UMLF : construction d'un lexique médical francophone unifié

Pierre Zweigenbaum,[a] **Robert Baud,**[b] **Anita Burgun,**[c] **Fiammetta Namer,**[d] **Éric Jarrousse,**[e] **Natalia Grabar,**[a] **Patrick Ruch,**[b] **Franck Le Duff,**[c] **Benoît Thirion,**[f] **Stéfan Darmoni**[f]

[a] *STIM/DSI, Assistance Publique – Hôpitaux de Paris, France*

[b] *DIM, Hôpitaux Universitaires de Genève, Suisse*

[c] *LIM, Centre Hospitalier Régional Universitaire de Rennes, France*

[d] *ATILF, Université Nancy 2, France*

[e] *VIDAL, Paris, France*

[f] *L@STICS, Centre Hospitalier Universitaire de Rouen, France*

Abstract

Medical Informatics has a constant need for basic Medical Language Processing tasks, e.g.*, for coding into controlled vocabularies, free text indexing and information retrieval. Most of these tasks involve term matching and rely on lexical resources: lists of words with attached information, including inflected forms and derived words, etc. Such resources are publicly available for the English language with the UMLS Specialist Lexicon, but not in other languages. For the French language, several teams have worked on the subject and built local lexical resources. The goal of the present work is to pool and unify these resources and to add extensively to them by exploiting medical terminologies and corpora, resulting in a unified medical lexicon for French (UMLF). This paper exposes the issues raised by such an objective, describes the methods on which the project relies and illustrates them with experimental results.*

Keywords

Natural Language Processing; Language; France; Controlled Vocabulary; Algorithms; Funding, Non-US Government

1 Introduction

Des ressources de base pour le traitement automatique de la langue naturelle comme celles que fournit le « Lexique Spécialiste » (Specialist Lexicon, [1]) de l'UMLS constituent un atout majeur pour l'informatique médicale. Des listes de mots avec des informations morphosyntaxiques associées (*e.g.*, *sténoses*, *nom pluriel*) peuvent être utiles pour extraire des

[0] Une description en anglais de ce travail a été présentée à la conférence MIE 2003.

termes à partir de textes médicaux [2], tâche où un étiquetage syntaxique précis est un point clé pour l'analyse correcte des textes. Relier les formes fléchies et les mots dérivés à leurs mots de base (*abdominaux – abdominal*, *diabétique – diabète*), accroît la puissance et la souplesse de l'appariement de termes, par exemple, pour indexer des textes par les concepts de l'UMLS avec MetaMap [2]. Cela améliore également la recherche d'information, en particulier pour les langues morphologiquement riches comme le français, par exemple pour projeter des requêtes vers le MeSH français dans CISMeF [3,4], ce qui permet une navigation « sémantique » au lieu d'une stricte navigation hiérarchique. Plus généralement, l'accès à des bases de connaissances, qu'elles soient indexées avec une terminologie contrôlée (*e.g.*, la base de médicaments VIDAL pour les intranets hospitaliers, www.vidalcim.net) ou pas (*e.g.*, la base de connaissances ADM sur les signes et les maladies [5]), est facilité par des connaissances lexicales. Elles constituent également un atout pour coder des diagnostics dans les classifications CIM-10 ou CIF de l'OMS.

De telles connaissances lexicales sont disponibles pour l'anglais médical dans le Lexique Spécialiste de l'UMLS [1] et pour l'anglais général (ainsi que pour le néerlandais et l'allemand) dans la base CELEX [6]. Un lexique médical a été constitué pour l'allemand [7] et un autre est en cours de constitution pour l'espagnol. En revanche, pour le français, des ressources lexicales existent, mais sont incomplètes et dispersées dans plusieurs équipes ; par exemple, des lexiques français ont été préparés pour divers projets de traitement automatique de la langue médicale [8,9,10,11,12] et incluent des ressources morphosyntaxiques [10,4]. Des méthodes ont été conçues pour apprendre des ressources lexicales à partir de terminologies [13,14,15] de corpus [16,17,18] et par amorçage à partir de lexiques flexionnels existants [19]. Ici encore, ces développements de ressources sont dispersés dans plusieurs équipes. De façon similaire, des outils linguistiques effectuant des traitements au niveau des mots existent dans ces équipes : par exemple, le lemmatiseur français FLEMM [20], ou un étiqueteur de textes médicaux français [21].

