The Development of a MeSH-based Biomedical Termbase at Hogeschool Gent

Joost Buysschaert
Hogeschool Gent
Departement Vertaalkunde, Groot-Brittannielaan 45, BE 9000 Gent, Belgium
joost.buysschaert@hogent.be

Abstract

This paper reports on an ongoing long-term project to build an English-and-Dutch termbase using the MeSH terms (Medical Subject Headings) as input. Although from the start NLP applications had been envisaged, the database has mainly been built according to the traditional principles of terminology management for human translation. With important parts of the project now nearing completion, the question arises whether and how the material could be made available in a traditional dictionary format as well as in a format that can be used in language technology applications. It is argued that the traditional detailed working method used, based on explicit evidence and recording a wealth of information on synonyms, variants, usage, and reliability, can also be profitable to NLP applications. It is unlikely, however, that a single format can be found to make the data available for all possible purposes. Rather, the current database will have to act as a common repository from which various extractions can be made, through conversion, for different applications. To facilitate conversions, it would be expedient for future projects to work towards a uniform standard from the start. It is speculated that TermBase eXchange may be the most promising emerging standard at the moment.

1. Existing Medical Glossaries with Dutch

Existing English-and-Dutch medical dictionaries are limited in scope, definitely when one confronts them with the vast wealth of medical terms found in thesauri like the Medical Subject Headings (MeSH, http://www.nlm.nih.gov/mesh/).

Among the bilingual sources we may mention two dictionaries in paper form, Kerkhof (2003) and Mostert (2002), both of them slim volumes uniting both language directions. Online lists like Taalvinder (http://www.ochrid.dds.nl/medici.htm) and Woordenboek Ziekenhuistermen (via http://www.ziekenhuis.nl) are very deserving but also limited in their number of entries as well as in the information provided.

An important multilingual list in which Dutch is also represented is the Multilingual glossary of technical and popular medical terms in nine European languages at http://allserv.rug.ac.be/~rvdstich/eugloss/welcome.html, developed at our college in cooperation with the Heymans Instituut voor Farmacologie. Yet, here too, a term like orthopaedic will obviously be found but a more technical item like orthomolecular will be absent.

2. A Bilingual Termbase Project

An obvious and undoubtedly rewarding way to increase the scope of a medical glossary is to take input from a detailed medical thesaurus like the MeSH. This idea was suggested to us by R. Vander Stichelen of the Heymans Instituut in 1987. His first suggestion was to provide Dutch equivalents for the MeSH subject headings so that, for example, the Dutch headings could be used to search the Index Medicus; or so that the Dutch as well as English headings could be used for indexing medical publications co-sponsored by his Institute. (On the topic of Cross-Language Information Retrieval see also Peter Schauble et al. and references there.)

By suggesting the idea to our School of Translation Studies (Hogeschool Gent), however, he had awakened another interest, viz. the development of a full-scale medical dictionary. This was to take the project beyond such applications like indexing, document retrieval and natural language processing (NLP) in general, to also make it useful for human translators dealing with a variety of medical texts. As will be indicated below, the interests of the NLP-specialists and traditional translators/terminologists do not always coincide but the confrontation of two parties can be wholesome.

Lack of adequate funding for the project meant that it was cut up in a large number of thesis subjects (over 130 to date). Students are each assigned a subchapter from the MeSH, so that they can concentrate on a specialist subject area. They liaise with one or more specialists of that subject area, preferably staff in the University Hospital, and they fill in (very) detailed records on each concept studied. Research involves primary texts as well as reference works and informants.

Work has been slow moving but thorough. The MeSH chapter on diseases has now been covered for 90% and the chapter on medical procedures and techniques is also nearing completion. Large sections of other chapters have also been dealt with but some need revision. In the last couple of years, work has started on adding French equivalents using the same detailed record, but here too progress is slow.

There are now plans to publish specific parts of the Dutch-and-English material, possibly on CD-ROM or on a protected website, and the project leaders are faced with a choice between a more traditional dictionary format that would undoubtedly be hailed by the human medical translators, or a machine readable format that would be welcomed by human language technologists - or both.

There can be no doubt that the way in which the material has been developed has been more slanted towards the traditional dictionary approach; yet it is believed to be sufficiently structured to allow conversion to an NLP-type glossary.

