Development of the Spanish Version of the Systematized Nomenclature of Medicine: Methodology and Main Issues.

Guillermo A. Reynoso[1], Alan D. March[2], Carolina M. Berra[1], Rosana P. Strobietto[1], Mariela Barani[1], Matías Iubatti [1], María P. Chiaradio[1], Debora Serebrisky[1], Andrea Kahn[1], Oscar A. Vaccarezza[1], Jorge L. Leguiza[2], Marcelo Ceitlin[1], Daniel A. Luna[2,3], Fernan G. Bernaldo de Quirós[3], María I. Otegui[3], María C. Puga[3], Mara Vallejos[3]

[1]Centro de Educación Médica e Investigaciones Clínicas (CEMIC), Buenos Aires, Argentina; [2]Departamento de Informática Médica, Universidad del Salvador, Buenos Aires, Argentina; and [3]Plan de Salud, Hospital Italiano de Buenos Aires, Buenos Aires, Argentina.

ABSTRACT

This presentation features linguistic and terminology management issues related to the development of the Spanish version of the Systematized Nomenclature of Medicine (SNOMED). It aims at describing the aspects of translating and the difficulties encountered in delivering a natural and consistent medical nomenclature. Bunge's three-layered model is referenced to analyze the sequence of symbolic concept representations. It further explains how a communicative translation based on a concept-to-concept approach was used to achieve the highest level of flawlessness and naturalness for the Spanish rendition of SNOMED. Translation procedures and techniques are described and exemplified. Both the computer-aided and human translation methods are portrayed. The scientific and translation team tasks are detailed, with focus on Newmark's four-level principle for the translation process, extended with a fifth further level relevant to the ontology to control the consistency of the typology of concepts. Finally the convenience for a common methodology to develop non-English versions of SNOMED is suggested.

INTRODUCTION

The need for a common terminology for healthcare has been extensively reviewed. [1,2] In order to fulfill the requirements of all audiences, controlled terminologies must be true concept representation systems. The development of the Spanish version of SNOMED [3] began in 1996, through an agreement with the College of American Pathologists (CAP). Following is the description of the theoretical and methodological background on which the process has been based.

BACKGROUND

From an epistemological perspective, symbolic concept representation may be viewed as a three-layered model.[4] The first layer is *physical*, and refers to the objects and events "out there", in the real world. The second layer is *conceptual*, and comprises individual human beings' mental representations of the artifacts in the physical layer. Finally, there is the *linguistic* (or symbolic) layer, made up of words, phrases, sentences and entire languages. These layers are bound to each other through special relationships: *designation*, by which linguistic categories identify their corresponding conceptual counterparts; *reference*, which binds concepts to their real-world subjects; and *denotation*, which is simply the union of designation and reference, and binds linguistic expressions to real-world entities.

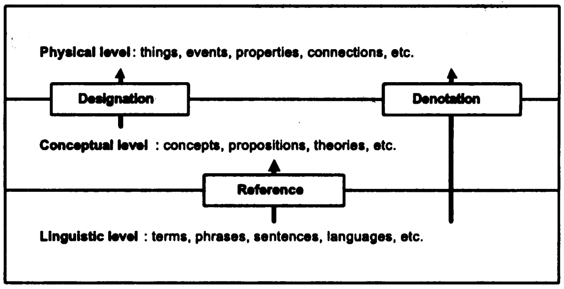
Physical level: things, events, properties, connections, etc.
Designation
Denotation
Conceptual level : concepts, propositions, theories, etc.
Reference
Linguistic level : terms, phrases, sentences, languages, etc.

Fig 1. Knowledge Representation Model (modified from Bunge).

Whereas the physical layer is easily understandable, the ontological status of the other two deserves a special comment. The conceptual layer is entirely personal, ideas which truly exist only in individual minds.[5] The linguistic layer is social, as languages exist for the ultimate purpose of communication. Symbols such as words will ensure proper communication only when performing their designational and denotational functions in a systematic manner: the same symbol must evoke

1067—5027/00/$5.00 © 2000 AMIA, Inc.

similar concepts in different people's minds and always refer to the same physical entities.

As for the translation field, two currents may be distinguished: semantic and communicative.[6] The former looks back on the author, tends to respect his linguistic style, and attempts to reproduce the pragmatic impact of the original text. It concentrates on the linguistic level. The communicative style is focused on the readership, by considering the message and the main ideas behind the text as the goal; it tends to be simple, clear and brief, and is always written in a natural and resourceful style. In his analysis of scientific language, Bunge [4] has pointed out that in the field of Science, as opposed to what occurs with the Arts, the main determinant in the idea-symbol pair is the idea. Whereas coding is a sensible solution, the fact that rubrics are still necessary to interpret codes requires that a special effort be made for the production of "idea-based" phrases. Since many coding systems do not include definitions or criteria for assignation, the meaning of rubrics must be self-evident and all-encompassing.