Les objectifs du présent travail sont de rassembler et d'unifier ces ressources, de les étendre à l'aide des méthodes mentionnées ci-dessus, et de les rendre disponibles, dans des formats standard, pour la recherche et l'industrie, sous la forme d'une Union de Lexiques Médicaux Francophones (UMLF). Ce travail s'effectue dans le cadre d'un projet subventionné par le ministère français pour la Recherche et l'Enseignement Supérieur (ACI UMLF, subvention #02C0163, 2002–2004). Nous décrivons d'abord les questions soulevées par nos objectifs et résumons les positions initiales du projet (section 2). Nous présentons ensuite des résultats expérimentaux en acquisition lexicale pour illustrer les méthodes sur lesquelles le projet s'appuie (section 3). Nous discutons pour finir des questions et perspectives complémentaires (section 4).

2 Problèmes et méthodes pour le développement d'un lexique médical

La conception du projet nous a permis de soulever un ensemble de problèmes qui doivent être abordés lors de la construction d'un lexique médical. Cette section expose ces problèmes et des solutions initiales, dont certaines font encore l'objet de débats pour le lexique UMLF.

2.1 Couverture

Une première question concerne la frontière entre langue générale et langue médicale. Même si certains mots sont clairement marqués comme appartenant à la langue médicale (*cœur*, *diagnostiquer*, *chirurgical*, *cliniquement*), d'autres sont couramment employés dans la langue médicale mais ne peuvent être considérés comme spécifiques à cette langue (*droite*, *accroissement*). Des facteurs tels que la fréquence et un sens spécifique au domaine seront pris en compte pour mettre au point une règle de décision pragmatique. Un équilibre doit être atteint entre la priorité à donner à des mots clairement médicaux et le soin de ne pas laisser de côté des mots utiles dans les textes médicaux. Par ailleurs, l'estimation d'un nombre de mots à atteindre ou de la couverture lexicale attendue sur des textes médicaux non vus auparavant nécessite au préalable une définition plus précise de ce qu'est un « mot » et une méthodologie sérieuse de mesure de couverture.

Un deuxième problème concernant la couverture est qu'un lexique ne peut jamais être exhaustif, particulièrement dans un domaine aussi vaste que la médecine. Constuire un échantillon représentatif de l'usage de la langue médicale est un enjeu en soi. Ce sera fait de deux façons. D'une part, en collectant des usages effectifs dans de grands corpus diversifiés représentant les spécialités médicales avec leur contact avec des champs proches (comme la biologie, les statistiques, les aspects légaux...), ainsi que divers genres (documents hospitaliers, manuels de cours, sites web médicaux, requêtes à des moteurs de recherche, etc.) [3,22] ; d'autre part, en compilant des vocabulaires médicaux contrôlés tels que thesaurus et classifications : par exemple, la CIM-10 [23], la CIF, le répertoire d'anatomopathologie de SNOMED [24] et l'ensemble de la SNOMED Internationale en français lorsqu'elle sera disponible, le Catalogue commun des actes médicaux (CCAM) français, les thesaurus Vidal (VidalCIM). Le MeSH français [25] et la terminologie WHO Adverse Drug Reaction [26] nécessiteront un traitement particulier car leur forme (majuscules non accentuées) n'est pas appropriée pour l'acquisition lexicale ; néanmoins, l'équipe CISMeF a déjà mis manuellement en casse mixte accentuée 30 % du MeSH ; et son accentuation complète, avec l'aide de méthodes automatiques, est en cours [27]. Un autre cas spécifique est celui de l'ADM [5], une base de connaissances riche qui mêle les propriétés d'un corpus, d'un lexique et d'une terminologie, et qui se trouve également en majuscules non accentuées. Les monographies de médicament du VIDAL constituent une instance supplémentaire de corpus de type « base de connaissances ».

