User Centred Networked Health Care
A. Moen et al. (Eds.)
IOS Press, 2011

doi:10.3233/978-1-60750-806-9-819

Assisting the Translation of the CORE Subset of SNOMED CT Into French

Hocine ABDOUNE[a], Tayeb MERABTI[b,c], Stéfan J. DARMONI[b,c], Michel JOUBERT[a,1]
[a] *LERTIM, Faculty of Medicine, University of Aix-Marseille 2, France*
[b] *CISMeF, Rouen University Hospital, France*
[c] *TIBS, LITIS EA 4108, Institute of Biomedical Research, University of Rouen, France*

Abstract. Background: the Core Subset of SNOMED CT is part of the UMLS-Core Project dedicated to study problem list vocabularies. SNOMED CT is not yet translated into French. Objective: to propose an automated method to assist the translation of the CORE Subset of SNOMED CT into French. Material: the 2009 AA versions of the CORE Subset of SNOMED CT and UMLS; use of four French-language terminologies integrated into the UMLS Metathesaurus: SNOMED International, ICD10, MedDRA, and MeSH. Method: an exact mapping completed by a close mapping between preferred terms of the CORE Subset of SNOMED CT and those of the four terminologies. Results: 89% of the preferred terms of the CORE Subset of SNOMED CT are mapped with at least one preferred term in one of the four terminologies. Discussion: if needed, synonymous terms could be added by the means of synonyms in the terminologies; the proposed method is independent from French and could be applied to other natural languages.

Keywords. Problem lists, SNOMED CT, UMLS, Translations

1. Introduction

Weed first introduced and has since popularized the concept of the problem-oriented medical record [1]. The problem-oriented record consists of four essential elements: the data base, problem list, detailed plans, and structured progress notes dealing with each of the identified problems. Problem lists data are often used to drive functions other than clinical documentation, e.g. generation of billing codes, supporting clinical research and quality assurance. In an ideal world, everybody should use a single, standardized problem list vocabulary. In reality, most institutions use their own local vocabularies.

The U. S. National Library of Medicine (NLM) started the UMLS-CORE Project to study problem list vocabularies [2]. The Unified Medical Language System (UMLS) is a valuable resource for terminology research. CORE stands for Clinical Observations Recording and Encoding, a mnemonic referring to the capture and codification of clinical information in the summary segments of the medical record such as the problem list, discharge diagnosis and reason for the encounter. The UMLS-CORE

[1] Corresponding author : Michel Joubert, Lertim, Faculté de Médecine, Université de la Méditerranée, 27 boulevard Jean Moulin, 13005, Marseille, France

Project has two goals: 1) to study and characterize the problem list vocabularies of large health care institutions in terms of their size, pattern of usage, mappability to standard terminologies and extent of overlap, and 2) to identify a subset of UMLS concepts that occur with high frequency in problem lists to facilitate the standardization of problem list vocabularies. A CORE Problem List Subset was derived based on datasets from several institutions. The most frequently used terms, about 14'000 in all, represented about 95% of the usage volume in each institution. These were mapped to 6'800 UMLS concepts, which formed the basis of the UMLS-CORE Subset. SNOMED CT covers a high percentage (81%) of the identified UMLS-CORE concepts [3].

Our aim is to propose an automated method to assist a translation of the CORE Subset of SNOMED CT (shortly, CORE Subset in what follows) into French. This study follows a work related to an assistance of an automated translation of SNOMED CT into French [4]. The translation of SNOMED CT is currently being performed in Canada by the Infoway institution in accordance with the IHTSDO organization [5].

2. Material

2.1. Unified Medical Language System

The UMLS project launched by the NLM integrates health terminologies in a single Metathesaurus [6]. To date, the UMLS Metathesaurus contains a hundred terminologies. More specifically, within the Metathesaurus we will be using: the MRCONSO table, which lists all the concepts incorporated in the UMLS with no duplication and in which each concept is attributed a unique identifier (CUI), and the MRREL table which describes explicit relationships, if any, between concepts in the original terminologies. Within MRREL, we only use the following explicit mappings: *primary_mapped_to/from, mapped_to /from*, *other_mapped_to/from* [7].

