

A Twofold Strategy for Translating a Medical Terminology into French

Louise Deléger, PhD¹, Tayeb Merabti, MSc², Thierry Lecrocq, PhD², Michel Joubert, PhD³,
Pierre Zweigenbaum, PhD¹, Stéfan Darmoni, MD, PhD²

¹CNRS, LIMSI, Orsay, F-91403 France ²CISMeF, Rouen University Hospital, France ³LERTIM, Marseille Medical University, France

Abstract

Objective. The goal of this study is to assist the translation of a medical terminology (MedlinePlus) into French. **Methods.** We combined two types of approaches to acquire French translations of English MedlinePlus terms. The first is knowledge-based and relies on the conceptual information of the UMLS metathesaurus. The second method is a corpus-based NLP technique using a bilingual parallel corpus. **Results.** The knowledge-based method brought translations for 611 terms, among which 67.6% were considered valid. The corpus-based approach provided translations for 143 terms of which 71.3% were considered valid. We thus acquired a total of 435 translated terms (51.3%). **Conclusion.** Combining two approaches allowed us to semi-automatically translate more than half of the terminology, while focusing on only one would have provided a more partial translation. From an applicative viewpoint, this French version is now integrated in the catalogue of online health resources CISMeF.

Introduction

Terminologies are useful to a variety of applications, including automated coding, free-text indexing and information retrieval. A large number of medical terminologies exists, but coverage varies according to languages. French, while being fairly well represented (twenty or so health terminologies exists in French, such as MeSH¹, ICD-10² and SNOMED International³), could still benefit from the addition of new terminologies.

The catalogue of online health resources in French (CISMeF)⁴ is an example of an application relying on French-language medical terminological resources. CISMeF was originally indexed on the basis of only one medical terminology (the MeSH thesaurus). Recently, though, the strategy has moved towards the use of the main health terminologies available in French for automatic indexing and information retrieval⁵. In this context, the addition of new French terminologies would be especially useful, for instance through the translation of some of the many existing English-

language standards. This paper focuses on the translation of one of these terminologies: the MedlinePlus Health Topics vocabulary.

Creating a complete translation of a vocabulary is a time-consuming task, which requires skilled and knowledgeable medical translators. Various studies have investigated automatic methods to assist the translation of medical terminologies or create multilingual medical vocabularies. Some rely on existing terminological resources, such as Joubert et al.⁶ who performed mapping of French terminologies to SNOMED CT using the semantic information provided by the UMLS Metathesaurus, Marko et al.⁷ who mapped monolingual medical lexicons using morphological decomposition in order to create a multilingual dictionary and Nyström et al.⁸ who used various parallel terminologies to build an English-Swedish medical dictionary. Other types of methods make use of text corpora to acquire translations of medical terms. These multilingual text corpora can be either parallel (i.e. texts in different languages that are translations of each other) such as those used by Widdows et al.⁹ to match English UMLS terms with their German translations or by Neveol and Ozdowska¹⁰ to find French translations of MeSH terms; or they can be comparable (i.e. texts addressing the same general topic in different languages) as used by Chiao and Zweigenbaum¹¹ to look for French translations of medical terms, Dejean et al.¹² to extend the German version of the MeSH and Morin et al.¹³ to build a Japanese-French terminology.

In this paper, we propose to combine two types of approaches to translate a medical terminology (the MedlinePlus Health Topics) from English to French: a knowledge-based approach relying on the UMLS Metathesaurus, and a corpus-based approach making use of a bilingual parallel corpus.

Material

UMLS

The *Unified Medical Language System*^{14,15} (UMLS) is a repository of biomedical vocabularies developed by the US National Library of Medicine. The UMLS

integrates over 2 million concepts (2,181,676 in the 2009AB version) from 129 biomedical vocabularies.

The UMLS is made up of three main knowledge components, but, for our purpose, we retain only the Metathesaurus: a very large, multi-purpose, and multi-lingual vocabulary database that contains information about biomedical and health related concepts, their various names, and the relationships among them. It is built from the electronic versions of many different thesauri, classifications, code sets, and lists of controlled terms used in patient care, health services billing, public health statistics, biomedical literature indexing and cataloging, and health services research.

Each concept has a unique identifier in the Metathesaurus (Concept Unique Identifier, CUI). This means that the same concept appearing in various terminologies, perhaps with various names and synonyms, has a unique entry in the Metathesaurus. Thus, the concept identifier allows to link the different terminologies included in the UMLS.

MedlinePlus

The MedlinePlus vocabulary is a terminology (included in the UMLS metathesaurus) designed for the MedlinePlus website¹⁶. This website provides a variety of health information to the general public, among which a series of short articles on specific topics, referred to as the *MedlinePlus Health Topics*. The MedlinePlus vocabulary is used to write these topics. It is developed by the National Library of Medicine (NLM) and exists in two languages: English and Spanish. This terminology (version of October 2009) contains 1,982 entries, corresponding to 848 preferred terms and 1,134 entry terms (synonyms).