Un facteur supplémentaire de non-exhaustivité dans un lexique est la génération productive de mots dérivés (*bronchiolite*, *bronchiolitique*), de mots composés (*iléojéjunostomie*) et d'acronymes (*ESB*), pour ne citer que les modes de formation de mots les plus courants, ainsi que les noms propres (*Babinski*). Tous doivent être pris en compte ; ceux déjà connus peuvent être listés dans le lexique, et des algorithmes de reconnaissance dynamique de mots inconnus doivent être fournis. Cependant, pour rester dans le cadre des ressources disponibles, le projet se focalise sur les mots dérivés[1].

2.2 Qu'est-ce qu'un mot ?

Une entrée dans le lexique associe des informations à un *lexème* — ce que nous appellons généralement un « mot ». Mais souvent, les lexèmes sont faits de plusieurs unités (*e.g.*, *veine cave*, *placenta prævia*), avec un sens global qui n'est pas complètement dérivable à partir du

[1] Un projet ultérieur, VUMeF (RNTS 2003, coordinateur S.J. Darmoni), doit couvrir la plupart des points restants.

sens des unités individuelles. Comme dans le Lexique Spécialiste de l'UMLS, les critères d'inclusion d'un lexème « multiterme » comprendront sa présence dans un dictionnaire, l'existence d'un synonyme ou d'une abréviation. Par exemple, *infarctus du myocarde* peut s'abréger en *IM* et a un terme « synonyme » *crise cardiaque*. De nouveau cependant, une attitude pragmatique doit être prise étant donné les ressources du projet. La phase actuelle d'UMLF vise à recenser les unités utiles pour la terminologie médicale ; elle ne peut omettre des unités lexicales fortement liées comme *veine cave* ; néanmoins, la description linguistique de base (morpho-syntaxe) d'un terme comme *infarctus du myocarde* est complètement dérivable de celle de *infarctus* et *myocarde*, et son sens est largement compositionnel ; sa présence est donc moins indispensable dans le lexique. Deux type d'entrées supplémentaires sont utiles pour nos objectifs : des affixes (*-al*, *-ique*, *de-*, *in-*, *hyper-*, *trans-*, *brady-*) et des éléments de composition « liés » (*adéno-*, *myo-*, *-carde*), qui ne peuvent apparaître seuls, mais constituent des éléments de base dans la formation de mots. Les derniers se distinguent par une « capacité référentielle » propre. Ils appartiennent tous deux à un espace différent du lexique.

2.3 Quelle information pour chaque lexème ?

Le présent travail se limite à la morphologie et à la syntaxe. Le lexique UMLF associera à chaque mot des informations catégorielles (nom, adjectif, etc.) et morphosyntaxiques (genre, nombre, etc.) lorsque c'est pertinent. Chaque forme fléchie doit être liée à sa ou ses formes canoniques, ou *lemme* (*e.g.*, l'adjectif féminin pluriel *muqueuses* à *muqueux*, le nom pluriel *muqueuses* à *muqueuse*). Chaque mot dérivé doit être relié à son mot de base (*e.g.*, l'adjectif *aortique* à *aorte*). Bien sûr, le sens (types sémantiques, relations hiérarchiques, synonymes non morphologiquement reliés) est ce que le traitement automatique de la langue médicale cherche réellement à traiter, et doit être abordé dans une phase ultérieure (VUMeF). Il sera utile, par exemple, d'associer des types sémantiques (*e.g.*, tirés du Réseau Sémantique de l'UMLS) aux lexèmes. Il faut noter cependant, comme mentionné plus haut, que les terminologies et plus largement le Métathesaurus de l'UMLS [28] traitent déjà certains de ces points. Rappelons également que le Lexique Spécialiste n'inclut pas de tels liens sémantiques.