3. NLP versus traditional terminology

As suggested earlier, cross-fertilization of terminology work for NLP on the one hand and traditional terminography on the other stands to benefit both parties. NLP adepts are typically interested in one-to-one term lists in machine readable form; whereas traditionalist terminologists tend to favour detailed records for each concept.

One-to-one conversions of the MeSH-thesaurus have been created for several languages (cf. http://www.nlm.nih.gov/research/umls/sources_by_categories.html). Some can be consulted via HONselect.
that traditional terminologists are apt to do, can also be relevant for the machine readable derivations.

3.3. The issue of usage

Dutch medical language, more so than English, has variants that can be termed either "technical" or "popular". The former terms (nausea) would be favoured in the scholarly literature, the latter (misselijkheid) would be used in the communication with patients and are therefore also eligible for use in patient information leaflets. In fact, the need for popular equivalents that could be used to make information leaflets more readable prompted the European Commission to sponsor the Multilingual glossary referred to above. (In the US, patient information does not enjoy the same status, mainly because of legal concerns; cf. Vander Stichele, 2004, 13ff.). Again, a translator would want to know both types of terms because s/he may find him/herself asked to translate texts of various types. But s/he would want the terms to be labelled, so that they are recognizable as being either more scholarly or more popular. Simple bilingual lists may not carry such indications. Yet the information is relevant and ways should be found to store it even in lists used for NLP purposes.

Level of formality or register is only one type of usage. Another type is regional usage. Occasionally, a Dutch medical term is only known, or favoured, in Belgium; the abbreviation MUG, for example, is a commonly understood name of a particular type of ambulance service in Belgium but would be unknown in Holland. The same holds true for British versus American English with institutional names like NHS or Medicare. But also spelling differences may be a question of regional usage (anaemia versus anemia); any system, whether meant for human translators or for machine purposes, would need to make both variants available but should also label them appropriately. The UMLS's Specialist Lexicon, clearly geared towards NLP, does list alternative spellings but does not label them in detail – cf. an example in Browne et al. 2000, 1:

```
{base=anaesthetic
 spelling_variant=anesthetic
 entry=E0008769
 cat=noun
 variants=reg
}
```

(The Specialist Lexicon can be downloaded at http://lexsrv3.nlm.nih.gov/SPECIALIST/Projects/lexicon/2006/release/LEXICON/LEXICON.)

A third type of usage information (which overlaps with the category of reliability below) is that of topicality. Some terms tend to become obsolete for a variety of reasons. The rapid evolution in genetics, for example, has meant that a number of vaguely named "factors" have at a later stage received more specific names. Sometimes also there are attempts at new classifications with new names, as has been the case with the vocabulary of epilepsy. The obsolete medical terms, however, have a strong tendency to survive anyway and to crop up regularly in texts. This means that at least the human translator needs to be aware of their existence but also of their status. Yet indexing systems, document retrieval systems, or machine
translation systems can equally well gain by the presence of obsolete terms in their lists, though there, too, it would arguably be interesting if those terms could be marked as special.

3.4. The issue of relative reliability
Unlabelled lists do not explain whether proposed translations are common terms or not. Yet this is crucial information. Traditional terminology work has often recognized this by adding reliability codes. All our projects in Ghent use the following codes (among some other ones):

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>nor</td>
<td>This term was found in a normative source</td>
</tr>
<tr>
<td>leg</td>
<td>This term is the legally used term</td>
</tr>
<tr>
<td>pri</td>
<td>This term has been found used in only one primary source</td>
</tr>
<tr>
<td>2pri</td>
<td>This term has been found used in only two primary sources</td>
</tr>
<tr>
<td>3pri</td>
<td>This term has been found used in at least three primary sources</td>
</tr>
<tr>
<td>sec</td>
<td>This term has been found used in only one secondary source</td>
</tr>
<tr>
<td>2sec</td>
<td>This term has been found used in only two secondary sources</td>
</tr>
<tr>
<td>3sec</td>
<td>This term has been found used in at least three secondary sources</td>
</tr>
<tr>
<td>neo</td>
<td>This term is a neologism created by the terminologist (and preferably sanctioned by a domain specialist).</td>
</tr>
</tbody>
</table>

Table 1. Reliability codes in GenTerm.

A term that comes with the code "nor 3pri 3sec" would therefore be a very reliable term; one with just "sec" would be more doubtful and "neo" serves as a firm warning that the term is a proposal only.

