Assisting the Translation of SNOMED CT into French using UMLS and four Representative French-language Terminologies

**Michel Joubert[a], PhD, Hocine Abdoune[a], Msc, Tayeb Merabti[b], Msc,
Stéfan Darmoni[b], MD, PhD, and Marius Fieschi[a], MD, PHD**
**[a] LERTIM, Faculty of Medicine, University of Aix-Marseille II, Marseille, France, and
[b] CISMeF, University Hospital of Rouen, France**

Abstract

Objective. To provide a semantics-based method to assist the translation of SNOMED CT into French. To do so, we selected four French-language terminologies: ICD-10, SNOMED International, MedDRA, MeSH, as they are dedicated to different uses – epidemiology, clinical medicine, adverse reactions, medical literature, respectively – in order to map them to SNOMED Clinical Terms (CT), and thus associate French terms with SNOMED CT concepts. In this way, we measured the number of SNOMED CT concepts to be found in French-language terminologies. Material and Method. We used the UMLS Metathesaurus. The mapping method was based on the coincidence of identifiers and on the explicit mappings present in the Metathesaurus. Results. The study dealt exclusively with preferred terms (PTs) in the terminologies. The terminologies are mapped with varying success as regards PTs mapped to SNOMED terms (from 52% to 96%). Conversely, 45% of SNOMED CT terms are mapped by uniting the four terminologies. Discussion. A more effective mapping technique than the current method is under consideration. Conclusion. The method presented will be refined. It could certainly provide useful assistance in the translation of SNOMED CT into French. Due to its general nature, it could be used to translate SNOMED CT into other languages than French.

Introduction

A large number of health terminologies are currently available. All were designed for different ends and uses: SNOMED International [1] then SNOMED Clinical Terms [2] for the clinical coding of pathologies and procedures; ICD-10 [3] for the coding of epidemiological data; ICPC [4] for the coding of patient data by GPs; ATC [5] for the coding of drug properties; WHO-ART [6] and MedDRA [7] for the coding of adverse effects; LOINC [8] for the coding of examination results; and MeSH [9] for the indexing of scientific literature. SNOMED CT is now acknowledged to be the reference in terms of health terminologies. SNOMED CT adopted a description logic foundation that has enabled its curators to formally represent concept meanings and relationships [10, 11]. Maintenance and diffusion were entrusted to the IHTSDO non-profit organization [12]. The various members (countries) within the organization are responsible for translating it into their own national language. The body entitled "Inforoute Santé Canada" (the French-language version of Infoway) is currently operating its French translation [13]. At the request of the French Ministry of Health, we are charged with studying possible collaboration with this institution in order to assist with the translation of SNOMED CT into French.

Currently, a few terminologies are already present in a French-language version. It should be noted, however, that French is well represented, behind English and Spanish, in a number of standard terminologies. We have selected four of these – ICD-10, SNOMED International, MedDRA and MeSH – as they were designed for different purposes in an attempt to map them to SNOMED Clinical Terms (CT), and thus associate French terms to SNOMED CT concepts. To do so, we are using UMLS [14], and, more precisely, its Metathesaurus [15]. Various studies have investigated terminology mapping using UMLS [16-18]. Our own study was inspired in part by their results.

Material and Method

UMLS draws on three knowledge resources: the Metathesaurus, the semantic network and the specialist lexicon. In this study, we are using only the Metathesaurus. More specifically, within the latter, we will be using both the MRCONSO table, which lists all the concepts incorporated in the UMLS with no duplication and in which every concept is attributed a unique identifier (CUI), and the MRREL table which describes the relationships, if any, between concepts in the original terminologies. Hence, a single concept can give rise to as many lines in MRCONSO as there are terminologies in which it can be identified, despite having been attributed a unique CUI. In each line of MRCONSO, one can determine whether one is presented with a preferred term (PT) or a synonym. Another UMLS resource that can be utilized for mapping is the explicit mapping relations provided by some source terminologies that are included in the MRREL table, e.g. ICD-9-CM mappings to SNOMED CT. Most of these mappings can be identified by their relationship attributes (e.g. *mapped_from / to*, *primary_mapped_from / to*, *other_mapped_from / to*). When an explicit mapping relationship exists between two concepts, CUI_1 and CUI_2, it is likely that all terms designating CUI_2 may be mapped to terms designating $CUI_{1,}$ whatever the terminologies and whatever the language in which they are formulated. In other words, explicit mappings between two terminologies can be "reused" for other terminologies by means of the UMLS concept structure [18]. In this study, we deal exclusively with PTs from SNOMED CT and from the French-language terminologies. Our study was conducted using the 2008 AA version of UMLS.

