Assisting the Translation of SNOMED CT into French.

Tayeb Merabti, Lina Soualmia, Julien Grosjean, Catherine Letord, Stéfan Darmoni

MEDINFO 2013
C.U. Lehmann et al. (Eds.)

doi:10.3233/978-1-61499-289-9-47

Assisting the Translation of SNOMED CT into French

Tayeb Merabti [a], Lina F. Soualmia [a,b], Julien Grosjean [a], Catherine Letord [a], Stéfan J. Darmoni [a,b]

[a] *CISMeF & TIBS, LITIS EA 4108, Rouen University Hospital, Rouen, France*
[b] *INSERM, Unité Mixte de Recherche en Santé (UMR_S) 872, équipe 20, Paris, France*

Abstract

The objective of this study is to evaluate to approaches assisting the translation of SNOMED CT into French. Two types of approaches were combined: a concept-based one, which relies on conceptual information of the UMLS Metathesaurus and a lexical-based one, which relies on NLP techniques. In addition to the French terminologies (whether included in UMLS or not). Using the concept-based approach, a set of 156,157 (39.4%) SNOMED CT terms were translated to at least one French term from UMLS. Expanded to the French terms from UMLS terminologies translated by CISMeF, 2,548 (+0.7%) additional SNOMED CT terms were translated to at least one French term. Using the lexical-based approach, a set of 145,737 (36.8%) SNOMED CT terms were translated to at least one French term from HeTOP. The qualitative evaluation showed that 44% of the translations were rated as "relevant". Overall, the two approaches have provided the translation of 168,750 (42.6%) SNOMED CT terms into French using different bilingual terminological sources included in UMLS or in HeTOP.

Keywords:

Semantic Interoperability. Mapping, Terminology as Topic, Coding System, Multilingualism.

Introduction

Health care systems use different biomedical terminologies in different languages, but their coverage varies. The French language, while being fairly well represented, could benefit from the addition of new terminologies such as the Foundational Model of Anatomy (FMA) or the Systematized Nomenclature of MEDicine Clinical Terms (SNOMED CT). The catalogue of online health resources in French (CISMeF) [1] is an example of an application based on French-language biomedical terminologies. It was originally indexed on the basis of the Medical Subject Headings (MeSH®) thesaurus. Since 2005, biomedical terminologies available in French have been used for indexing and retrieval. The addition of other existing standards, currently available in English only, would be useful. The SNOMED CT is a good example of such a terminology not yet translated into French. The SNOMED CT in French would be useful to index and to search clinical resources through the CISMeF or any clinical system in French.

Background

Various studies have investigated automatic methods to assist the translation of biomedical terminologies or create multilingual biomedical vocabularies. Some of these methods used rewriting rules to translate biomedical terms: in [2] the authors proposed translating biomedical terms from Portuguese into Spanish. Their method is also applied for information retrieval [3]. However, as stated in [4], the rules used are hand-coded, which renders this approach and makes it nontransferable to other languages and domains. The method proposed in [4] relies on an automatic process that infers rewriting rules from examples. These examples represent a list of paired terms in two studied languages (pair terms from Masson medical dictionary and from the Unified Medical Language Systems (UMLS) Metathesaurus). An automatic method was proposed that relies on machine learning [4]. It can infer transducers from examples of bilingual word pairs without any additional resource or knowledge. In contrast, some methods used existing terminological resources to translate biomedical terminologies: in a previous work [5] a semantic-based method was proposed to assist the translation of SNOMED CT into French. The four French terminologies included in UMLS Metathesaurus were used. Recently, a UMLS-based approach and a corpus-based approach were combined to translate MEDLINEPlus® Topics from English into French [6]. This UMLS-based approach was used in BabelMeSH [7] to automatically translate a query from French, Spanish and Portuguese into English to allow querying of MEDLINE® via PubMed® directly in these languages. In order to create a multilingual dictionary, the authors in [8] mapped monolingual medical lexicons can use morphological decomposition. In [9], the authors proposed a method that uses various parallel terminologies to build an English-Swedish medical dictionary. Other types of methods are based on text corpora to acquire translations of medical terms. Approaches developed in our study are mapping methods developed before and regarding the creation of mappings between terms from different terminologies [10].

