UMLF: A Unified Medical Lexicon for French

Pierre Zweigenbaum,[a1] Robert Baud,[b] Anita Burgun,[c]

Fiammetta Namer,[d] Éric Jarrousse,[e] Natalia Grabar,[a] Patrick Ruch,[b]

Franck Le Duff,[c] Jean-François Forget,[e] Magaly Douyère,[f]

Stéfan Darmoni[f]

[a]STIM/DSI/Assistance Publique–Hôpitaux de Paris,
91, boulevard de l'Hôpital, 75634 Paris Cedex 13, France
{pz,ngr}@biomath.jussieu.fr

[b]DIM, Hôpitaux Universitaires de Genève,
Rue Micheli-du-Crest 24, CH-1211 Genève 14, Switzerland
{Robert.Baud,Patrick.Ruch}@dim.hcuge.ch

[c]LIM, Centre Hospitalier Universitaire de Pontchaillou,
rue Henri Le Guilloux, 35033 Rennes, France
{Anita.Burgun,Franck.Leduff}@univ-rennes1.fr

[d]ATILF & Université Nancy 2,
CLSH, BP 3397, 54015 Nancy Cedex, France
Fiammetta.Namer@univ-nancy2.fr

[e]VIDAL SA,
21 rue Camille Desmoulins, 92789 Issy les Moulineaux Cedex 09, France
{Eric.Jarrousse,Jean-Francois.Forget}@medimedia.fr

[f]L@STICS, Centre Hospitalier Universitaire de Rouen,
1 rue de Germont, 76031 Rouen, France
{Magaly.Douyere,Stefan.Darmoni}@chu-rouen.fr

[1]Corresponding author.

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

Basic natural language resources such as those in the UMLS Specialist Lexicon [1] are a key asset for Medical Informatics. Lists of words with attached morphosyntactic information (*e.g.*, *stenoses*, *noun*, *plural*) can be useful for extracting terms from medical texts [2], where accurate syntactic tagging is instrumental to successful text analysis. Relating inflected forms and derived forms to their base words adds power and flexibility to term matching: *e.g.*, mapping into UMLS with Metamap [2]. This also enhances information retrieval, especially with inflected languages such as French, for instance when mapping into French MeSH in CISMeF [3, 4], allowing 'semantic' navigation instead of a restrictive hierarchical navigation. More generally, access to knowledge bases, whether indexed with controlled vocabularies (*e.g.*, the VIDAL drug knowledge base for hospital intranets, www.vidalcim.net) or not (*e.g.*, the ADM knowledge base on diseases [5]), is facilitated by lexical knowledge. This is also an asset for coding diagnoses into WHO's ICD-10 or ICF classifications.

Such lexical knowledge is available for medical English in the UMLS Specialist Lexicon [1] and for general English (as well as Dutch and German) in the CELEX base [6]. A medical lexicon has been started for German [7] and one is planned for Spanish. In contrast, for the French language, some lexical resources do exist, but they are incomplete and scattered in multiple teams; for instance, French lexicons have been devised for various medical natural language processing (NLP) projects [8, 9], including morphosyntactic resources [4, 8]. Methods have been designed for acquiring lexical resources from terminologies [10, 11], from corpora [12, 13, 14] and by bootstrapping from existing inflectional lexicons [15]. Again, these resource development methods are scattered over different teams. In much the same way, language-aware tools for performing word-level operations exist in these teams: for instance, a French lemmatizer FLEMM [16], or a French medical text tagger [17].

The objectives of the present work are to pool and unify these resources, to complete them using the above-mentioned methods, and to make them widely available, in standard formats, for research and industry, in the form of a Unified Medical Lexicon for French (UMLF). It is performed in the framework of a project funded by the French Ministry for Research and Education (ACI UMLF, grant #02C0163, 2002–2004). We first describe more precisely the issues raised by our objectives and outline the initial positions of the project (section 2). We then present experimental results in lexical acquisition in order to illustrate the methods on which the project relies (section 3). We finally discuss some further issues and perspectives (section 4).

2 Issues and methods in the development of a medical lexicon

The design of the project has allowed us to pinpoint a series of issues that must be addressed when building a medical lexicon. This section exposes these issues and initial solutions, some of which are still debated for the UMLF lexicon.

2.1 Coverage

A first issue is to draw the line between general language and medical language. Although some words are clearly marked as related to the medical domain (*heart*,[1] *diagnose, surgical, clinically*), others are heavily used in medical language but cannot be said to be specific to it (*right*, *enlarged*). Factors such as frequency and domain-specific meaning will be taken into account to design a pragmatic decision rule. A balance must be struck between the priority to be given to clearly medical words and the care not to omit words useful in medical texts. Besides, giving an estimate either for the number of words or for the expected coverage of unseen medical texts must wait for both a more precise definition of 'word' and a serious methodology for measuring coverage.