We worked with the 2009 AA version of UMLS. Our mappings operate exclusively on preferred terms (PTs) of each French-language terminology: SNOMED International (107'900 PTs), ICD10 (9'306 PTs), MeSH (25'186 PTs), et MedDRA (18'209 PTs).

2.2. CORE Subset of SNOMED CT

SNOMED CT is a hierarchical structure of concepts. It contains 310'074 terms in the 2009 version integrated into UMLS. These terms are organized along axes. The most representative axes are: *disorder* (73'006 terms), *procedure* (53'119 terms), *finding* (33'626 terms).

CORE Subset version 2009AA is a set of SNOMED CT concepts which represent the most frequently used (14'000 terms) in the databases of the institutions studied by the NLM [3]. These terms have been mapped by NLM to 6'800 UMLS concepts, and more than 5'000 to SNOMED CT concepts. They are principally distributed along the following axes: *disorder* (3'794 concepts), *finding* (752 concepts), *procedure* (396 concepts).

3. Method

The mapping method is as follows: suppose two terms t_1 and t_2 of two different terminologies, 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: 1) CUI_1=CUI_2 (in MRCONSO), this corresponds to an exact mapping, and/or 2) there is an explicit mapping between CUI_1 and CUI_2 (in MRREL). The algorithm is run sequentially, all the possible mappings, exact and explicit, are tried to align each couple of terms.

When an explicit mapping relationship exists (e.g. SNOMED CT to ICD-9-CM [8]) between two concepts, CUI_1 and CUI_2, it is likely that all terms designating CUI_2 can 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 [9].

4. Results

Table 1 shows the contribution of each of the four French-language terminologies with regard to the three most representative axes of the CORE Subset. For instance, 3'277 terms of SNOMED International map *disorder* concepts of the CORE Subset. They represent 86% of the 3'794 terms of the CORE Subset of this axis.

Table 1: Contribution in number and percentage for each terminology by axis in the CORE Subset.

Terminologies	Disorder	Finding	Procedure
SNOMED Int.	3,277 / 86%	522 / 69%	262 / 66%
ICD10	2,733 / 72%	477 / 63%	7 / 2%
MeSH	2,151 / 57%	364 / 48%	118 / 30%
MedDRA	2,505 / 66%	495/ 66%	162 / 41%

Table 2 shows the the number of PTs of the union of the four French-language terminologies mapped to CORE Subset PTs (concepts) with regard to the three studied axes. For instance, the *disorder* axis shows 3'463 of the union of terminologies mapped to 3'794 CORE Subset concepts, they represent 91% of them. In the end, the method allows the translation of 89% of CORE Subset terms along these three axes.

Table 2: Number of PTs in the union of French-language terminologies aligned by axis with PTs of the CORE Subset.

Axes	# of PTS of French Terminologies	# of PTs of the CORE Subset	% of PTs of the Core Subset
Disorder	3, 463	3, 794	91%
Finding	632	752	84%
Procedure	291	396	73%
Total	4,386	4,942	89%

5. Discussion and Conclusion

Table 1 shows that the contribution of SNOMED International for translating terms is about 80% of the terms of CORE Subset along the three axes, and that ICD10 contribution is 63%. These results may be explained by the fact that 91% of SNOMED International terms are integrated into SNOMED CT, and that 87% of ICD10 terms are also integrated into SNOMED CT [4]. Considering the three axes in Table 2 (*disorder*, *finding*, and *procedure*), it is possible to propose at least one proposal for the translation of 4'386 of the 4'942 CORE Subset terms, that means 89%. Terminologies are integrated into the UMLS Metathesaurus by experts by means of exact and explicit mappings. Then we can expect that terms of different terminologies referring to a same biomedical concept are attached to a same Metathesaurus concept. So, the mapping we operate does not need validation in our mind, because they have been made previously.