SNOMED CT. There are 311,313 SNOMED CT concepts qualified as PTs in the UMLS Metathesaurus. SNOMED CT concepts are either "primitive" or "fully defined". A fully-defined concept can be differentiated from its parent and sibling concepts by virtue of its relationships with other concepts. Otherwise it is primitive. A concept definition is the list of its relationships to other concepts. The nature of the relationship between two concepts may be : "defining" (920,146), "qualifying" (314,681), "historical" (75,387), and "additional" (47,505) (numbers concern pairs of PTs only). For instance, Table 1 lists the ten most frequent defining and additional relations.

Relationship	Frequency
Isa	501,826
Has_finding_site	84,798
Has_associated_morphology	58,077
Has_method	51,190
Part_of	47,505
Has_direct_procedure_site	31,881
Interprets	24,965
Has_causative_agent	21,641
Has_active_ingredient	18,514
Has_dose_form	9,131

Table 1. The ten most frequently found relations in SNOMED CT.

For example, *Acute infarct* and *Myocardium structure* are primitive concepts which serve to define the fully-defined concept *Acute myocardial infarction* by means of the relationship *Has_associated_ morphology* and *Has_finding_site* respectively. Furthermore, *Acute myocardial infarction "Isa" Myocardial infarction* and "*Isa*" *Acute heart disease*. There are 311,313 PTs in SNOMED CT. There are 261,264 (84%) primitive concepts and 50,049 (16%) fully-defined concepts. Concepts are organized in classes, i.e. entities that share common properties. Table 2 shows the number and percentage of concepts in the fourteen most frequently used classes.

Class	PTs	% PTs
Disorder	74,993	24.0
Procedure	50,253	16.1
Finding	32,630	10.5
Organism	27,643	8.9
Body structure	25,478	8.2
Substance	22,767	7.3
Product	18,530	6.0
Qualifier value	8,583	2.8
Event	8,415	2.7
Observable entity	7,749	2.5
Situation	4,863	1.6
Morphological abnormality	4,746	1.5
Physical object	4,489	1.4
Occupation	4,084	1.3

Table 2. Number and percentage of concepts per class in SNOMED CT.

Four French-language terminologies: ICD-10, SNOMED International (SNMI), MedDRA, MeSH. The French versions of ICD-10 and SNMI are not integrated into the Metathesaurus. However, it is easy to map them to the English-language versions thanks to their common code identifiers, which comes down to integrating them into the Metathesaurus. For these four terminologies, we are concerned only with the most precise descriptors, those which are generally used for coding and indexing, and not intermediary descriptors in terminology hierarchies. Here, once again, we are interested only in the PTs. The number of descriptors is as follows: ICD-10: 9,308, SNMI: 107,900, MedDRA: 17,867, MeSH: 24,767. The union of these four terminologies represents 159,842 PTs and, after elimination of duplicates, 137,300 CUIs.

Mapping. The mapping method we used in this study was as follows: suppose two descriptors t_1 and t_2 of two terminologies T_1 and T_2, respectively; suppose CUI_1 and CUI_2, the respective projections of t_1 and t_2 in the Metathesaurus, then t_1 and t_2 are mapped if

- CUI_1=CUI_2 (in MRCONSO), or
- there is an explicit mapping between CUI_1 and CUI_2 (in MRREL).