[1]Although the project works on the French language, for ease of understanding, examples in this paper are given in English or French as suits best.

A second issue for coverage is that a lexicon can never be exhaustive, especially in a large domain such as medicine. Our rationale is to sample medical language use. This will be done in two ways. On the one hand, by collecting diversified text corpora, representing medical specialties as well as their contact with related fields (such as biology, statistics, law...) and representing diverse genres (hospital documents, textbooks, medical web sites, queries to search engines, etc.) [3, 18]. Such corpora display actual language use by health care professional and other medical language users.

On the other hand, by compiling existing controlled medical vocabularies such as thesauri and classifications: *e.g.*, ICD-10, ICF, French SNOMED Microglossary and full French SNOMED when available, French Catalogue of Procedures (CCAM), VIDAL thesauri (VidalCIM). Thesauri and classifications provide a very dense population of medical terms: they may represent rare terms which cannot all be expected to be found in the collected corpora, whatever their size (see, *e.g.*, [19] for insights into the nature of word frequency distributions). Specific provision must be made for the French MeSH and WHO Adverse Drug Reaction terminology whose form (unaccented uppercase letters) is not suitable as is for lexical acquisition. Nonetheless, CISMeF has manually provided 30% of the MeSH with accentuated lowercase letters; machine-aided accentuation of the full MeSH has then been performed, both by learning methods (STIM, [20]) and against other medical lexicons (HUG). The combination of these methods has resulted in an accentuated version of the French MeSH thesaurus. Another specific case is that of ADM [5], a rich knowledge base which mixes properties of a corpus, a lexicon and a terminology, and which is also in unaccented uppercase letters. The VIDAL drug monographs are an additional instance of ‘knowledge-base’-type corpus.

A further factor of non-exhaustivity in a lexicon is the productive generation of derived words (*bronchiolite*, *bronchiolitique*), compound words (*ileojejunostomy*) and acronyms (*BSE*, *ESB*), to cite the most prevalent word formation devices, as well as proper nouns (*Babinski*). All these must be dealt with; those already seen may be listed in the lexicon, and algorithms to help recognize unseen ones dynamically must be provided. In order to keep within resources though, the project focusses on derived words.

2.2 What is a word?

An entry in the lexicon associates information with a *lexeme*—what we generally call a ‘word’. But often enough, lexemes are made of several tokens (*e.g.*, *veine cave*, *vena cava*, *part of speech*), with a global meaning which is not fully derivable from the meanings of the individual tokens. As in the UMLS Specialist Lexicon, criteria for entering a multitoken lexeme will include its presence in a dictionary, the existence of a synonym or of an abbreviation. For instance, *myocardial infarction* can be abbreviated as *MI*, *infarctus du myocarde* has a ‘synonym’ term *heart attack*. Here again though, a pragmatic position must be found given project resources. The current UMLF phase aims at compiling the tokens useful for medical terminology; it cannot drop strongly dependent lexical units such as *veine cave*; however, the basic linguistic description (morpho-syntax) of a term such as *myocardial infarction* is fully derivable from that of *myocardial* and *infarction*, and its meaning is by and large compositional, so that its presence is less mandatory in the lexicon. Two additional kinds of entries are useful for our purposes: affixes (-*al*, -*ique*, *de*-, *in*-) and ‘bound’ compound elements (*myo*-, -*carde*), which cannot occur alone, but are basic elements in word formation. Both belong to a different space in the lexicon.

2.3 Which information for each lexeme?

The present work is limited to morphology and syntax. The UMLF lexicon will provide each word with part-of-speech information (noun, adjective, etc.) and with number and gender features where relevant. Each inflected form must be related to its canonical form(s) or *lemma* (*e.g.*, plural feminine adjective *muqueuses* to *muqueux*, plural noun *muqueuses* to *muqueuse*). Each derived word must be linked to its base word (*e.g.*, adjective *aortique* to *aorta*). Again, meaning (semantic types, hierarchical relations, non-morphologically-related synonyms) is basically what medical NLP aims to deal with, and must be addressed in a later phase. It will be useful, for instance, to assign semantic types (*e.g.*, drawn from the UMLS Semantic Network) to lexemes. It must be noted though, as mentioned above, that terminologies and more broadly the UMLS metathesaurus already address some of these issues. Again, the Specialist Lexicon does not include such semantic links.

2.4 Static or dynamic specification of a lexicon

Two main approaches have been proposed for specifying a lexicon: explicit word lists or decomposition rules and tools; both approaches have their pros and cons. Explicit lists of inflected and derived words can be validated by humans and can yield faster processing time. Rules applied dynamically by morphological analysis tools can deal with unseen words and reduce memory requirements. A hybrid method using general rules together with lists of exceptions has proven effective both for lemmatization and stemming [1, 16]. Finite state transducers are another general method for implementing both static and dynamic recognition of inflected forms and derived words [8].