With the intent of improving the assistance of the translation of CORE Subset, we would like to propose a set of French-language terms and of synonyms to an original English set. This proposal is based on the construction of the UMLS Metathesaurus itself: the Metathesaurus is a terminology integration system, in which synonymous terms from various terminologies are clustered into concepts, allowing for seamless mapping between terms from different terminologies through a UMLS concept [10, 11]. For instance, The CORE Subset concept "acute myocardial infarction" is translated to *infarctus aigu du myocarde* in the French ICD10, and into the same MedDRA PT with synonyms "acute myocardial infarction, unspecified site**"**, **"**acute myocardial infarction, unspecified site, episode of care unspecified" (expressed in English). Let remark that this concept is not mapped to MeSH.

Moreover, synonymy is a symmetric relationship between terms. In order that transitivity can be applied: a synonymous term of another term is considered a synonym of the synonyms of the latter synonym. Hence, it is possible to build a set of terms for a term made of preferred terms originating from different terminologies and via synonyms in these terminologies. As such, MeSH can largely contribute thanks to its 97'000 synonyms, not counting more than 20'000 French synonyms added by the CISMeF team (Rouen University Hospital, France), not yet integrated in the French translation of MeSH.

As previously proposed for assisting the translation of SNOMED CT into French [4], our method could be improved by exploiting hierarchical relationships within some terminologies and propose more generic terms for the translation of more specific ones when exact and explicit mappings are not successful. This refinement seems promising but collides with two difficulties: 1) it requires a human expertise to validate a translation proposal, and 2) some research studies have shown the possible confusion that may occur in some terminologies in the interpretation of hierarchical relationships, notably between *IS_A* and *PART_OF* relationships [12, 13]. Moreover this kind of inheritance due to hierarchies does not apply to the concept of synonymy described above.

The automated method we propose for assisting the translation of the CORE Subset terms is not dependent on French, since it works at a conceptual level and not at a lexical one. Hence, it can be reused for another natural language than French, on condition that terminologies in this language are sufficiently integrated in the Metathesaurus.

Acknowledgements: The authors thank the National Library of Medicine of the United States who provided them with the UMLS knowledge sources and the CORE Subset of SNOMED CT. The authors are also grateful to Richard Medeiros, Rouen University Hospital Medical Editor, for editing the manuscript.

References

[1] Weed LL. Medical records that guide and teach. *N Engl J Med* 1968; 278: 593-600 and 652-7.
[2] Fung KW, Mc Donald C, Strinivasan S. The UMLS-CORE project: a study of the problem list terminologies used in large healthcare institutions. *JAMIA* 2010; 17(6): 675-80.
[3] The CORE Problem List Subset of SNOMED CT. http://www.nlm.nih.gov/research/umls/Snomed/core_subset.html
[4] 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. *Proc. AMIA Annu Symp 2009*; 2009:291-5.
[5] Canada Health Infoway. http://www.ihtsdo.org/members/ca00/
[6] National Library of Medicine. UMLS Metathesaurus. http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html
[7] Fung KW, Bodenreider O. Utilizing the UMLS for semantic mapping between terminologies. *Proc AMIA Annu Symp*. 2005: 266-270.
[8] Imel M. A closer Look: The SNOMED Clinical Terms to ICD-9-CM Mapping. *Journal of AHIMA* 2002; 73; 66-69.
[9] Bodenreider O, Nelson SJ. Beyond synonymy: Exploiting the UMLS Semantics in Mapping Vocabularies. *Proc AMIA Annu Symp* 1998: 815-9.
[10] McCray AT, Nelson SJ. The Representation of meaning in the UMLS. *Methods Inf Med* 1995; 3:193-201.
[11] Bodenreider 0. Biomedical Ontologies in Action: Role in Knowledge Management, Data Integration and Decision Support. *Yearb Med Inform*, 2008: 67-79.
[12] Cimino JJ, Min H, Perl Y. Consistency across the hierarchies of the UMLS Semantic Network and Metathesaurus. *J Biomed Inform* 2003; 36: 450-61.
[13] Ceusters W, Smith B, Kumar A, et al. Mistakes in medical ontologies: where do they come from and how can they be detected? *Stud Health Technol Inform* 2004; 102: 145–63.