2.4 Spécification statique ou dynamique d'un lexique

Deux approches principales ont été proposées pour la spécification d'une lexique : une liste explicite (« statique ») de mots ou des règles et des outils de décomposition (« dynamique ») de mots. Les deux approches ont leurs avantages et leurs inconvénients. Une liste explicite de formes fléchies et de mots dérivés peut etre validée par des humains et peut permettre un temps de traitement plus rapide (accès direct dans une table). Des règles appliquées dynamiquement par des outils d'analyse morphologique peuvent traiter des mots inconnus et réduire les besoins en mémoire. L'emploi de méthodes hybrides, utilisant des règles générales complétées par des listes d'exceptions, est une voie qui a montré son efficacité [1,20] aussi bien en lemmatisation qu'en racinisation. La compilation de listes de mots sous forme de transducteurs [29,10] est une autre méthode générale pour obtenir les avantages des deux approches.

3 Expériences et résultats d'acquisition lexicale

Les méthodes pour collecter des connaissances lexicales (*méthodes d'acquisition lexicale*) se divisent en deux grandes classes. D'une part, les méthodes à base de connaissances [10,20] supposent disponibles des connaissances a priori et les appliquent à une source donnée. Par exemple, un lemmatiseur [20] représente des connaissances linguistiques sur le calcul du lemme (forme non fléchie, *e.g.*, *abdominal*) d'une forme fléchie d'un mot (*e.g.*, féminin pluriel *abdominales*). D'autre part, les méthodes de découverte [16,17] supposent que peu de connaissances sont disponibles, et mettent en jeu des processus d'apprentissage. Par exemple, [15] détecte des mots dérivés (*e.g.*, l'adjectif *abdominal*) en relation avec des mots de base (*e.g.*, le nom *abdomen*). Évidemment, ces deux types de méthodes peuvent se compléter (les deux sont illustrées ci-dessous), et sont à employer en sus des ressources lexicales existantes (section 3.4).

3.1 Listes de mots

L'étape initiale dans la compilation d'un lexique consiste à collecter des listes de mots à partir d'échantillons représentatifs de langue médicale : des terminologies et des corpus médicaux (voir section 2.1). L'origine des mots (de quel type de texte) comme leur fréquence doivent être enregistrées. À cette étape du traitement, ce que l'on obtient est des formes (potentiellement fléchies) de mots plutôt que des lemmes non fléchis. De plus, ces mots peuvent inclure du bruit (nombres ou résidus de conversion de pages web, tels que des composants d'URL) qui doit être filtré dans une étape ultérieure. Par exemple, le MeSH français fournit 21 475 formes de mots uniques (58 912 occurrences); une étude de 108 660 requêtes (29 092 uniques) envoyées sur une période de cinq mois au moteur de recherche de CISMeF a observé 21 112 formes uniques (131 570 occurrences). La collecte de 2 338 pages web indexées par CISMeF sous le terme MeSH « Signes et symptômes, états pathologiques », complétées par leurs voisins immédiats sur le web (*[CISMeF-signes]*, au total 9 787 pages), une fois converties en texte simple, a produit 142 545 formes de mots (bruitées ; 5 204 901 occurrences).

3.2 Catégories syntaxiques et connaissances flexionnelles

Le premier type d'information lexicale qui peut être appris est la catégorie syntaxique (nom, adjectif, etc.) de chaque mot. On peut l'obtenir en exploitant le contexte d'usage de chaque mot dans un corpus. Un *étiqueteur morphosyntaxique* [30,31,21] peut non seulement étiqueter les mots dont il dispose dans son lexique interne, mais aussi suggérer l'étiquette la plus probable en contexte pour un mot inconnu. Sous cet angle, il fournit une méthode de découverte. Le lemme (forme non fléchie) de chaque forme de mot peut être obtenu avec un lemmatiseur [20], généralement avec l'aide de sa catégorie syntaxique. Certains lemmatiseurs emploient une approche hybride de type découverte et base de connaissances avec à la fois des règles générales et des exceptions, ce qui leur permet de traiter des mots inconnus [1,20]. *[CISMeF-signes]*, une fois étiqueté avec TreeTagger [31] (www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html) et lemmatisé avec FLEMM [20] (www.univ-nancy2.fr/pers/namer/Telecharger_Flemm.htm), fournit (entre autres catégories) 21 659 adjectif lemmatisés uniques (507 162 occurrences) et 38 025 noms (1 188 574 occurrences). Par effet de bord, ce processus relie les formes fléchies à leur lemme, fournissant ainsi des connaissances flexionnelles.