Methods

Our strategy to translate the MedlinePlus terminology is twofold: it combines a knowledge-based approach on the one hand, and a corpus-based NLP method on the other hand.

Knowledge-based Approach

This method implies that each term to be translated must be included into the Metathesaurus. We use four French terminologies: MeSH, SNOMED International (SNMI), MedDRA¹⁷ and WHO-ART¹⁸. The French version of SNMI is not integrated into the UMLS Metathesaurus. However, its terms are tagged with UMLS Concept Unique Identifiers (CUIs), which comes down to integrating them into the Metathesaurus. The number of preferred terms is as follows:

MeSH: 25,588, MedDRA: 17,867, SNMI: 107,900 and WHO-ART: 1,532. We also use 12,893 CISMef French synonyms added to the MeSH¹⁹. The union of these four terminologies represents 133,326 distinct UMLS concepts, after elimination of duplicates.

The principle of the method is based on the conceptual construction of the UMLS metathesaurus. For each English MEDLINEPlus term included in the UMLS we extract its Concept Unique Identifiers. For example, for the MEDLINEPlus code “T120” corresponding to the English term “Alcoholism”, there are two corresponding UMLS concepts “C0001973” (“Alcoholic Intoxication, Chronic”) and “C0740858” (“Substance abuse problem”). The next step of this method collects for each UMLS concept all French terms corresponding to the given concept, that is all French terms possessing the same CUI. For the example of the MEDLINEPlus term “Alcoholism”, three French terms are associated with the two UMLS concepts.

Corpus-based Approach

The corpus-based approach (described in more detail in Deleger et al.²⁰) relies on a Natural Language Processing technique called *alignment*, which is the matching of units (sentences, words, etc.) that are translations of one another. Here, alignment is performed in a parallel English/French corpus at the sentence and word levels, and medical terms are selected from the results of this process.

The Health Canada Parallel Corpus. We gathered our parallel corpus from a bilingual (English/French) health website : the *Health Canada* website²¹. This website is mostly dedicated to the general public. For this experiment, we randomly selected a subset of 760 pairs of English and French documents (a total of approximately 1.29 million words) from this website. Documents were converted from HTML to text format as a preprocessing step.

Aligning Medical Terms. As a first step, we perform sentence alignment of the corpus. This step is necessary because there is never a full one-to-one correspondence between the sentences of two parallel documents. We use a state-of-the-art sentence aligner called GMA^{22,23}.

We then align words using a suite of tools combining linguistic and statistical knowledge: the ITools^{24,25}. We thus obtain English words paired with their candidate French translations. These words can be either single words (e.g. *anxiety*) or multi-word units (e.g. *mental retardation*). This is especially important since

terminologies are for a large part composed of multi-word expressions.

Since we are only looking for translations of MedlinePlus terms, we then select from the word pairs only those where the English word matches an entry from the MedlinePlus terminology. This matching is done using the MetaMap program²⁶ on the English part of the word pairs. Further filtering is performed by removing verbs (since terminologies mostly contain noun phrases).

A first manual review of the candidate translations is then performed to remove incorrect translations (alignment errors), as well as some obvious inappropriate candidates (linguistically correct translations, but unsuitable medical terms). The reviewed translations are then given to medical experts for validation and inclusion in the French version of MedlinePlus.

Results

The number of translations as well as the number of different translated English terms obtained with each approach are reported in Table 1. The knowledge-based method brought the most translations: 611 MedlinePlus terms were translated, among which 67.6% were actually retained as part of the French version of MedlinePlus. The corpus-based approach provided translations for 143 terms of which 71.3% were considered valid. Combining the two methods allowed us to acquire a total of 435 translated terms corresponding to 51.3% of the MedlinePlus terms. The remaining terms (413 = 48.7%) for which translations were missing were manually translated by a medical expert and a chief librarian.

	Translated terms	Validated translated terms
K-based	611	413 (67.6%)
C-based	143	102 (71.3%)
All	614	435

Table 1: Number of translated terms (K-based = Knowledge-based approach, C-based = Corpus-based approach)

Table 2 shows the number of terms (translated and validated) for which the two approaches proposed concurrent (different) translations. For the same set of terms, the corpus-based approach has a higher acceptance rate than the knowledge-based approach, which means that translations obtained with the corpus-based

approach are more often valid.

Translated terms	Validated (K-based)	Validated (C-based)
58	34 (58.6%)	38 (65.5%)

Table 2: Number of terms for which the two approaches proposed a different translation (concurrent translations)

Table 3 shows the number of terms (translated and validated) for which the two approaches proposed the same translation. The acceptance rate for those terms is slightly lower than the total acceptance rates given in Table 1. This means that, contrary to what could have been expected, redundant translations obtained with both approaches are not more often valid than other translations. We have not come up with a satisfying explanation for this observation.

