*, 1999, Ljubliana, IOS Press, p. 891-894.

Synapse, "Dictionaire Synapse médical (CDROM)". Toulouse**,** 1997.

Tournier, J., *Introduction descriptive à la lexicogénétique de l'anglais contemporain.* Paris/Genève, Champion-Slatkine,1985.

Tran, T. D., A. Burgun, *et al*., "Acquisition semi-automatique de terminologie bilingue en biologie moléculaire à partir des corpus comparables", *TIA*, 2003, Strasbourg, p. 166-175.

Warren, B., "The importance of combining forms", *Contemporary Morphology*, W. U. Dressler *et al*, Berlin, New York, Mouton - Walter de Gruyter, 1990, p. 111-132.

Xu, J. and B. W. Croft, "Corpus-based stemming using co-occurrence of word variants." *ACM Transactions on Information Systems*, **16,**(1), 1998, p. 61-81.

Zweigenbaum, P., R. Baud, *et al*., "UMLF: a unified medical lexicon for French." *International Journal of Medical Informatics*, **74,**(2-4), 2005, p. 119-124.

Zweigenbaum, P. and N. Grabar, "A contribution of medical terminology to medical language processing resources: Experiments in morphological knowledge acquisition from thesauri", *Proceedings of the Conference on Natural Language Processing and Medical Concept Representation*, 1999, Phoenix, AZ, IMIA WG6, p. 155-167.

Zweigenbaum, P. and N. Grabar, "Corpus-based associations provide additional morphological variants to medical terminologies", *Actes AMIA Annual Fall Symposium 2003*, 2003, Washington, DC, AMIA, 768-772.