The translation of SNOMED CT has been initiated in few countries. The International Health Terminology Standards Development Organization (IHTSDO) maintains a complete Spanish translation of SNOMED CT [11]. Denmark completed a systematic and quality assured translation of the major part of SNOMED CT in 2009 [12]. The Swedish translation was completed in 2010 and both countries used the same workflow that is now an IHTSDO standard for translation. Translation guidelines have been elaborated by the Translation Special Interest Groups of IHTSDO[1]. "Inforoute Santé Canada" (the French-language version of Infoway) operates the translation of SNOMED CT into French [13]. This Canadian Extension of SNOMED CT contains 35,220 active concepts[2] (2011).

[1] http://www.ihtsdo.org/fileadmin/user_upload/Docs_01/About_IHTSDO/Publications/IHTSDO_Translation_Guidelines_v2.00_20100407.pdf

[2] http://www.ihtsdo.org/fileadmin/user_upload/Docs_01/Members/Canada/IHTSDO_Annual_Report_2010_2011_04_10_CAN.pdf

In this study, we proposed combining two approaches to automatically translate the SNOMED CT into French: a concept-based approach relying on the UMLS [14], and a Natural Language Processing (NLP) approach. The latter relies on the 45 biomedical terminologies and ontologies (BMTO) included in the Cross-lingual Health Multiple Terminologies and Ontologies Portal (HeTOP [15]).

The contribution of this paper is to go one-step further than our previous work [5], using concept methods and compare with a lexical-based approach to determine the strengths and weaknesses of each approach.

Material

The UMLS Metathesaurus

The UMLS Metathesaurus [14], developed by the US National Library of Medicine (NLM®), integrates over 2 million concepts (2,669,267 in its 2012 version) from 159 biomedical vocabularies. The MRCONSO table, which lists all UMLS concepts, was used in this study. Only four terminologies of the 159 are included with their French version in the UMLS Metathesaurus: the MeSH thesaurus, the World Health Organization Adverse Reaction Terminology (WHO-ART), the WHO International Classification of Primary Care (ICPC2), and the Medical Dictionary of Regulatory Activities (MedDRA). However, five BMTO that have an existing official French version are included in the UMLS, but without their French version: the International Statistical Classification of Diseases (ICD10), the Systematized Nomenclature of MEDicine (SNOMED Int), Logical Observation Identifiers Names and Codes (LOINC), the International Classification of Functioning, Disability and Health (WHO-ICF) for handicap and the International Classification for Nursing Practice (ICNP). Furthermore, the CISMeF team has partially translated BMTO included in the UMLS only in English: 24,563 synonyms and 689 ambiguous acronyms of the MeSH Descriptors, 163 synonyms of the MeSH Qualifiers, 20,887 MeSH Supplementary Concepts, and, 847 MEDLINEPlus terms and 12,700 FMA terms.

In the next sections, a distinction will be made between "French BMTO Set 1" which corresponds exclusively to French terms included in the UMLS (N_{CUI1}=81,506 (3.7%)), "French BMTO Set 2" which corresponds to all French terminologies in the UMLS integrated with French terms or not (N_{CUI2}= 222,171 (10.09%)), and "French BMTO Set 3" (N_{CUI3}= 266,768 (12.12%)) which corresponds to all French terms from UMLS terminologies with those translated "only" by the CISMeF team.

The Health Multiple Terminologies and Ontologies Portal (HeTOP)

A generic meta-model was designed in order to fit all 45 terminologies into one global structure. The HeTOP [15] is connected to this meta-model to search concepts from all health terminologies available in French (or in English and translated into French) included in this portal and, to browse it dynamically. This allows to:

a) Manual or automatic indexing of resources for the catalogue;
b) Retrieval of resources;
c) Teaching or performing audits in terminology management.