In this case, T_1 is one of the four French-language terminologies used and T_2 is SNOMED CT. As an application of the above, even if the explicit mapping comes from another terminology T_3 not part of the set of terminologies used, it still applies to t_1 since it is established between CUI_1, to which t_1 is attached, and CUI_2, to which t_2 is attached.

Results

We mapped each of the French-language terminologies to SNOMED CT using the above method in order to assess the presence of each within the latter. Table 3 shows the numbers and percentages of PTs from each terminology mapped to at least one PT from SNOMED CT. The mapping score of the PTs from the union of the four terminologies with SNOMED CT is 82%.

We then went on to map SNOMED CT with the union of the four French-language terminologies with a view to measuring the amount of SNOMED CT mapped to these terminologies. We found 141,068 (45%) SNOMED CT PTs mapped to the union of these four terminologies. Thus, there remain 170,245 (55%) unmapped PTs. Among these, we counted 146,603 (47%) primitive concepts and 23,642 (8%) fully-defined concepts. Table 4 shows the distribution of the unmapped PTs according to the SNOMED CT classes. The percentages are given relative to the number of initial concepts and not to the number of concepts in the classes in order to allow a comparison with Table 2. For instance, there are 24.0% PTs from SNOMED CT in the *Disorder* class, and 8.6% PTs are unmapped.

Terminology	PTs	Mapped PTs	% PTs
ICD-10	9,308	8,949	96
SNMI	107,900	98,590	92
MedDRA	17,867	9,359	52
MeSH	24,767	14,024	57

Table 3. Number and percentage of PTs mapped with SNOMED CT.

Discussion

The results shown here were obtained with raw data found in the UMLS Metathesaurus by using a simple direct mapping method via the coincidence of CUIs in MRCONSO, or via explicit mappings supplied by MRREL. One can note that ICD-10 and SNMI are almost totally mappable in SNOMED CT, and that MedDRA and MeSH are only half mappable (Table 3). Conversely, one observes that the SNOMED CT concepts have 45% of mapped PTs in the united four terminologies. While we have no statistical evidence, this score can be compared to the difference between the number of PTs in the union of the four terminologies (137,300) and the number of mapped SNOMED CT PTs (141,068). The difference between them (3,768) is mainly due to the fact that several different PTs in SNOMED CT are attached to the same CUIs in the Metathesaurus. For instance, the two different PTs *Impending infarction* and *Preinfarction angina* are attached to a same CUI via different SNOMED CT codes.

Class	PTs	% PTs
Disorder	26,683	8.6
Procedure	31,432	10.1
Finding	22,322	7.2
Organism	10,057	3.2
Body structure	15,842	5.1
Substance	8,357	2.7
Product	12,601	4.0
Qualifier value	7,287	2.3
Event	7,056	2.2
Observable entity	6,023	1.9
Situation	3,250	1.0
Morphological abnormality	1,586	0.5
Physical object	3,486	1.1
Occupation	2,709	0.8

Table 4. Distribution of unmapped PTs according to SNOMED CT classes.

If one examines in the unmapped SNOMED CT PTs the percentages of fully-defined concepts and of primitive concepts, one notes that they are roughly identical to those in SNOMED CT (an approximate ratio of 1:6). This fact would lead one to think that this characteristic has no impact on mapping SNOMED CT to other terminologies. One can also observe that it is in the classes with the largest number of PTs (e.g. *Disorder*, *Procedure*, *Finding*) that one finds the largest number of unmapped PTs (Table 4). These are highly important classes which represent 51% of the PTs in SNOMED CT. These three classes alone contain 26% of unmapped initial concepts, or almost half such concepts (55%).