According to Newmark, [6] four levels are simultaneously brought into play during the translation process, all of which directly refer to the abovementioned three-layered model:

1) The *level of language* (or textual level), both the foundation and the source. On this level, certain conversions are intuitively made; the source language (SL) grammatical structures are rendered into their ready target language (TL) equivalents, and the lexical units are translated into their immediate contextually appropriate sense.

2) The *referential level*, the level of objects and events, real or imaginary, which has to be progressively visualized and built up, and is an essential part, first of the comprehension, then of the reproduction process. Any sentence/concept is intrinsically linked to its reference. Both the physical and the conceptual level are continuously at play, but translating occurs on the linguistic level, where the greatest possible referential and pragmatical correspondence with the words and sentences of the source language (SL) text is sought.

3) The *cohesive level*, which is more general and strictly linguistic, traces the train of thought and the various presuppositions of the original text. This level encompasses both comprehension and reproduction. It makes up a structure through connective words (conjunctions, repetitions, definite articles, general words, referential synonyms, punctuation marks). The cohesive level is a regulator in that it secures coherence and adjusts emphasis.

4) The *level of naturalness*, of common language appropriate to the writer or the speaker in a certain situation. The level of natural usage is grammatical as well as lexical. In all communicative translation, 'naturalness' is essential.

Finally, the revision procedure, which constitutes at least half of the complete process, is taken up. At this stage, the translated version is revised after some time, again having the textual, referential, cohesive, and natural levels in mind. It is at this point that most problems and inaccuracies are identified and solved.

TRANSLATION PROCEDURES

Creating a text that reads smoothly in the TL requires syntactical and structural changes to regroup the sequence of phrases and words, so as to preserve and accurately communicate the meaning and intent of the SL text. Several translation techniques have been used in the Spanish version of SNOMED, all of them framed within a context of respect for the original and focus on the terminology user.

Transference (or Borrowing): Due to the leading political, scientific and medical position of the English-speaking world, other languages often incorporate new terms by direct transference, that is simply the use of a word from the SL in the TL. Specific instances of this phenomenon are words such as *stress* or *rash*, frequently used in Spanish-speaking medical communities. Although functional equivalents such as *tensión* and *sarpullido* for the aforementioned do exist, the transferred forms were included in order to comply with the criteria of frequency of usage.

Naturalization: As a consequence of this frequency of usage, transferred terms from a SL are usually adapted to the TL, first to its normal pronunciation, and later to its normal morphology. Such is the case with the previous example of *stress*, which in Spanish has been metamorphosed to *estrés*. Considering that Transference and Naturalization are, in a way, sides of the same coin, both transferred and naturalized terms were included in the Spanish edition of SNOMED.

Synonymy: Synonyms do not occur in a one-to-one relation · across languages. In translating SNOMED, synonyms were often encountered without direct equivalents in Spanish. The opposite --synonyms to Spanish terms not found in English--was also a frequent instance. In every case, synonyms were added or removed according to normal usage in the Spanish language. No transfer or naturalization was exercised at this point, in order to preserve naturalness.

Through-translation: Common collocations, names of organizations, components of compounds, and other terms are often the object of literal transfer from a SL to a TL, in a process known as through-translation or loan translation. Normally this type of translation is only reserved for widely recognized terms. In the medical domain, the use of the acronym *HIV* is commonplace, although its formal Spanish equivalent *VIH* exists. Widely accepted English loan terms were included in our edition.

Shifts or transpositions: Because of differences in grammatical structure, translation often requires compromise involving both naturalness and grammaticality. Shifts [6] or transpositions [7] imply a change in syntax or structure to enhance meaning and reflect correct usage in another code. Several types of transpositions exist and were applied during the development of the Spanish edition of SNOMED:

The first type, the change from singular to plural, or the position of an adjective is automatic and offers the translator no choice.

A second type of shift is required when a SL grammatical structure does not exist in the TL. A typical instance of this is the use of the English gerund and present participle, often misused in Spanish. Gerunds were translated as nouns ("Operating an inguinal hernia"- Operación [surgery] de una hernia inguinal), infinitives (...operar [to operate] una hernia inguinal), and present participles as subordinate clauses ("Conditions causing complications in pregnancy"- condiciones que causan [that cause] complicaciones en el embarazo), adjectives ("mining technician"- técnico minero), or prepositional phrases ("dispatching and receiving clerk"- empleado de despacho y recepción de mercadería), according to their naturalness for each case.

