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

Fiammetta Namer

UMR "ATILF" CNRS & Université Nancy2

Mots Clefs : morphologie, sémantique, multilinguisme, composition savante, relation lexicale, terminologie médicale

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

Résumé : 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'intéraction 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 décrite 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.

1 Variation terminologique et morphologie

Dans le domaine bio-médical, comme dans toute langue de spécialité, l'extraction de variantes terminologiques constitue un enjeu important (Bourigault *et al.* 2001). L'objectif que la démarche présentée ici vise à atteindre, est de mettre à contribution la morpho-sémantique pour maîtriser l'appariement terminologique bilingue, voire translinguistique. 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 (Gaussier 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 (Daille 2001; Jacquemin, Tzoukermann 1999) d'autres utilisent plutôt des systèmes statistiques fondés sur l'apprentissage de règles (Hathout 2003; Grabar, Zweigenbaum 2000). Enfin, des travaux ont été menés dans le but de faire coopérer morphologie et terminologie bilingue, entre autre par (Chiao, Zweigenbaum 2003). Notre approche se situe plutôt dans la lignée des systèmes basés sur l'application de contraintes. L'analyseur morphologique DériF[1] (Namer 2003), qui constitue l'une des étapes de notre système, a été récemment adapté pour l'analyse du vocabulaire bio-médical, dans le cadre des projets UMLF et Vumef. Comme nous allons le voir, DériF se fonde sur la transposition de connaissances linguistiques pour apparier les lexèmes spécialisés au moyen de relations de trois types : (1) il relie le lexème analysé à sa base (*bactérien, bactérie*) même si celle-ci est d'origine gréco-latine (*hépatique, foie*), (2) ce lien est annoté au moyen d'une pseudo-définition (*amyotrophie* : "absence de développement des muscles") ; (3) grâce à l'adaptation des ressources spécifiques au domaine bio-médical, DériF calcule les relations de synonymie, hyponymie et approximation[2] entre les mots composés dits-savants (Fradin 2000; Warren 1990) : *hystérorragie* y est vu comme synonyme de *métrorragie*, hyponyme de *hystérorrhée*, voisin de *colporragie* ; l'intérêt de ces composés savants est qu'ils constituent à eux seuls près de la moitié des néologismes recensés dans les textes médicaux (Lovis et al. 1998). 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[3] ; 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, sur un lexique d'environ 29000 termes, et donne lieu à l'émergence d'environ 3000 familles lexicales. L'article s'organise comme suit. Nous présentons 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) nous développons la méthode utilisée pour réaliser cet objectif, et nous présentons (§4) les résultats obtenus en français. Ces résultats conduisent naturellement à une discussion et à des perspectives (§5) qui clôtureront cette présentation.

2 Genèse

Comme annoncé en §1, la méthode proposée s'appuie sur la synergie entre un analyseur morphologique basé sur règles, une table qui classe et annote les racines gréco-latines

---

[1] DériF a été développé lors des projets ACI MorTAL (G. Dal, CNRS) et UMLF (P. Zweigenbaum, INSERM), et RNTS VumeF (S. Darmoni, L@stics et JF Forget, Vidal) (Zweigenbaum et al. 2003; Darmoni et al. 2003).

[2] L'approximation (voisinage) subsume les notions de co-méronymie, de co-hyponymie et de compatibilité.

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

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 théories en morphologie lexicale[4] 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, nous nommerons éléments de formation, notés EFs. Un EF partage sa catégorie et son sens avec l'entrée lexicale contemporaine auquel il supplée (ainsi, *gastr-* signifie *estomac*$_{FR}$, 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[5] correspond à *vascul--ite*$_{FR}$, *Vascul--itis*$_{DE}$, *vascol--ite*$_{IT}$ et *vascul--itis*$_{ES/EN}$. (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 : 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*. On établit alors quatre types de relations lexicales entre les EFs : synonymie, notée = (OPT=OPHTALM, *vision*), hyponymie, notée < (BLAST *cellule embryonnaire* < CYT *cellule*), méronymie, notée ← (CORO *pupille* ← OCUL *œil*) et approximation, notée ~ (RHIN *nez* ~ OTO *oreille*).