SNOMED CT is more complete than any of the other terminologies used. Specifically, it draws on precise concept definitions that the other terminologies do not contain. For instance, *Acute posterior myocardial infarction* in SNOMED CT has no equivalent in ICD-10. One method for mapping it would consist of providing an approximate mapping rather than a precise one. In this case, the solution would consist of mapping *Acute posterior myocardial infarction* in SNOMED CT to *Acute myocardial infarction* in ICD-10, since *Acute posterior myocardial infarction* "*Isa*" *Acute myocardial infarction* in SNOMED CT. We tested this method using only the parents of concepts from SNOMED CT, and not all the ancestors, in the different hierarchies. So far, the results have been only partially validated. However, we may anticipate in order to show the usefulness of the technique. Thus, we obtain 98% and 99% of PTs from ICD-10 and SNMI, respectively, mapped to SNOMED CT PTs, as compared with the results in Table 3, i.e. 96%

and 92%, respectively. MedDRA and MeSH, for their part, both saw their mapping scores increase from 52% to 67% and from 57% to 78%, respectively. Thus the PTs from the combined four terminologies are mapped 92% to the SNOMED CT PTs versus 82% previously. Conversely, 72% of the SNOMED CT PTs, versus 45% previously, are mapped to PTs from the combined terminologies. If retained, this will constitute a third step in our method: if there is neither a direct mapping between two PTs nor an explicit mapping between them, then one can try to map a father of one of these two PTs in their respective hierarchies by means of the *Isa* relationship.

In our future investigations, we will ensure that good use is made of the *Isa* relationship in SNOMED CT. Two studies have shed significant light on the use of this relationship. As a general rule, it may give rise to misunderstanding as a result of confusions [19], particularly with the *Par_of* relationship, which is also used in SNOMED CT as an additional relationship. Another study has shown the overabundance of *Isa* relationships (Table 1) and the infrequent use in SNOMED CT of qualifying relationships [20], which were introduced into the terminology in order to facilitate post-coordination with interface terminologies rather than to ensure pre-coordination with standard terminologies. For example, the two concepts *Heart disease* and *Acute heart disease* are both present in SNOMED CT, to which the concept *Acute myocardial infarction* is connected by *Isa* relationships. One might have assumed that *Acute myocardial infarction* would be connected to *Heart disease* via the *Isa* relationship qualified by *Acute*, and, as a result in this case, dispense with the *Acute heart disease* concept. The problem here lies in the auditing of SNOMED CT [21-23] which does not fall within the scope of our concerns.

Conclusion

We intend to assess these two methods and endeavour to apply them generally to all the French-language terminologies integrated – directly or indirectly by means of their equivalent English ones – into the UMLS in order to enrich the mappings between French-language terminologies and SNOMED CT. Our experience to date has taught us that a perfect mapping method remains a mirage, all the more so as the different terminologies are not all designed for the same purposes and as priorities will possibly be established according to the classes of terms requiring to be mapped.

The concept-based approach is a key feature of the translation into French of SNOMED CT. The aim is not to translate English terms, but to associate French terms with concepts the meaning of which are described not only by English terms but also by their relationships with other concepts [24]. Our thinking is totally in line with this approach and our method could certainly provide useful assistance. In the absence of a gold standard, it is difficult to assess the quality of the translation achieved. Moreover, it should be borne in mind that different expressions may be used from one country to another on account of local language habits and medical practices. Some terms used in Canada, for instance, are not current in France or Switzerland, and vice versa. This phenomenon is not specific to French as one finds English versions of ICD-10 adapted to the country of usage, e.g. Australia with ICD-10-AM [25] and Canada with ICD-10-CA [26].

Due to its general nature, the proposed method could be used for translations of SNOMED CT into other languages than French, on condition that terminologies expressed in these languages were integrated directly or indirectly into the UMLS Metathesaurus.

Acknowledgements

This study forms part of the InterSTIS project which is partly funded by the French Agency for Research under grant: ANR-07-TECSAN-010. The authors wish to thank the U.S. National Library of Medicine for kindly providing versions of the UMLS.

References

[1] Lussier YA, Rothwell DJ, Cote RA. The SNOMED model: a knowledge source for the controlled terminology of the computerized patient record. Meth Inf Med. 1998; 37(2): 161-4.