Some terminologies and classifications are included in the UMLS Metathesaurus (N=9) but the majority are not (N=36), e.g. ORPHANET for rare diseases or WHO-ATC for drugs. Currently, HeTOP integrates 1,296,049 concepts in English, 704,166 in French, and 932,095 relations.

The SNOMED CT

The SNOMED CT includes 395,349 concepts, organized hierarchically in its UMLS Metathesaurus 2012 version. The SNOMED CT offers a terminological foundation for Electronic Health Records and other health Information Technology systems. The international release of the terminology is managed by the IHTSDO founded by nine countries. It is currently translated into several languages [16, 17]. The SNOMED CT is organized along hierarchy. The most representative concepts are: disorder (73,006 terms), procedure (53,119 terms) and finding (33,626 terms).

Methods

The strategy to translate the SNOMED CT terminology is twofold: it combines concept-based and lexical-based approaches.

Concept-based approach

The mapping method is as follows: suppose two terms t1 and t2 of two different terminologies, suppose CUI1 and CUI2, the respective projections of t1 and t2 in the Metathesaurus, then t1 and t2 are mapped if: CUI1=CUI2 (in the MRCONSO table which contains concepts names and sources). The algorithm is run sequentially and all possible exact mappings are aligned with each pair of terms.

Lexical-based approach

In this approach, all terms in English from all bilingual terminologies (English and French) were normalized, and we applied an algorithm to find terms in target terminologies that were the most lexically similar. When a correspondence was found, the translation of the English target term was proposed as one possible translation of the SNOMED CT term. This algorithm was exploited in several previously reported studies to map external French and English terminologies to UMLS and HeTOP. In this method, we used the normalization program ("Norm") included in the UMLS [19]. The Normalization process involved stripping genitive marks, transforming plural forms into singular, replacing punctuation, removing stop words, lower-casing each word, breaking a string into its constituent words, and sorting the words into their alphabetic order. A list of all stages for normalizing the SNOMED CT term "Presence of urinary reducing substances - finding" is available in the Figure 1. Mapping used by this approach provided three types of correspondences between all terms:

a. Exact correspondence: if all the words composing the two terms were exactly the same;
b. Single to multiple correspondences: when the source term could not be mapped by one exact target term, but can be expressed by a combination of two or more terms;
c. Partial correspondence: in this type of mapping only part of the source term was mapped to one or more target terms.

Table 1 contains examples of these three types of correspondences. In this work, only exact correspondences were considered. This type of mapping is easy to evaluate in English and the "not exact" correspondence is useful for the translation of English terms into French. For example, based on this approach, the SNOMED CT term "Thymic branches of internal thoracic artery" was normalized into "artery branch internal thoracic thymic", which is mapped to the SNOMED International term "Thymic branches of internal thoracic artery". Finally, the corresponding French SNOMED International term "Rameaux thymiques de l'artère thoracique interne" was subsquently proposed as a possible translation of the English

SNOMED CT term "Thymic branches of internal thoracic artery".

Figure 1 - Example of Normalization process for the SNOMED CT term "Presence of urinary reducing substances -finding"

Remove genitives	Presence of urinary reducing substances - finding
Replace punctuation with spaces	Presence of urinary reducing substances finding
Remove stop words	Presence urinary reducing substances finding
Lowercase	presence urinary reducing substances finding
Uninflect each word	presence urinary reduce substance find
Word order sort	find presence reduce substance urinary

Table 1 - Examples of the three types of mappings using lexical approach

Type of correspondance	SNOMED CT term	French term(s) (*English term(s)*)
Exact	Dolasetron mesylate	Mésilate de dolasétron
Single to Multiple	Left Dorsal scapular artery	Artère scapulaire postérieure (*Dorsal scapular artery*) *and* (+) Gauche (*Left*)
Partial	Abdominal extraperitoneal fascia	Fascia de l'abdomen (*Fascia of abdomen, nos*)

Manual translation of SNOMED CT

Since the integration of the SNOMED CT into the HETOP portal, a manual translation of 4,353 SNOMED CT terms has been performed by a pharmacist (CL). These terms correspond to active ingredients of drugs. For the majority of terms, the expert used automatic mappings provided by the lexical-based approach between SNOMED CT and the International Nonproprietary Name[3]. However, the expert went beyond the translation by adding synonyms and mapping SNOMED CT concepts corresponding to drug commercial names for example, the SNOMED CT term "Aciclovir 5% cream (product)" was mapped to the pharmaceutical specialty "ACICLOVIR MYLAN 5 % cream".