3 Démarche

Rappelons que notre objectif est l'appariement multilingue de termes médicaux composés savants au moyen de relations lexicales calculées au cours de l'analyse morphologique de ces termes. Notre 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 mots composés d'un corpus à partir de relations de base établies entre les EFs qui constituent ces termes, et réunies 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 2003). Les résultats des analyseurs basés sur contraintes ont l'avantage d'associer à une décomposition hiérarchique la

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

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

définition du mot analysé en fonction du procédé morphologique identifié. Ainsi, le sens d'un lexème obtenu par affixation 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". Comme l'illustre la Fig.1, les EFs apparaissent très fréquemment dans la formation de termes suffixés, préfixés ou composés, toutes langues confondues. Contrairement à l'affixation, la composition construit un nom ou un adjectif en associant deux constituants (chacun peut être un lexème autonome ou un EF, et le cas échéant, d'origine grecque ou latine). Dans les langues romanes, la composition savante (*saxifrage*$_A$) se distingue de la composition dite populaire (*casse-pierre*$_A$) par la place occupée par le constituant tête (noté X), placé à droite du constituant modifieur, noté Y. Le sens du composé est fonction, entre autres, de sa catégorie et du rapport sémantique entre X et Y : le composé peut être de type **(a)** additif (*buccodentaire*$_A$ caractérise ce "qui concerne la bouche : *bucc* et les dents), **(b)** endocentrique (*gastralgie*$_N$ est hyponyme de douleur : *algie*, et affecte l'estomac : *gastr*) ou **(c)** exocentrique (*brachycéphale*$_A$ n'est pas hyponyme de tête : *céphal*(e), mais désigne ce(lui) "qui a une tête : *céphal* courte : *brachy*").

Lang.	Affixation[6]	*traduction*	EF abs.	Composition (type)	*traduction*	EFs abs
IT	epat#ico	*hépatique*	HEPAT	gastro--ectomia **(b)**	*gastrectomie*	GASTR, ECTOMI
FR	bucc#al	*buccal*	BUCC	bucco--dent(aire) **(a)**	*buccodentaire*	BUCC
EN	an#alges(ic)	*analgésique*	ALGES	thermo--algesia **(b)**	*thermoagésie*	THERM, ALGES
DE	Hypo#thermie	*hypothermie*	THERM	Thermo--taxis **(b)**	*thermotaxie*	THERM, TAXI
ES	intra#cefal(ico)	*intracéphalique*	CEPHAL	braqui--cefalo **(c)**	*brachycéphale*	BRACHY,CEPHAL

Figure 1 : Eléments de Formation et procédés morphologiques

3.2 Table multilingue des Eléments de Formation

EF (1)		Instanciation (2) Anglais	Allemand	Français[7]	Italien	Espagnol	CAT (3)	Chapitre SNOMED (4)	Relation lexicale (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)

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. A chaque représentation abstraite d'un EF (col.1) correspondent

[6] Les frontières base-affixe sont marquées '#'

[7] 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.

sa catégorie grammaticale (col.3), la tête de chapitre SNOMED où il apparaît (col.4), et les relations lexicales de base (col.5) dont les symboles sont expliqués en §2, et que l'EF abstrait entretient avec d'autres EFs abstraits présents dans la table. Par exemple, GASTR est synonyme de STOMAC, appartient à ABDOMIN, et a à voir avec HEPAT (*foie*), ENTER (*intestin*) et PANCREAT (*pancréas*). Enfin, la col.2 décrit les instances de l'EF pour chaque langue prise en compte : chaque instance couple la réalisation du symbole abstrait, avec sa traduction. Ainsi, ALGI est instancié par *algia/algy*$_{EN}$: 'pain', *algie*$_{DE}$: 'Schmerz', *algie*$_{FR}$: 'douleur', *algia*$_{IT}$: 'dolore', *algia*$_{ES}$: 'dolor'. L'ajout d'une nouvelle langue dans le système suppose donc uniquement l'insertion d'une nouvelle sous-colonne dans la col.2.