"Primary" sources are defined as texts written by and for domain specialists (in our medical project: doctors writing for doctors or at least for trainee medical staff). "Secondary" sources are either reference works (especially dictionaries, which tend to be compilations with the editor not always being a specialist in every subdomain) or texts written for a lay audience (a website for sufferers of arthritis).

Some terms are typical "dictionary terms" and appear not to be used in practice; other, usually very specialist terms, are well-represented in primary documents but have somehow escaped the attention of dictionary compilers. The choice of "3" as a threshold (in the codes 3pri and 3sec) is admittedly debatable and dates back to the days when only paper sources were available. The presence of a technical term in at least three different sources was then deemed to be sufficient evidence of good reliability. In the days of the internet, it has become much easier to find 3 google hits even of a not so common term. Yet it is not clear what an alternative threshold could be. Much depends on the language and the specialist domain. Five hits for a Dutch term in a not very commonly practised specialism is a lot. Only five hits for an English term in the context of a widely practised specialism makes one go and look for a better synonym.

The relevance of reliability labels is considerable in translation work. A drawback of using MeSH as input of our termbase is that its tree structure contains a number of artificial terms entered to fill the gaps in the system. These include the so called "NON MeSH" terms (fortunately labelled as such) like neoplasms by site. In our project, this particular term has received the code "nor pri", which indicates to the reader that although this term is in the MeSH tree (= nor), it occurs only once in a primary source and was not found recorded in the reference works.

Another example is an extensive list of artificial names ending in "surgical procedures" (in the E04-chapter of MeSH), meant to refer to the actual performance of surgery, and not to the branch of medicine (which is a different chapter in MeSH). In actual practice, sources would say that obstetric surgery was performed, not that obstetric surgical procedures were performed. The latter is once again a creation for the sake of a well-designed concept tree but not an actually used term. The reliability codes fortunately help to make this clear.

The codes can also help the translator decide on a synonym: ambulatory surgery, with "nor 3pri 3sec" will be preferred to day-case surgery, which is rated as "3pri sec" only. But what is true of human translation, is obviously also true of, for example, machine translation. If Systran's Dictionary Manager had both ambulatory surgery and day-case surgery in its English-to-Dutch list, it would be able to recognize both in a source text; but its Dutch-to-English conversion should be coded in such a way that the former is presented as the preferred option.

Some terms "do occur" but should be warned against because they are not very rare use compared with established alternatives. Some other terms need to be deprecated because they are very obsolete, carry undesirable overtones, or violate established spelling rules (like Dutch arthrose, a relatively frequent misspelling of arthroscopy). In our project, we relegate such terms to a field "ExtraSyn". Again, it means that they are retrievable, but that suitable warning is given.

3.5. The issue of standards
Whereas traditional terminology excels in the areas of giving evidence and providing information on synonyms, usage and reliability (as well as other categories, like grammatical information, information on collocations and even pronunciation - all of which are present in our termbase), it has a very poor record when it comes to observing ICT-related standards. NLP-related terminology work, on the contrary, has had to ask itself from the start what formal criteria it had to fulfil for its word lists to be compatible with a number of computerized tasks.

The only standards issues that have half-affected traditional termbase builders, are exchange formats. Yet in practice, few terminologists were originally interested in a proposed standard like Martif (Machine Readable Terminology Exchange Standard), simply because they regarded their termbases as their own assets and felt no need to ever exchange them. In recent years, however, translators have increasingly been asked to contribute to large translation projects, and sharing terminological databases is often a must in these cases. The newer TBX standard (TermBase eXchange, an open-source XML-based standard, cf. http://www.lisa.org/standards/tbx/)
stands a better chance of fulfilling these needs. TBX makes it possible to convert material from one terminology management system (TMS) to another. A TBX-file is a tagged file like the fragment below:

```
<languageGrp>
  <language type="English" lang="EN-US" />
  <termGrp>
    <term>patient identifier</term>
  </termGrp>
  <descrGrp>
    <descr type="PartOfSpeech">noun</descr>
  </descrGrp>
</languageGrp>
```

Table 2 Fragment from a Medtronic record in TBX (taken from http://www.lisa.org/standards/tbx/samples/#medtronic)