Quantitative evaluation

Coverage of the two approaches according to the number of SNOMED CT terms translated into French was investigated. Both approaches (lexical-based approach limited to the exact correspondence) were compared, using the number of different SNOMED CT terms translated by each approach.

Qualitative evaluation

Evaluation, which was blind to the method used to translate, was performed on 1,414 SNOMED CT translations by the same pharmacist from lexical approach. In order to evaluate the translations, a five level scale for rating their quality was used: (a) "relevant" if the French translation corresponded exactly to the English SNOMED CT term; (b) Broader than (BT-NT) if the French translation was rated as broader than the SNOMED CT English term; (c) Narrower than (NT-BT) if the French translation was rated as narrower than the SNOMED CT English term; (d) "irrelevant" if the French translation was false and (e) "can not say" when the expert could not evaluate the translation using the other levels. For each evaluated translation, the expert can proposed a correct translation even if the French translation was not in any of the terminologies. Examples of each evaluation are listed in Table 2.

Table 2 - Examples of each type of evaluation

English term	Translation into French	Evaluation
Entire upper gastrointestinal tract	tube digestif supérieur	Relevant
Sennoside	Sennosides A et B	BT-NT
Interferon beta-1a preparation (product)	toxine botulinique de type B	NT-BT
Botulinum toxoid type B (substance)	toxine botulinique de type B	irrelevant the correct term: toxoïde de Clostridium botulinum type B, but it doesn't exist in any French terminology)
Complete luxation of lens (disorder)	subluxation du cristallin	CNS (term proposed: Luxation complète du cristallin)

Results

Table 3 describes the number of SNOMED CT translated terms according to three different sets of French BMTO for the two methods (conceptual and lexical). Using the conceptual approach, the results were 15.5% for French BMTO Set 1, 39.4% for French BMTO Set 2 and 40.1% for French BMTO Set 3. The French BMTO Set 3 allowed translation of 2,548 additional SNOMED CT terms. Using the lexical-based approach, a total of 145,737 (36.8%) SNOMED CT terms were translated to at least one French term from HeTOP (see Table 3). The union of the two approaches (conceptual and lexical) provided translation of 168,750 SNOMED CT terms (42.6%) to at least one French term. Compared to the set of 4,353 SNOMED CT terms translated manually, 1,436 (33%) SNOMED CT terms were identical to those translated by the concept-based approach and 1,424 (32.7%) terms were identical to those translated by the lexical-based approach.
Table 4 displays the number of SNOMED CT translated terms according to each approach and each French BMTO, including those not in the UMLS Metathesaurus.

[3] http://www.who.int/medicines/services/inn/en/index.html

Table 3 - Number of SNOMED CT translations found according to each approach and to each set of terms in French

BMO	Conceptual			Lexical	UNION
	French BMO Set 1	French BMO Set 2	French BMO Set 3	HETOP	
Number SNOMED CT	61,370 **15.5%**	156,157 **39.4%**	158,705 **40.1%**	145,737 **36.8%**	168,750 **42.6%**

Table 4 - Contribution in number for each terminology in the SNOMED CT for each approach and the three sources

Source	Terminologies	Conceptual	Lexical
∈ French BMO Set 1	MedDRA	31,461	24,363
	MeSH	14,726	14,657
	WHO-ART	3,149	2,226
	WHO-ICPC2	629	282
∈ French BMO Set 2	ICD-10	10,586	7,606
	ICNP	1,081	2,584
	LOINC	8,448	4,068
	SNOMED Int.	96,699	95,892
	WHO-ICF	307	274
∈ French BMO Set 3	FMA	5,409	5,301
	MEDLINEPlus	659	675
	MeSH SC	3,701	4,487
∉ UMLS	BNP		3,331
	HPO		1,874
	ORPHANET		3,302
	WHO-ATC		2,578

Quantitative evaluation

The lexical approach found 9,559 translations and 23,013 were found only by the conceptual approach according to the French BMTO Set 3.