3.3 Connaissances dérivationnelles

Des listes de mots dérivés avec leur mots de base peuvent être obtenues en appliquant un analyseur morphologique (« raciniseur ») créé manuellement [10,18] à des listes de mots trouvés dans un corpus, de la même façon que l'était un lemmatiseur dans l'étape précédente, pour repérer de nouveaux mots dérivés. Elles peuvent aussi être découvertes à partir de terminologies structurées en comparant des mots similaires dans des termes liés [15]. Par exemple, 1 042 mots dérivés avec leurs mots de base ont été obtenus (après validation) à partir de la CIM-10 et du répertoire d'anatomopathologie de la nomenclature SNOMED Internationale [15]. Finalement, nous avons commencé à expérimenter des méthodes de découverte de mots dérivés suivant les principes proposés par [16]. Les résultats initiaux sur *[CISMeF-signes]* [32] montrent une très bonne précision [33]. Les exemples suivants, qui sont les 26 premiers couples Nom-Adjectif sur les 3 891 proposés dans cette expérience, montrent la variété et la qualité (et illustrent le type de bruit) des dérivations adjectivales proposées par cette méthode : *diabète–diabétique, asthme–asthmatique, urine–urinaire, cellule–cellulaire, kyste–kystique, douleur–douloureux, tuberculose–tuberculeux, grippe–grippal, cancer–cancéreux, vaccin–vaccinal, glomérule–glomérulaire, commission–communautaire, déficience–déficient, allergie–allergique, chirurgie–chirurgical, oesophage–oesophagien, handicap–handicapé, aphasie–aphasique, vaccin–vacciné, veine–veineux, articulation–articulaire, aliment–alimentaire, infection–infectieux, potassium–potassique*. Ces couples de mots sont en cours de validation.

3.4 Fusion et validation d'informations lexicales

Nous avons déjà construit des lexiques médicaux au cours de projets antérieurs. Les ressources préexistantes et nouvellement créées selon les méthodes décrites ci-dessus seront unifiées et validées. Tout d'abord, ces ressources doivent suivre la même « ontologie » pour les informations syntaxiques (étiquettes de catégories et de traits syntaxiques). L'expérience de projets d'unification antérieurs (*e.g.*, l'évaluation GRACE d'analyseurs morphosyntaxiques du français, www.limsi.fr/TLP/grace/) a montré qu'un format commun pouvait être conçu pour représenter de façon unifiée les diverses conventions de modélisation des informations morphosyntaxiques issues de modèles syntaxiques différents. Pour la distribution, plusieurs formats peuvent être générés à partir du format commun. La fourniture d'un format de distribution compatible avec le Lexique Spécialiste de l'UMLS rendra possible l'usage d'outils UMLS avec des ressources francophones. Ensuite, le statut de chaque entrée lexicale doit être documenté : importée de ressources antérieures des équipes participantes, collectée à partir d'un corpus donné, d'une terminologie, etc., validée ou simplement proposée. De cette manière, on assurera la traçabilité de l'origine des entrées lexicales, ce qui permettra de justifier leur inclusion dans le lexique final. Enfin, la validation inclura aussi bien des vérifications automatiques de cohérence qu'une relecture humaine. Entre autres, des entrées multiples pour les mêmes formes fléchies ou les mêmes lemmes pourront être détectées et présentées à la relecture ; les lemmes qui ne diffèrent que d'une lettre peuvent révéler de réelles variantes orthographiques, ou au contraire des fautes d'orthographe dans les documents source. Chaque entrée fera l'objet d'une validation croisée par deux équipes différentes afin d'assurer la meilleure qualité aux ressources produites. L'avis de Sociétés savantes sera demandé selon la nécessité et la possibilité.

4 Discussion et perspectives

Nous avons présenté des méthodes et des expériences initiales pour la collecte d'un grand lexique de mots médicaux français incluant des informations morphologiques pour aider au traitement automatique de la langue naturelle dans diverses tâches telles que l'appariement de termes (codification) et la recherche d'information. Plusieurs aspects demandent certes à être encore affinés, et des types d'information utiles (par exemple, les synonymes) ne peuvent être traités actuellement avec les ressources disponibles ; le projet UMLF doit être considéré comme une première étape vers des ressources lexicales étendues pour faciliter le traitement du français médical. Le projet VUMeF (RNTS 2003) [34] doit couvrir l'étape suivante.