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, 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) [8].

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 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 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 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 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 un analyseur morpho-sémantique, la table multilingue des EFs et les règles

[8] Une version monolingue des règles et de la table des EF est présentée dans (Namer, Zweigenbaum 2004).

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 de synonymie =, hyponymie < et approximation ~. L'analyseur décompose le lexème d'entrée, en identifiant s'il y a lieu, les EFs qui le constituent[9]. 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 29000 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 [10]. La réalisation de la chaîne de traitement en français est rendue possible avant tout par l'existence de l'analyseur morpho-sémantique DériF ("Dérivation en français"). 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 est 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*$_{NOM}$, 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 candidates de l'input (i.e. dans l'exemple, *gastralgie*). Tout d'abord, (**R1**) X est conservé, et Y est remplacé par tous les EFs trouvés dans la table avec lesquels Y est en relation ; ceux-ci sont restitués sous leur forme de réalisation en français, et la relation potentielle est calculée selon **R1** (e.g. dans la Fig. 4 `eql:stomach/algie`[11]) ; ensuite, (**R2**), X est remplacé par chacun de ses synonymes dans la Table des EFs, et l'opération de substitution de Y est identique à ce qui se passe avec **R1** (e.g. `isa:abdomin/odynie`) ; puis les rôles de Y et X sont inversés, lors de l'activation de **R3** (e.g. `see:gastr/ite`) et de **R4** (e.g. `see:stomach/ite`).

---

[9] 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.

[10] 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/Terminologie/Menu/terminologie__medicale_menu.htm)

[11] 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'`.

```
      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:hépat/algie, see:hépat/algo, see:hépat/algés,
see:hépat/odyn, see:pancréat/algie, see:pancréat/algo, see:pancréat/algés,
see:pancréat/odyn, see:stomac/ite, see:stomach/ite)
```

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

A partir de cet ensemble de relations candidates, le système ne garde que les relations concernant les termes attestés. Pour *gastralgie,* et étant donné le contenu du lexique de 29000 entrées du français, on s'attend à ce que seuls les éléments soulignés 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 notre 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 associé sa famille lexicale, regroupant l'ensemble des composés du corpus avec lesquels A entretient l'une des relations de synonymie, hyponymie et approximation. La Fig. 5 reproduit la famille lexicale de *gastralgie*.[12]

```
(11565) gastralgie/NOM (maladie)    "  douleur (du -- liée au) estomac  "
      gastralgie/NOM: synonym of  gastrodynie/NOM, stomacalgie/NOM, stomacodynie/NOM,
stomachodynie/NOM, (gastralgique/ADJ)
      gastralgie/NOM: subtype of  abdominalgie/NOM
      gastralgie/NOM: see also  entéralgie/NOM, entérodynie/NOM, gastrite/NOM,
hépatalgie/NOM, hépatodynie/NOM, pancréatalgie/NOM
```

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 17240 des 29000 lexèmes du corpus de travail[13]. La chaîne de traitement enfin produit plus de 3000 familles lexicales à partir des lexèmes composés du corpus, générant au total des liens entre 7438 ADJs et/ou NOMs distincts.

5 Discussion, perspectives

L'utilisation de la morphologie des mots composés dans le but d'optimiser la recherche d'information en biomédecine a déjà fait l'objet d'expérimentations, entre autre par (Schulz et

[12] D'autres termes sont ajoutés à la famille suivant des crières morphologiques. Notamment, un adjectif relationnel (e.g. *gastralgique*) est considéré comme 'synonyme' de son nom base.