Translated terms	Validated translated terms
104	68 (65.4%)

Table 3: Number of terms for which both approaches proposed the same translation (redundant translations)

Table 4 gives examples of translations acquired through the corpus-based approach. The last three lines show rejected translations. The first of these translations corresponds to the non-medical sense of the word *labor*. The second one (*MTS* for *STD*) is only used in Canadian French, the standard abbreviation in French being *MST*. The last example is a case where the expression (*ulcère d'estomac*) is used in lay language but is not medically accurate.

English	French
viral hepatitis	hépatite virale
snuff	tabac à priser
generalized anxiety disorder	trouble anxieux généralisé
screening	dépistage
frostbite	engelure
labor	*main d'oeuvre
STDs	*MTS
peptic ulcer	*ulcère d'estomac

Table 4: Examples of acquired translations (* indicates a rejected translation)

Discussion

Our strategy was to rely on two different automatic approaches, each presenting a number of advantages and drawbacks. The knowledge-based approach is straightforward and easy to implement. It also allows to acquire good quality translations since they are already part of existing medical terminologies. With this method we were able to obtain translations for a large part of the MedlinePlus vocabulary. However this approach is heavily dependent on the availability of terminologies within a knowledge base such as the UMLS. Mapping between terms of different languages might vary in coverage depending on the terminology to be translated and on the target language. Here, MedlinePlus is a small terminology and French is a language rather well represented in the UMLS which explains our good results but we can imagine that a language such as Chinese would have a lower coverage, especially when translating a large terminology where some concepts may not be present in any of the existing terminologies.

In contrast, the use of text corpora offers a more extensive field to look for translations. Moreover, for a patient-oriented terminology such as MedlinePlus, the corpus-based approach offers the advantage of providing terms potentially more adapted to patients, since they come from corpora primarily aimed at the general public, while the UMLS metathesaurus brings more technical vocabulary. Indeed, the translations proposed with this method obtained a higher acceptance rate (71.3% vs. 67.6% for all terms and 65.5% vs. 58.6% for the same set of translated terms). For a more specialized vocabulary, though, this approach might bring less relevant terms or might need to use different parallel corpora. A limitation of this approach is the availability of such parallel corpora. Though the number of multilingual texts keeps growing, their availability varies according to domains and languages. Acquiring parallel corpora of specific medical specialties could prove challenging. Another weak point of the experiment is the rather low quantity of acquired translations. This is explained in part by the small size of the corpus (1.2 million words) which is only a subset of the whole Health Canada website (27.7 million words). Future experiments will use a greater part of the corpus. Also, we processed a corpus randomly gathered from the many documents of the website, but it would be interesting to characterize the content of the documents so as to select the most relevant texts and thus process more focused data. A simple way of doing that would be to detect the source terms present in the corpus beforehand and to keep only the docu-

ments containing those terms.

Combining the two approaches is a way of overcoming the limits of both methods by maximizing the sources of translations. It allowed us to semi-automatically translate more than half the MedlinePlus vocabulary, while focusing on only one of these approaches would have provided a more partial translation. Nevertheless, we were still missing translations for a fair part of the terminology (48.7%). This could be resolved in part by processing a larger and more accurate corpus as pointed out previously.

An advantage of our approach is that it provides access to previously translated texts. Instead of starting from scratch, we can re-use previous work and identify attested translations that a human translator might not have thought of, especially if translating terminologies without textual context. Also, it is semi-automatic and saves time compared to a fully manual approach, although this might not be obvious in the case of the MedlinePlus terminology, given its rather small size. The experiment described here is just an example and the approach is meant to be applied to other terminologies as well. For instance, it would be especially interesting for the translation of SNOMED CT: the very large size of this terminology would make the use of a semi-automatic approach clearly beneficial.

From an applicative point of view, this new French version is now integrated in the CISMéF French catalogue of online health resources to be used as an additional terminology allowing automatic multi-terminology indexing²⁷ and multi-terminology information retrieval²⁸. To address these two goals, a multi-terminology health portal²⁹ was developed with an access to 15 health terminologies in French.

In future work, this French version of MedlinePlus will be used in the French Infobutton³⁰ (now known as *CiSMéF InfoRoute*), which is now integrated to the Multi-Terminology Health Portal.

Conclusion

In this paper, we presented a strategy to translate the MedlinePlus terminology into French. This is the first translation of this vocabulary undertaken by a different institution than the NLM (who provided the English and Spanish versions). We combined two methods, a knowledge-based method and a corpus-based Natural Language Processing method, which allowed us to automatically translate over 50% of the terminology. This French version is now used in the CISMéF

catalogue for the indexing and retrieval of health documents on the Web.