3 Experiments and results in lexical acquisition

Methods for collecting lexical knowledge (*lexical acquisition methods*) can be divided into two broad classes. On the one hand, knowledge-based methods [8, 16] assume some prior knowledge is available, and apply it to a given source. For instance, a lemmatizer [16] embodies linguistic knowledge about how to compute the lemma (uninflected form, *e.g.*, *abdominal*) of an inflected word form (*e.g.*, feminine plural *abdominales*). On the other hand, discovery methods [12, 13] assume little prior knowledge is available, and involve some learning process. For instance, [11] guesses relations between derived words (*e.g.*, adjective *abdominal*) and base words (*e.g.*, noun *abdomen*). Obviously, these two sorts of methods can complement each other (both are illustrated below), and are to be used on top of existing lexical resources. A processing chain has been implemented which applies a series of existing and new tools to process text corpora and prepare lexical information as stated above.

3.1 Word lists

The initial step in the compilation of a lexicon is to collect word lists from representative samples of medical language: medical terminologies and text corpora (see section 2.1). Both the origin of the words (from which text) and their frequencies must be recorded. At this step of processing, what is obtained is (potentially inflected) word forms rather than uninflected lemmas. Besides, these words may include noise (numbers or residues from Web page conversion, such as URL components) which must be filtered in a later step. For instance, the French MeSH yields 21,475 unique word forms (58,912 tokens); a study of 108,660 queries (29,092 unique) sent over five months to the CISMeF search engine observed 21,112 unique word forms (131,570 tokens). A collection of 2,338 Web pages indexed in CISMeF by the MeSH term 'Pathological Conditions, Signs and Symptoms', completed with their immediate Web neighbors (*[CISMeF-signs]*, total 9,787 pages), once converted to text format, provided 142,545 (noisy) word forms (5,204,901 tokens). The above-mentioned VIDAL drug monographs constitutes yet another corpus *[VIDAL-monographs]* which contains 66,319 word forms (6,823,787 tokens). These two example show a contrast between a diversified corpus *[CISMeF-signs]*, with a large vocabulary, and a more focussed corpus *[VIDAL-monographs]*, with a probably more specialized but at the same time less diverse lexicon.

3.2 Part-of-speech and inflectional knowledge

The first kind of lexical information that can be acquired is the part-of-speech (POS: noun, adjective, etc.) of each word. It can be obtained by exploiting the context of use of each word in a corpus. A *part-of-speech tagger* [17] can not only tag words that are listed in its internal lexicon, but also suggest the most probable tag in context for an unknown word. In that respect, it is a discovery method. The lemma (uninflected form) of each word form can be obtained with a lemmatizer [16], often with the help of its part-of-speech. Some lemmatizers use a hybrid knowledge-based and discovery approach with both general rules and exceptions, which allows them to handle unseen words [1, 16]. *[CISMeF-signs]*, once POS-tagged with TreeTagger [21] (www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html) and lemmatized with FLEMM [16] (www.univ-nancy2.fr/pers/namer/Telecharger_Flemm.htm), displays (among other categories) 21,659 unique, lemmatized adjectives (507,162 tokens) and 38,025 nouns (1,188,574 tokens). *[VIDAL-monographs]*, going through the same process, yields 10,086 unique, lemmatized adjectives (572,603 tokens) and 29,840 nouns (1,867,545 tokens). As a side-effect, this process relates inflected forms to their lemmas, thus providing inflectional knowledge.

3.3 Derivational knowledge

Lists of derived words with their base words can be obtained by applying a hand-crafted morphological analysis tool (linguistic knowledge-based 'stemmer') [8, 14] to lists of words found in a corpus, just as lemmatizers were in the previous step, to spot new derived words. For instance, the DériF morphological analyzer [22] applies to part-of-speech-tagged, lemmatized words as produced in the previous step. It handles affixal derivation for a set of general French suffixes and prefixes, and is being extended to cover more phenomena specific to medical language, including neoclassical compounds [23]. It yields a decomposition of complex words such as *désintoxication/N* (where *N* = noun) into their base words:

désintoxication/N ⇒ [[[*dé* [*in* [*toxique* A] (*er*) V]] A] *tion* N],

(*désintoxication/N*, *désintoxiquer/V*, *intoxiquer/V*, *toxique/A*)

("(Action|result of) *detoxicate*")

This morphological analysis step is included in our current processing chain.