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

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 entre les EFs (donc ne calculent pas de relations lexicales), 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, notre approche 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 les relations lexicales de base dans la table des EFs. Notamment, il faut éviter des annotations trop spécifiques sur des EFs polyréférentielles en médecine. Ainsi, étiqueter LABI (*lèvre*) comme partie-de BUCC (*bouche*) entraînerait un codage pour le moins curieux d'adjectifs comme *inguino-labial*, relatif à la gynécologie.[14]

Les améliorations prioritaires de la démarche présentée (en dehors de l'évolution de DériF, qui ne concerne que le français) passent tout d'abord par l'ajout de nouvelles règles de 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*. 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*. Enfin, mais cela conduira à ajouter un nouveau module monolingue au système, on pourrait générer automatiquement (dans la langue de son choix) les termes correspondant aux relations lexicales possibles d'un terme A, absents du corpus de travail et morphologiquement plausibles: pour le français, cela reviendrait, par exemple (Fig. 4), à générer *abdominodynie, pancréatodynie, stomac(h)ite* qui sont non seulement absents du corpus de travail mais aussi introuvables sur Internet.

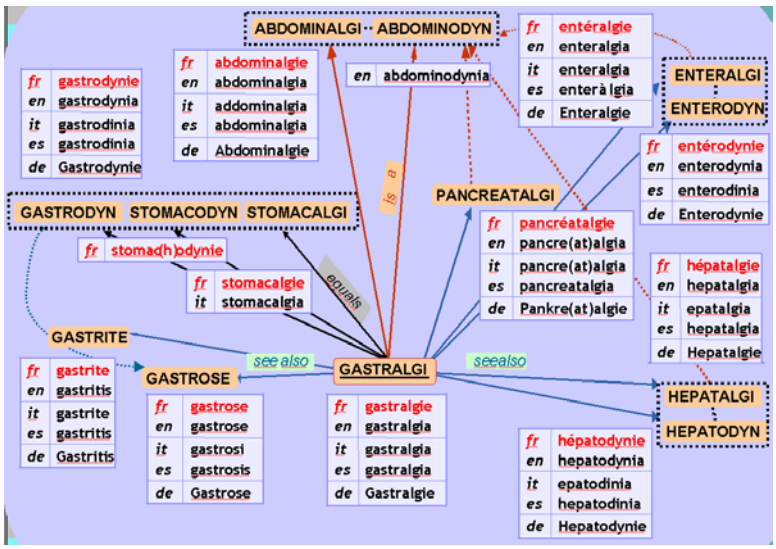

Figure 5 : Une Famille Lexicale Abstraite: celle de GASTRALGI

La réalisation dans d'autres langues que le français[15] de l'approche présentée ne nécessite pratiquement que de disposer d'un parseur morphologique pour chaque nouvelle langue. En un premier temps, celui-ci peut être extrêmement rudimentaire (e.g. un simple raciniseur)

---

[14] Plus généralement, les EFs ambigus nécessitent un traitement par DériF qui exploite des listes d'exception : e.g. *péd* signifie pied (*pédologue, pédicure, pédestre*) ou enfant (*pédagogue, pédophile, pédiatre*).

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

dans la mesure où sa tâche fondamentale est d'identifier les composants X et Y d'un composé savant. Une fois cet analyseur disponible e.g. pour les cinq langues qui nous ont servi à illustrer notre démarche, la chaîne de traitement (à l'exception de l'analyseur) ne manipule plus que des données abstraites. On obtient alors un ensemble de familles lexicales abstraites à l'image de ce qu'illustre la Fig. 5. Dans chacune d'elles, les noms partageant la même structure et les mêmes composants dans différentes langues sont identifiés par une étiquette abstraite ; ce sont ces étiquettes qui sont alors reliées entre elles par les relations lexicales selon le même mécanisme que celui que nous avons décrit pour le français au §4.