Les objectifs actuels de ce travail laissent également à des travaux futurs la dimension multilingue des lexiques médicaux ; de fait, à part l'anglais [1], des ressources existent aussi pour l'allemand médical [7,35], et un travail sur l'espagnol médical est en cours aux États-Unis à la National Library of Medicine. Néanmoins, certaines des méthodes de découverte présentées ici sont applicables à d'autres langues [15]. L'alignement de langues est aussi une tâche importante, pour laquelle diverses méthodes ont été proposées [13,35,36].

Le site web UMLF témoignera de l'avancée du projet. Une structure de maintenance sera également préparée en parallèle avec le travail technique. Le projet UMLF se terminera en 2004, date à laquelle il rendra ses ressources lexicales librement disponibles à des fins de recherche — et trois ans plus tard pour tous usages.

Références

[1] McCray AT, Srinivasan S, et Browne AC. Lexical methods for managing variation in biomedical terminologies. In: Proc Eighteenth Annu Symp Comput Appl Med Care, Washington. Mc Graw Hill, 1994; pp. 235–9.

[2] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. *J Am Med Inform Assoc* 2001;8(suppl).

[3] Darmoni SJ, Leroy JP, Thirion B, et al. CISMeF: a structured health resource guide. *Methods Inf Med* 2000;39(1):30–5.

[4] Zweigenbaum P, Darmoni SJ, et Grabar N. The contribution of morphological knowledge to French MeSH mapping for information retrieval. *J Am Med Inform Assoc* 2001;8(suppl):796–800.

[5] Seka L, Courtin C, et Le Beux P. ADM-INDEX: an automated system for indexing and retrieval of medical texts. *Stud Health Technol Inform* 1997;43 Pt A:406–10.

[6] Burnage G. *CELEX - A Guide for Users*. Nijmegen: Centre for Lexical Information, University of Nijmegen, 1990.

[7] Weske-Heck G, Zaiß A, Zabel M, et al. The German Specialist Lexicon. *J Am Med Inform Assoc* 2002;8(suppl).

[8] Baud RH, Rassinoux AM, et Scherrer JR. Natural language processing and semantical representation of medical texts. *Methods Inf Med* 1992;31:117–25.

[9] Zweigenbaum P et Consortium MENELAS . MENELAS: an access system for medical records using natural language. *Comput Methods Programs Biomed* 1994;45:117–20.

[10] Lovis C, Baud R, Rassinoux AM, Michel PA, et Scherrer JR. Medical dictionaries for patient encoding systems: a methodology. *Artif Intell Med* 1998;14:201–14.

[11] Bodenreider O et McCray AT. From French vocabulary to the Unified Medical Language System: A preliminary study. In: Cesnik B, Safran C, et Degoulet P, eds, Proc 9 th World Congress on Medical Informatics, 1998.

[12] Zweigenbaum P. Ressources pour le domaine médical : terminologies, lexiques et corpus médicaux. *Lettre de l'ELRA* 2001;6(4):8–11.

[13] Baud RH, Lovis C, Rassinoux AM, Michel PA, et Scherrer JR. Extracting linguistic knowledge from an international classification. In: Pappas C, Maglaveras N, et Scherrer JR, eds, Proceedings of MIE'97, Thessaloniki, Grece. IOS Press, 1997.

[14] Zweigenbaum P et Courtois P. Acquisition of lexical resources from SNOMED for medical language processing. In: Cesnik B, Safran C, et Degoulet P, eds, Proc 9 th World Congress on Medical Informatics, 1998; pp. 586–90.

[15] Grabar N et Zweigenbaum P. A general method for sifting linguistic knowledge from structured terminologies. *J Am Med Inform Assoc* 2000;7(suppl):310–4.

[16] Xu J et Croft BW. Corpus-based stemming using co-occurrence of word variants. *ACM Transactions on Information Systems* 1998;16(1):61–81.