Qualitative evaluation

For the 1,414 translations of SNOMED CT terms evaluated by the expert, 628 (44%) translations were evaluated as "relevant" (see Table 5). A total of 306 translations were evaluated as "NT-BT". For example, the translation of the SNOMED CT term "Sea bass - dietary" by the French MeSH term "serrans" (Bass) was evaluated as "NT-BT". In 32% of cases translations were evaluated as "irrelevant" (see Table 5).

Table 5 - Evaluation of 1,414 SNOMED CT translations

	Number of translations	Number of SNOMED CT terms
Relevant	628 (44%)	628
BT-NT	28 (2%)	24
NT-BT	306 (22%)	294
Irrelevant	450 (32%)	251
Can Not Say	2	2

Discussion

The goal of this study was to compare two approaches to translate SNOMED CT terms from English into French. The concept-based approach was straightforward and easy to implement. This approach benefited from the knowledge domain included in the UMLS. In spite of the small number of French terminologies extracted from UMLS, the concept-based approach allowed the acquisition of good quality translations. The number of translations has been increased (+0.7%), using French BMTO Set 3. The lexical approach was more difficult to implement but benefited from the large number of French biomedical terminologies included in HETOP but not yet included in the UMLS. Unlike our method, which is a semasiological approach with a linguistic expression as starting point, the approach used in "infoway" [13] is an onomasiological approach based on concept to translated SNOMED CT with a concept as starting point to translate terms. According to Table 4, several translations were provided by terminologies not included into the UMLS, such as ORPHANET and WHO-ATC. Qualitative evaluation showed 44% of translations were rated as "relevant" and more than 28% of translations were rated as "irrelevant". It is difficult however to perform manual evaluation of a large number of translations. Implicitly, evaluation of SNOMED CT terms translated automatically can help us to validate multiple English to French translations, because SNOMED CT contains a very high proportion of all English terminologies included in UMLS or HETOP. The SNOMED CT terms translated manually corresponded to drugs. It was very difficult to perform automatic translation on such terms using a lexical-based approach and also due to the low number of bilingual terminologies and classifications of drugs integrated into the UMLS. Evaluation also showed that in several cases the results were NT-BT or BT-NT. This problem was due to the kind of BMTO used in this study which was either more specialized (FMA) or more general (ICD10) than SNOMED CT. In contrast, the use of approaches such as a corpus-based approach or a statistical-based approach could offer more accurate translations [6]. However, these approaches are limited since such parallel corpora are not widely available and generate low quantities of translations. Nevertheless, a word-by-word translation of terms might be a possible complementary approach. Using UMLS Semantic Types, when both terms are included in the UMLS Metathesaurus, could help in solving problems due to ambiguous acronyms or to terms which are lexically close but with different meaning (e.g. sterile as an "aseptic technique" and sterility as "Infertility").

Conclusion

In this paper, a methodology to translate SNOMED CT terms into French was presented. Two approaches were used, a concept-based one and a lexical-based one. The approaches allowed translating automatically 42.6% of the SNOMED CT terms from English into French. The automatic and manually translations will be integrated into the HeTOP and the majority of these translations will be also validated by the CISMeF experts to improve their quality and will be used to translate other English terminologies into French.

References

[1] Darmoni S, Leroy J, Thirion B, Baudic F, Douyère M, Piot J. CISMeF: a structured health resource guide. Meth Inf Med. 2000;39(1):30-5.