Derived words can also be discovered from structured terminologies by comparing similar words in related terms [11]. For instance, 1,042 derived words with their base words were obtained (after validation) from the French ICD-10 and SNOMED Microglossary for Pathology [11]. We also designed corpus-based methods for the discovery of derived words, inspired by [12]. Initial results on *[CISMeF-signs]* [24] show a very good precision. For instance, the following are the top 26 Noun-Adjective pairs (out of 3,891) collected by this method: *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*. These discovery methods also help to collect derivation rules which, once validated, can be included into knowledge-based tools such as DériF.

3.4 Fusion and validation of lexical information

We already have assembled medical lexicons during the course of former projects. Both pre-existing and newly-produced resources resulting from the above-mentioned methods need to be unified and validated. First, these resources must use the same ‘ontology’ of syntactic information (part-of-speech tags, morphosyntactic features). Experience in previous unification projects (*e.g.*, the GRACE evaluation of French morphosyntactic analyzers, www.limsi.fr/TLP/grace/) has shown that a common format could be designed to represent in a unifying way the various conventions for modeling morphosyntactic information in different syntactic models. For distribution, several formats can be generated from the common format. Providing a distribution format compatible with the UMLS Specialist Lexicon will enable the use of UMLS tools with French resources.

Second, the status of each lexical entry must be documented: imported from former resources of the participating teams, collected from corpus, from terminology, etc., validated or still only proposed. This way, care will be taken to ensure the traceability of lexical entry origin so that inclusion in the final lexicon can be properly motivated. Our processing chain records the corpora and positions of occurrence of each word collected. Finally, validation will involve both automated consistency checking and human review. Among other points, multiple entries for the same inflected forms or lemmas can be detected and presented for human review; lemmas which differ by only one letter may reveal either actual spelling variants or spelling errors in the source documents. All entries will be cross-validated by two different teams to ensure the highest quality to the resulting resources. Advice will also be asked from Medical Societies where needed and possible.

4 Discussion

We have shown methods and initial experiments to collect a large lexicon of French medical words, including morphological information suitable for helping language processing in various tasks such as term matching and information retrieval. Indeed, several aspects still need to be worked out, and useful types of information (*e.g.*, synonyms) cannot be addressed currently for want of larger resources; the present project must be considered as a first step towards extensive lexical resources for easier processing of French medical language. A follow-up project, VUMeF [25], is to cover some of the remaining work.

The current goals of this work also leave for further investigation the multilingual dimension of medical lexicons; as a matter of fact, apart from English in the UMLS Specialist Lexicon [1], resources also exist for medical German [7, 26]. Relations have been established with these groups to foster the sharing of respective experiences. Some of the discovery methods presented here are also applicable to further languages [11]. Language alignment is also an important task, for which various methods have been proposed [10, 26, 27].

5 Conclusion

The UMLF project aims to compile and distribute a French medical lexicon. This paper presented the main issues involved, the path followed to obtain such a lexicon, and the current software tools assembled or developed to implement the chosen methods. A first version of the lexicon is scheduled for alpha-testing within the UMLF consortium in the first quarter of 2004.

The UMLF web site (www.biomath.jussieu.fr/umlf/) will keep track of project progress. Provision for a maintenance structure will also be prepared in parallel with technical work. The UMLF project will end in fall 2004, where it will make its lexical resources freely available for research purposes. Availability for all uses will be granted three years later.

References

[1] McCray AT, Srinivasan S, and Browne AC. Lexical methods for managing variation in biomedical terminologies. In: Proc 18th Annu Symp Comput Appl Med Care, Washington. Mc Graw Hill, 1994:235–9.

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

[9] Zweigenbaum P. Resources for the medical domain: medical terminologies, lexicons and corpora. *ELRA Newsletter* 2001;6(4):8–11.

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

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

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

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

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

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

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

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

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

[19] Baayen H. *Word Frequency Distributions*. Kluwer Academic Publishers, Dordrecht & Boston, 2001.

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

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

[22] Namer F. Acquisition automatique de sens à partir d'opérations morphologiques en français : études de cas. In: Pierrel JM, ed, Proceedings of TALN 2002 (Traitement automatique des langues naturelles), Nancy. ATALA, ATILF, June 2002.

[23] Namer F and Zweigenbaum P. Acquiring meaning for french medical terminology: contribution of morphosemantics. In: Musen M, ed, Proceedings 10 th World Congress on Medical Informatics, San Francisco, Ca. 2004. *Submitted*.

[24] Zweigenbaum P and Grabar N. Corpus-based associations provide additional morphological variants to medical terminologies. In: Musen M, ed, Proceedings AMIA Annual Fall Symposium 2003, Washington, DC. AMIA, November 2003. *To appear*.

[25] Darmoni SJ, Jarrousse E, Zweigenbaum P, et al. Extending the French part of the UMLS. In: Musen M, ed, Proceedings AMIA Annual Fall Symposium 2003, Washington, DC. AMIA, November 2003. *To appear* (poster).

[26] 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.

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