References

1. National Library of Medicine. Medical Subject Heading. URL: <http://www.nlm.nih.gov/mesh>.
2. OMS O. Classification statistique internationale des maladies et des problèmes de santé connexes, 1993. Dixième révision.
3. Coté RA, Rothwell DJ, Patolay J, Beckett R, and Brochu L. The systematized nomenclature of human and veterinary medicine: Snomed international. Technical report, College of American Pathologists, 1993.
4. Darmoni S, Leroy J, Thirion B, et al. CISMef: a structured health resource guide. *Meth Inf Med* 2000;39(1):30–5.
5. Darmoni S, Sakji S, Pereira S, et al. Multiple terminologies in an health portal: automatic indexing and information retrieval. In: Artificial Intelligence in Medicine, Lecture Notes in Computer Science, Verona, Italy. Springer, July 2009:255–9.
6. Joubert M, Abdoune H, Merabti T, Darmoni S, and Fieschi M. Assisting the translation of SNOMED CT into French using UMLS and four representative French-language terminologies. In: Proceedings of the AMIA Annual Symposium, 2009:291–5.
7. Markó K, Baud R, Zweigenbaum P, et al. Towards a multilingual medical lexicon. In: AMIA Annu Symp Proc, 2006:534–8.
8. Nyström M, Merkel M, Peterson H, and Ahlfeldt H. Creating a medical dictionary using word alignment: the influence of sources and resources. *BMC Med Inform Decis Mak* 2007;7.
9. Widdows D, Dorrow B, and Chan C. Using parallel corpora to enrich multilingual lexical resources. In: Proceedings of the Third International Conference on Language Resources and Evaluation, Las Palmas, Spain. ELRA, May 2002:240–4.
10. Névéal A and Ozdowska S. Extraction bilingue de termes médicaux dans un corpus parallèle anglais/français. In: Actes de Extraction et Gestion des Connaissances (EGC'05), 2005:655–64.
11. Chiao YC and Zweigenbaum P. Looking for French-English translations in comparable medical corpora. In: Proc AMIA Symp, 2002:150–4.
12. Déjean H, Gaussier E, Renders JM, and Sadat F. Automatic processing of multilingual medical terminology: applications to thesaurus enrichment and cross-language information retrieval. *AIM* February 2005;33(2):111–24.
13. Morin E, Daille B, Takeuchi K, and Kageura K. Bilingual terminology mining – using brain, not brawn comparable corpora. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07), Prague, Czech Republic. 2007:664–71.
14. Lindberg D, Humphreys B, and McCray A. The Unified Medical Language System. *Methods Inf. Med* 1993;32:281–91.
15. <http://www.nlm.nih.gov/research/umls/>.
16. <http://www.nlm.nih.gov/medlineplus/healthtopics.html>.
17. Brown E, Wood L, and Wood S. The medical dictionary for regulatory activities (meddra). *Drug Saf* 1999;2:109–17.
18. World Health Organization. Adverse Reactions Terminology. URL: <http://www.nlm.nih.gov/research/umls/>.
19. Douyère M, Soualmia L, Névéal A, et al. Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway. *Health Info Libr J* dec 2004;21(4):253–61.
20. Deléger L, Merkel M, and Zweigenbaum P. Translating medical terminologies through word alignment in parallel text corpora. *J Biomed Inform* 2009;42(4):692–701.
21. <http://www.hc-sc.ca>.
22. Melamed ID. Bitext maps and alignments via pattern recognition. In: Véronis J, ed, *Parallel Text Processing: Alignment and use of translation corpora*. Dordrecht, 2000.
23. <http://nlp.cs.nyu.edu/GMA/>.
24. Merkel M, Petterstedt M, and Ahrenberg L. Interactive word alignment for corpus linguistics. In: Proceedings of Corpus Linguistics, Lancaster, UK. 2003:533–42.
25. Merkel M and Foo J. Terminology extraction and term ranking for standardizing term banks. In: 16th Nordic Conference of Computational Linguistics, Tartu, Estonia. 2007:349–54.
26. Aronson A. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. In: AMIA Annu Symp Proc, 2001:17–21.
27. Pereira S, Sakji S, Névéal A, et al. Multi-terminology indexing for the assignment of mesh descriptors to medical abstracts in french. In: AMIA Annu Symp Proc, 2009:521–5.
28. Sakji S, Massari P, Letord C, et al. Evaluation of multi-terminology information retrieval in a medical catalog. In: AMIA, 2010. Submitted.
29. <http://dcweb7.chu-rouen.fr/pts>.
30. Darmoni S, Pereira S, Névéal A, et al. French infobutton: an academic and... business perspective. In: AMIA Annu Symp Proc, 2008.