UMLF: a Unified Medical Lexicon for French

Pierre Zweigenbaum,[1] Robert Baud,[2] Anita Burgun,[3] Fiammetta Namer,[4] Éric Jarrousse,[5] Natalia Grabar,[1] Patrick Ruch,[2] Franck Le Duff,[3] Benoît Thirion,[6] Stéfan Darmoni[6]
[1]STIM/DSI, Assistance Publique – Hôpitaux de Paris, France;
[2]DIM, Hôpitaux Universitaires de Genève, Suisse; [3]LIM, Centre Hospitalier Régional Universitaire de Rennes, France ; [4]ATILF, Université Nancy 2, France;
[5]VIDAL, Paris, France; [6]L@STICS, Centre Hospitalier Universitaire de Rouen, France

Lexical resources for medical language, such as lists of words with inflectional and derivational information, are publicly available for the English language with the UMLS Specialist Lexicon. The goal of the UMLF project is to pool and unify existing resources and to add extensively to them by exploiting medical terminologies and corpora, resulting in a Unified Medical Lexicon for French. We present here the current status of the project.[1]

DEVELOPMENT OF A MEDICAL LEXICON

Basic natural language resources such as those in the UMLS Specialist Lexicon[1] are a key asset for Medical Informatics. Beyond the Specialist Lexicon, a medical lexicon has been started for German; for the French language, some lexical resources do exist, but they are incomplete and scattered in multiple teams,[2,3] hence the objectives of the present project.

To build this lexicon, medical language use will be sampled by analyzing large, diversified corpora, representing diverse medical specialties and genres, and by compiling existing controlled medical vocabularies, *e.g.*, ICD-10, ICF, French SNOMED Microglossary and full French SNOMED when available, French Catalogue of Procedures (CCAM), VIDAL thesauri (VidalCIM) as well as reaccented French MeSH. Words in the lexicon will be single words, but also multi-word terms (*"veine cave"*) when such terms are strongly associated. The UMLF lexicon will provide each word with part-of-speech information (noun, adjective, etc.) and with number and gender features where relevant, relating inflected forms to canonical forms and linking derived words to base words (*e.g.*, adjective *"aortique"* to noun *"aorte"*). Further information (compounds, acronyms) will be left for follow-up projects.

LEXICAL ACQUISITION EXPERIMENTS

To collect the target lexical and morphological knowledge, the project will use both methods which already embody linguistic knowledge[2,4] and discovery processes[3] to complement existing lexical resources. As an illustration, word lists have been collected from diverse sources: French MeSH (21,475 unique word forms), queries to the CISMeF search engine (21,112 unique word), medical Web pages (142,545 (noisy) word forms). In corpora, a part-of-speech tagger can suggest the most probable part-of-speech in context for an unknown word. The lemma (uninflected form) of each word form can be obtained with a lemmatizer.[4] Our corpus of Web pages produced (among other categories) 21,659 unique, lemmatized adjectives (507,162 occurrences) and 38,025 nouns (1,188,574 occurrences). Lists of derived words with their base words can be obtained by applying a handcrafted morphological analysis tool[2] to word lists, or they can be discovered from structured terminologies by comparing similar words in related terms;[3] initial experiments with corpus-based discovery of derived words also show a very good precision. Both preexisting and newly-produced resources resulting from the above-mentioned methods will be unified and validated. Providing a distribution format compatible with the UMLS Specialist Lexicon will enable the use of UMLS tools with French resources.

The UMLF project will end in 2004, where it will make its lexical resources freely available for research purposes—and three years later for all uses.

REFERENCES

1. McCray AT, Srinivasan S, and Browne AC. Lexical methods for managing variation in biomedical terminologies. In: Proc Eighteenth Annu Symp Comput Appl Med Care, Washington. Mc Graw Hill, 1994:235–9.
2. 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.
3. 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.
4. Namer F. FLEMM : un analyseur flexionnel du français à base de règles. *Traitement Automatique des Langues* 2000;41(2):523–47.

[1] UMLF is funded by the French Ministry for Research and Education (ACI grant #02C0163, 2002–2004).