The third type of shift occurs when literal translation is grammatically possible but may not accord with natural usage in the TL. Such is the case with *ventricular hypertrophy by EKG*, which could be translated as *hipertrofia ventricular por ECG* only at the expense of clearness: whereas an expert may readily grasp its meaning through context, the phrase is still professional shorthand lacking naturalness and is more correctly translated as *signos electrocardiográficos de hipertrofia ventricular (electrocardiographic signs of ventricular hypertrophy)* or *electrocardiograma con signos de hipertrofia ventricular (EKG with signs of ventricular hypertrophy)*.

Modulation: It refers to changing the conceptual approach to a unit of meaning. There are also several types of modulation, such as substituting an abstract term for a concrete one, or a noun for an adjective ("kindergarten teacher"- maestra jardinera).

Insertion/Omission: It refers to adding or deleting words or phrases in order to clarify meaning and ensure accurateness and naturalness during language transfer. Insertion involves the replacement of a virtual lexical gap by a structure that may allow for correct grammaticality in the TL. Example of such occurrences is *choking due to food in the larynx*, translated as *ahogamiento por presencia de alimento en la laringe (choking due to the presence of food in the larynx)*. Omission is often used to avoid redundancy in the TL ("social context condition"- condición social [social condition]), or simply because there exists a one-word option in the TL having the same meaning as the SL two-or-more-word term ("internal medicine specialist"- internista).

Translation methodology

SNOMED's monumental size and a clear need for productivity has refrained the authors from beginning with the phrase-by-phrase strategy common to full-text translation. During the first stage of the translation process, SNOMED International version 3.2, provided by the College of American Pathologists, was used as the initial work field. The full list of terms was used to generate a database with one unique identifier for every item and checksums of the Termcode and Enomen fields, in order to detect changes in future versions. The initial letter of all terms was turned into lowercase. The preliminary experience with off-the-shelf machine translation software was disappointing and the marginal utility was largely neutralized by the necessary corrections (mainly because of the verbless phrases and term categories which represent a great disadvantage for this kind of software). After performing an analysis of the word frequency, the software was trained with unrecognized medical words having at least 3 occurrences in the nomenclature, resulting in the addition of 8200 medical terms to the medical language module of the software (Globalink Power Translator, Globalink Inc.). All terms were pre-translated with the software, using a term-by-term translation. Morphemes (roots, prefixes, and suffixes in decreasing order of frequency) were later on replaced by their corresponding translation: --ectomy is always transcoded as –ectomía, acetyl as acetil, etc. As a result of this initial work, usual phrase tables were generated, taking the criterion of frequency as the leading parameter, and a pattern matching translation was done to avoid manual work on repetitive strings and to ensure consistency throughout the entire work. This was achieved with translation memory tools developed ad-hoc. After having followed these steps, about 20% of the terms had been adequately matched and required only manual confirmation; about 25% of the terms needed minor manual corrections, and the remaining terms had to be submitted to manual interactive translation processes.

The following steps uncovered a series of issues which required decisions and compromises: The terms generated at the pre-translation stage were initially evaluated by a group of both physicians and medicine students with knowledge of and practice in the English language, who applied a term-to-term approach. The inconsistency of that approach and the heterogeneity of problem-solving criteria yielded disappointing results. The problem needing readjustment at this stage was not the comprehension of the English original, but its rendering into the Spanish language. The analysis of a pre-release model of the French version of the Microglossary of Pathology (kindly provided by Dr. Roger Coté), the development of new software tools to make concept-by-concept translation easier within a Controlled Translation Environment (CTE), and the incorporation of a team of professional scientific translators to the project converged to set up a landmark which initiated a period of common criteria adoption and consensus as to the approach to conflict and problem solving. The in-house memory translation tool was used to provide for the re-use of words and phrases sharing the same conceptual context, and to allow for the development of terminological databases of both highly specialized terms and medical phraseology for the posterior development of auxiliary tables that permit an easier terminology retrieval. The final result was the use of a methodology based on iterative translation and revision cycles.

During the first stage of the task, the scientific team (physicians and students, who had received previous training and directions as to the common criteria adopted) worked on levels 1 and 2 --textual and referential levels--, and yielded a fully translated draft version, manually performed. A restricted access web site was used as a communication and discussion forum, created to solve daily difficulties related to the project. Periodic meetings were held, to unify criteria and to emphasize the importance of scientific rigor at the rubric translation process.

The above mentioned levels plus level 3 --cohesion--, were undertaken by a revision team of two professional scientific translators, who revised grammatical structures and checked for coherence and consistency. At this point, problem concepts were labeled, and terms were classified as real synonym, lexical variant, graphic variant, borrowing, naturalized term, foreign acronym, Spanish acronym, and adjectival form. Translation revision and problem solving were simultaneously carried out, as well as checks for denotation through consultation with domain experts. This procedure was also deemed necessary for coverage control and coverage adequacy. At this stage, the team also identified problems requiring solutions at later iterations, such as the criteria for considering synonyms in complex concepts (those made up by more than one single primitive concept), and flagged concepts that were considered locale-sensitive or variable within the Spanish-speaking regions (in general these were not scientific concepts, but ordinary, commonly used terms, such as 'eyeglasses').