[2] Spackman K. SNOMED Clinical Terms Fundamentals. 2007. URL: http://www.ihtsdo.org/fileadmin/user_upload/Docs_01/SNOMED_Clinical_Terms_Fundamentals.pdf

[3] World Health Organization. International Classification of Diseases. URL: http://www.who.int/classifications/icd/en/

[4] Lamberts H, Wood M. International Classification of Primary Care (ICPC). Oxford University Press; 1987.

[5] World Health Organization. The Anatomical Therapeutic Chemical classification. URL: http://www.who.int/classifications/atcddd/en/

[6] World Health Organization. Adverse Reactions Terminology. URL: http://www.umc-products.com/graphics/ 3149.pdf

[7] Brown EG, Wood L, Wood S. The medical dictionary for regulatory activities (MedDRA). Drug Saf. 1999; 20(2): 109-17.

[8] Logical Observation Identifiers Names and Codes. URL: http://loinc.org/

[9] National Library of Medicine. Medical Subject Heading. URL: http://www.nlm.nih.gov/mesh/

[10] Spackman KA, Campbell KE. Compositional concept representation using SNOMED: towards further convergence of clinical terminologies. Proc AMIA Annu Symp 1998: 740-4.

[11] Spackman KA, Dionne R, Mays E, Weis J. Role grouping as an extension to the description logic of Ontylog, motivated by concept modeling in SNOMED. Proc AMIA Annu Symp 2002: 712-6.

[12] International Health Terminology Standards Development Organization.
URL: http://www.ihtsdo.org/

[13] Canada Health Infoway.
URL: http://www.ihtsdo.org/members/update-from-members-2007/canada-english/

[14] Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Meth Inf Med 1993; 32(4): 281-91.

[15] UMLS Metathesaurus. URL:
http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

[16] Cimino J, Johnson S, Peng P, Aguirre A. From ICD9-CM to MeSH using the UMLS: a how-to guide. Proc Annu Symp Comput Appl Med Care 1993: 730-4.

[17] Bodenreider O, Nelson SJ. Beyond synonymy: Exploiting the UMLS Semantics in Mapping Vocabularies. Proc AMIA Annu Symp 1998: 815-9.

[18] Fung KW, Bodenreider O. Utilizing the UMLS for semantic mapping between terminologies. Proc AMIA Annu Symp. 2005: 266-70.

[19] Guarino N. Some ontological principles for designing upper level lexical resources. Proc 1st international conference on language resources and evaluation. 1998: 527-34.

[20] Cornet R. Do SNOMED CT Relationships Qualify? Stud Health Technol Inform. 2008; 136: 785-90.

[21] Bodenreider O, Smith B, Kumar A, Burgun A. Investigating subsumption in SNOMED CT: an exploration into large description logic-based biomedical terminologies. Artif Intell Med. 2007; 39(3): 183-95.

[22] Jiang G, Chute CG. Auditing the Semantic Completeness of SNOMED CT Using Formal Concept Analysis. J Am Med Inform Assoc. 2009; 16(1): 89–102.

[23] Schulz S, Suntisrivaraporn B, Baader F, Boeker M. SNOMED reaching its adolescence: Ontologists' and logicians' health check. Int J Med Inform. 2009; 78: 86-94.

[24] Fabry P, Lemieux R, Grant A. Vers une version française de la SNOMED CT [Towards a French version of SNOMED CT]. Proc 13rd Journées Francophones d'Informatique Médicale; 2009. Springer Verlag. In press.

[25] International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Australian Modification. URL: http://nis-web.fhs.usyd.edu.au/ncch_new/2.aspx#

[26] The Canadian Enhancement of ICD-10. URL: http://secure.cihi.ca/cihiweb/en/downloads/codingclass_icd10enhan_e.pdf

Address for correspondence

Michel JOUBERT
LERTIM
Faculte de Medecine
27 boulevard Jean Moulin
13005 Marseille
France
mjoubert@ap-hm.fr
http://www.lertim.org