[17] Jacquemin C. Guessing morphology from terms and corpora. In: Actes 20th ACM SIGIR, Philadelphia, PA. 1997; pp. 156–67.

[18] Hathout N, Namer F, et Dal G. An experimental constructional database: the MorTAL project. In: Boucher P, ed, *Many morphologies*. Cascadilla Press, Somerville, MA, 2002; pp. 178–209.

[19] Gaussier E. Unsupervised learning of derivational morphology from inflectional lexicons. In: Kehler A et Stolcke A, eds, ACL workshop on Unsupervised Methods in Natural Language Learning, College Park, Md. juin 1999.

[20] Namer F. FLEMM : un analyseur flexionnel du français à base de règles. *Traitement Automatique des Langues* 2000;41(2):523–47.

[21] Ruch P, Baud R, Bouillon P, et Robert G. Minimal commitment and full lexical disambiguation: Balancing rules and hidden markov models. In: Cardie C, Daelemans W, Nedellec C, et Tjong Kim Sang E, eds, Proc CoNLL-2000 and LLL-2000, Lisbon, Portugal. 2000; pp. 111–4.

[22] Zweigenbaum P, Jacquemart P, Grabar N, et Habert B. Building a text corpus for representing the variety of medical language. In: Patel VL, Rogers R, et Haux R, eds, Medinfo, 2001.

[23] Organisation mondiale de la Santé, Genève. Classification statistique internationale des maladies et des problèmes de santé connexes — Dixième révision, 1993.

[24] Côté RA. Répertoire d'anatomopathologie de la SNOMED internationale, v3.4. Université de Sherbrooke, Sherbrooke, Québec, 1996.

[25] Institut National de la Santé et de la Recherche Médicale, Paris. Thésaurus Biomédical Français/Anglais, 2000.

[26] WHO Collaboration Centre for International Drug Monitoring, Uppsala, Sweden. French translation of the WHO Adverse Reaction Terminology (WHOART), 1997. http://www.who-umc.org/.

[27] Zweigenbaum P et Grabar N. Restoring accents in unknown biomedical words: application to the French MeSH thesaurus. *International Journal of Medical Informatics* 2002:113–26.

[28] Lindberg DAB, Humphreys BL, et McCray AT. The Unified Medical Language System. *Methods Inf Med* 1993;32(2):81–91.

[29] Silberztein M. *Dictionnaires électroniques et analyse automatique de textes : le système INTEX*. Masson, Paris, 1993.

[30] Brill E. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics* 1995;21(4):543–65.

[31] Schmid H. Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK. 1994; pp. 44–9.

[32] Hadouche F. Acquisition de resources morphologiques à partir de corpus. DESS d'ingénierie multilingue, Institut National des Langues et Civilisations Orientales, Paris, 2002.

[33] Zweigenbaum P, Hadouche F, et Grabar N. Apprentissage de relations morphologiques en corpus. In: Daille B, ed, Actes de TALN 2003 (Traitement automatique des langues naturelles), Batz-sur-mer. ATALA, IRIN, juin 2003; pp. 285–94.

[34] Darmoni SJ, Jarrousse E, Zweigenbaum P, et al. Extending the French part of the UMLS. In: Musen M, ed, Actes AMIA Annual Fall Symposium 2003, Washington, DC. AMIA, novembre 2003. *À paraître*.

[35] Schulz S, Romacker M, Franz P, et al. Towards a multilingual morpheme thesaurus for medical free-text retrieval. In: Proceedings of MIE'99, Ljubliana, Slovenia. IOS Press, 1999.

[36] Chiao YC et Zweigenbaum P. Looking for French-English translations in comparable medical corpora. *J Am Med Inform Assoc* 2002;8(suppl):150–4.

Adresse de correspondance

Pierre Zweigenbaum, Mission de recherche en Sciences et Technologies de l'Information Médicale,
STIM/DSI/Assistance Publique–Hôpitaux de Paris, 91, boulevard de l'Hôpital, 75634 Paris Cedex 13, France
et ERM 202 de l'INSERM
E-mail: pz@biomath.jussieu.fr Url: http://www.biomath.jussieu.fr/~pz/