Considering that the aforementioned stages can generate grammatically correct and semantically coherent translations, but lacking in naturalness and usage, a review at level 4 –naturalness—was conducted by three physicians with a background in medical terminology.

Newmark's four-level description of the translation process [6] was adapted for SNOMED with the addition of a fifth level for the ontology, to control the consistency of the typology of concepts within the hierarchy. The model was also modified to make way for a recycling, ever-continuing process. Ideally, a term should not represent more than one single concept, but a concept may be represented by various terms. Although this is not always the case, there is an instance of control by context ('dressing' represents different concepts within the frame of 'Food', 'Procedures', or 'Physical Agents'). Nevertheless, the consistency parameter is at stake when a particular term represents not only a concept, but also its father: "Pathology" as a medical specialty and "Surgical Pathology" both translate to "Patología", and "Clinical Pathology" represents a concept not existing in the Spanish culture and therefore lacking a one-to-one match translation.

The translation process is backed up by version control and updating processes paralleling the English counterpart, quality controls by means of both systematic (procedures to find inconsistencies, spelling mistakes, double spaces, punctuation errors, and so on) and manual methods (parallel translation by the same or different teams of physicians and translators), and backtranslation (translation into English of already translated terms, to check for reproducibility). The results of all these quality control procedures and the Spanish version validation will be the subject of a forthcoming paper.

DISCUSSION

The three categories of medical terminology described by Newmark are to be encountered in the *corpus* of terms:

Academic: mainly transferred terms from Latin or Greek. Eg: phlegmasia alba dolens, tetanus;

Professional: formal terms used by experts. Eg: epidemic parotiditis, vericella, scarlatina;

Popular: lay terms. Eg: mumps, chicken-pox, lockjaw.

Whereas no special complications arise with the first two groups, *popular* terms are highly local and the most suitable match may vary depending on its localization. The common dichotomies between American and British English (U.S. incubators and water baths are *insulated*, while in Britain they are *lagged*) are also to be found in Spanish and are observed in all levels of language (Eg. Peninsular Spanish -spoken in Spain- favors *gafas* over the Americanized form *anteojos* for the English term *eyeglasses*).

One of the main problems in the field of terminology is the multitude of designations referring to the same concept. A term-to-term (literal) translation of SNOMED thus poses epistemological difficulties. For example, in its English version, SNOMED provides only one term for "tract" (neurological pathway), whereas at least two terms exist and are used in Spanish: *tracto* and *vía*. The opposite also occurs, as the Spanish word "sueño" simultaneously stands for the English terms *sleep* and *dream*. Consequently, common structures can be made unnatural or incorrect by term-to-term transcoding. One may even fall within an instance of what is called *translationese,* that is, a literal translation that does not produce the appropriate sense of the original.

The use of CAT proved to be time saving in the initial pre-translation phase. The use of custom memory translation tools during the manual interactive stage provided for re-use of previous translations of repetitive patterns, and enabled consistency, bookmarking of pending issues and version control during the team's daily work.

CONCLUSION

As previously referred, naturalness and allowance for regional differences have been the authors′ main goals while the convenience of a common methodology for the development of non-English versions of SNOMED is considered of paramount importance.

The fact that syntax and semantics assist each other in achieving genuine concept representations should be born in mind, and an adequate equilibrium must be sought. Machine-readability and expressiveness must be simultaneously controlled, so that none of them may cancel out the benefits of the other.

ACKNOWLEDGEMENTS

Special thanks to Dr. Roger Côté, Dr. Kent Spackman and Karen Kudla, from the College of American Pathologists for their guidance and continuous support during the development of the Nomenclatura Sistematizada de Medicina.

REFERENCES

1. Chute CG, Cohn SP, Campbell JR. A framework for comprehensive terminology systems in the United States: development guidelines, criteria for selection, and public policy implications. J Am Med Inform Assoc. 1998.

2. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med. 1998;5:503-10.

3. Cote RA, Rothwell DJ, Beckette R, Palotay J (eds.) SNOMED International. College of American Pathologists, Chicago, 1993.

4. Bunge M. Sense and Reference. Boston: Reidel; 1974.

5. Bunge M. La Investigación Científica. Barcelona: Ariel; 1981.

6. Newmark, P. A textbook of translation. UK: Prentice Hall; 1988

7. Catford JC. A Linguistic Theory of Translation. Oxfod: OUP; 1915.

8. Vinay JP and Darbelnet J. Stylistique comparée du francais et de l'anglais. Paris : Didier; 1965.