[2] Schulz S,Markó K, Sbrissia E, Nohama P, Hahn U. CognateMapping a heuristic strategy for the semi-supervised acquisition of a Spanish lexicon from a Portuguese leed lexicon. In: Proceedings of the 20th International Conference on Computational Linguistics, COLING'04. Geneva, Switzerland; 2004. p. 813 – 819.

[3] Markó K, Schulz S,Medelyan O, Hahn U. Bootstrapping dictionaries for cross-language information retrieval. In: Proceedings of the 28th International Conference on Research and Development in Information Retrieval, SIGIR05. Salvador, Brasil; 2005. p. 528 – 535.

[4] Claveau V. Translation of biomedical terms by inferring Rewriting Rules. Information Retrieval in Biomedicine: natural language processing for knowledge integration.V Prince MR, editor. IGI - Global; 2009.

[5] Joubert M, Abdoune H, Merabti T, Darmoni S, Fieschi M. Assisting the translation of SNOMED CT into French using UMLS and four representative French-language terminologies. In: Proc. AMIA Symp. 2009; 2009. p. 291–295.

[6] Deléger L, Merabti T, Lecroq T, Joubert M, Zweigenbaum P, Darmoni S. A twofold strategy for translation a medical terminology into French. In: Proc. AMIA Symp. 2010; 2010. p. 152–6.

[7] Liu F, Funtelo P,, Ackerman M. BabelMeSH: Development of a cross-language tool for MEDLINE/PubMed. In: AMIA Annu Symp Proc. 1012; 2006.

[8] Markó K, Baud R, Zweigenbaum P, Borin L, Merkel M, Schulz S. Towards a multilingual medical lexicon. In: AMIA Annu Symp Proc; 2006. p. 534–8.

[9] Nyström M, Merkel M, Peterson H, Ahlfeldt H. Creating a medical dictionary using word alignment: the influence of sources and resources. BMC Med Inform Decis Mak. 2007;7.

[10] Merabti T, Soualmia LF, Grosjean J, Joubert M & Darmoni SJ. Aligning Biomedical Terminologies in French: Towards Semantic Interoperability in Medical Applications. In Book: Medical Informatics, pp. 41-68, InTech, 2012.

[11] Reynoso GA, March AD, Berra CM, Strobietto RT, Barani M, Lubatti M et al. Development of the Spanish Version of the systematized nomenclature of medicine: methodology and main issues, Proc AMIA Symp; 694-8, 2008.

[12] Høy, A. Coming to terms with SNOMED CT® terms: linguistic and terminological issues related to the translation into Danish. In: Budin G, Laurén C, Picht H et al., Terminology Science and Research, dec. 2006.

[13] Fabry P, Lemieux R, Grant A. Vers une version française de la SNOMED CT. In: Risques, Technologies de l'Information pour les Pratiques Médicales. 2009. p. 69-78.

[14] Lindberg D, Humphreys B, McCray A. The unified medical language system. Methods Inf Med. 1993;32(4):281–291.

[15] Grosjean J, Merabti T, Dahamna B, Kergouraly I, Thirion B, Soualmia L, et al. Health multi-terminology portal: a semantic added-value for partient safety. In: PSIP Workshop; 2011. p. 129–138.

[16] Andersen U, Lerche J, Petersen PG, Bernstein K. Adapting SNOMED CT for use in Denmark-the tools and the process of concept based translation. In: Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems; 2007. p. 2613.

[17] Klein G, Chen R. Translation of SNOMED CT-strategies and description of a pilot project. Studies in health technology and informatics. 2009;146:673.

[18] Fung K, Bodenreider O. Utilizing UMLS for semantic mapping between terminologies. In: Proc AMIA Symp; 2005. p. 266–270.

[19] Browne AC, Divita G, Aronson AR, McCray AT. UMLS language and vocabulary tools. AMIA Annu Symp Proc. 2003;p. 798.

Address for correspondence

Tayeb Merabti
CISMeF Team, Rouen University Hospital
1, Rue de Germont – 76000 Rouen, France
Cours Leschevin, Porte 21
E-mail : tayeb.merabti@gmail.com