6 Conclusion

Nous avons 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. Les résultats obtenus en français sont utilisés pour étendre le système d'extraction de variantes terminologiques à de nouveaux liens simples : *maladie du foie / maladie hépatique* mais aussi plus complexes : *traitement contre la douleur à l'estomac / traitement antigastralgique*. Nous testons également la réutilisabilité des règles de calcul des relations lexicales pour établir des liens de synonymie, hyponymie et approximation entre les termes polylexématiques. L'idée est de vérifier la validité des appariements du type : *douleur à l'estomac* < *douleur au ventre*. On pourrait également envisager une utilisation en analyse du discours des relations d'hyponymie (*abdominalgie* < *douleur*) pour la recherche des liens anaphoriques[16].

Les applications multilingues de la démarche présentée (selon 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, qu'illustre la Fig.5. Chaque étiquette abstraite y regroupe les noms qui ont été effectivement rencontrés dans les corpus spécialisés de chaque langue. On note, à ce sujet, que *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 part de leur étiquette abstraite ; 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.

Références

BOURIGAULT, D., JACQUEMIN, C., et al. 2001. *Recent Advances in Computational Terminologies*. Amsterdam/Philadelphia: John Benjamins.

CHIAO, Y-C., ZWEIGENBAUM, P. 2003. The effect of a general lexicon in corpus-based identification of French-English medical word translations. *Actes MIE*, Amsterdam:.

CORBIN, D. 1987. *Morphologie dérivationnelle et structuration du lexique.* Lille: PUL.

[16] Merci au relecteur anonyme pour cette suggestion.

DAILLE, B. 2001. L'identification en corpus d'adjectifs relationnels: une piste linguistique pour l'extraction automatique de terminologie. In *T.A.L*, 42/3, Paris, Hermès:815-832.

DARMONI, S. J., JARROUSSE, E., et al. 2003. VumeF: Extending the French part of the UMLS. *Proceedings of the AMIA Symposium*, Washington, DC:824

FRADIN, B. 2000. Combining forms, blends and related phenomena. In *Extragrammatical and Marginal Morphology*, München: Lincom Europa: 11-59

GAUSSIER, E. 2001. General Considerations on Bilingual Terminology Extraction. In *Recent Advances in Computational Terminology*. Amsterdam/Philadelphia: John Benjamins: 167-184

GRABAR, N., ZWEIGENBAUM, P. 2000. A general method for sifting linguistic knowledge from structured terminologies. *Journal of AMIA* 7(suppl):310-314.

HAHN, U., HONECK, M., et al. 2001. Subword segmentation: Leveling out morphological variations for medical document retrieval. *Journal of AMIA* 8(suppl):229-233.

HASPELMATH, M. 2002. *Understanding Morphology*. London: Arnold.

HATHOUT, N. 2003. L'analogie, un moyen de croiser les contraintes et les paradigmes. Acquisition de connaissances à partir de dictionnaires de synonymes, *RIA*. 17(5-6),:923-934.

IACOBINI, C. 2003. Composizione con elementi neoclassici. In *La formazione delle parole in italiano*, Tübingen: Niemeyer: 69-96

JACQUEMIN, C., TZOUKERMANN, E. 1999. NLP for term variant extraction: A synergy of morphology, lexicon, and syntax. In *NLP and Information Retrieval*, Boston, Kluwer: 25-74.

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

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

NAMER, F. 2003. Automatiser l'analyse morpho-sémantique non affixale: le système DériF. In *Cahiers de Grammaire*. Toulouse: ERSS: 31-48.

NAMER, F., ZWEIGENBAUM P., 2004 Acquiring meaning for French Medical Terminology: contribution of Morphosemantics. in *11th MEDINFO*. 2004. San Francisco, CA:535-539.

SCHULZ, S., ROMACKER, M., et al. 1999. Towards a multilingual morpheme thesaurus for medical free-text retrieval. *Proceedings of MIE*, Ljubliana, Slovenia: 891-894.

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

WARREN, B. 1990. The importance of combining forms. In *Contemporary Morphology*, Berlin, New York: Mouton - Walter de Gruyter: 111-132.

ZWEIGENBAUM, P., BAUD, R., et al.. 2003. Towards a unified medical lexicon for French. In *Actes MIE*, Amsterdam: IOS Press: 415-420.