The terminological record used at Hogeschool Gent (nicknamed the GenTerm record) is modelled on the import format of the "old" Multiterm, the best-known TMS among human translators. It was chosen in the hope that Multiterm would in its later developments be standard-conscious. The old Multiterm input format has a "flat" structure, as the following opening fragment of a GenTerm-record illustrates.

```
**<Vakgebied>neurologie
<BSO:437.50
<UDC:616.8
<Project>"MeSH E8 CiV 4" - ^UPDATE MeSH E5 JY 6^
<Werkcode>C10.228.140.163.520
<Update Werkcode>C10.228.140.163.474.450 [UPDATE JY]
<Begrip>zeldzame erfelijke metabole aandoening van de hersenen bij zuigelingen (jongetjes) en die gekenmerkt wordt door een stoornis in de koperopneming</Begrip>
<Intemat>
<NI-term>kroeshaarsyndroom
<Equiv>
<En-term>kinky hair syndrome
<Equiv>
<Beeld>

<English>kinky hair syndrome [UPDATE JY]
<Trefwoord>kinky [UPDATE JY]
<Betroub>nor pri 3sec [UPDATE JY]
<Woordsrst>sub [UPDATE JY]
<Genus>
 ...
```

Figure 2. Fragment of a GenTerm-record.

The "new" Multiterm, originally named Multiterm iX, introduced a more structured XML architecture. A conversion module allows the transition from the old to the new format, which looks very much like a TBX-file. Yet, closer inspection has shown that there are obstacles in the way of converting iX to TBX (cf. Reineke, 2005). On the other hand, recent examples on the LISA website (http://www.lisa.org/standards/tbx/samples/) have demonstrated that conversion to TBX from a variety of terminological sources (varying from an XML-type lexicon from Medtronics to even a simple excel spreadsheet) is possible, giving hope that the trick can also be performed on Multiterm-data or indeed on the original GenTerm records. As traditional terminologists are becoming more standard conscious, it is legitimate to hope that TBX will at some stage provide the key to opening up their archives to NLP-minded colleagues. Yet at the NLP end, there have also been discussions about a common standard for lexical description; the emerging standard here could be LMF, Lexical Markup Framework (future ISO 24613, cf. Francopoulo et al. 2006.) Whether it is possible to link TBX with NLP remains to be seen. An XML specification of NLP is still in preparation.

3.5 The issue of variable grammatical forms

Another area in which human language technologists have been better than traditional TMS users, is that of recording alternative forms of terms: plurals alongside singular forms, for example, or inflected forms of verbs. Traditional term records often ignore these forms (GenTerm does not: it records them in a field called "Flexie"). Traditonalists have often regarded this information as obvious, relying as they do on the human user's language knowledge and therefore only recording exceptional forms. When using their TMS in conjunction with a translation memory, they rely on the fuzzy terminology recognition facility to spot the plural term even if only the singular term is in their list (with the risk that the fuzzy recognition will confuse palpitation with palpitation.)

Another aspect that at least some NLP lexicons have been better at is the recording of the syntactic potential of lexical items, for example the possible arguments of verbs.

4. Conclusions and recommendations

Research into medical terminology has so far been either geared towards preparing medical dictionaries for human use or towards readying machine readable lists for NLP purposes. This paper argues that it should be possible for the two to meet up. The point is illustrated with examples from our English-and-Dutch MeSH-based termbase. The following general conclusions may be drawn from the discussion:

1. It may not be realistic to try and design one termbase, in one uniform format, that will directly be able to serve the human user as well as various NLP users at the same time. A more realistic alternative is that one common core database is drawn up storing all the relevant information, from which various extractions and conversions can be made to serve NLP needs like indexing, multilingual document retrieval, automatic translation etc., and from which also a traditional dictionary can be derived.

2. The core database should be detailed from the start. It should:
- contain evidence in the form of contexts and references to sources
- list true synonyms as well as near synonyms, deprecated variants (including obsolete terms), alternative spellings, alternative grammatical forms
- label variants so that their usage status is made clear (register, regional usage etc.)
- (ideally) give information on the syntactic potential and pronunciation of lexical items
- give an indication of the relative reliability of the term, clearly distinguishing neologisms and rare terms from common terms.

(3) The core database should from the start adopt a design that is compatible with an agreed norm. The present relative enthusiasm for TBX, and the fact that conversion experiments from various existing formats to this norm look promising, make it a good choice. Yet it is unclear whether a bridge can be built between TBX and another emerging standard, LMF.

A slightly adapted version of this text has been submitted for publication in the journal Equivalences.

References


HONSelect. http://debussy.hon.ch/cgi-bin/HONselect?browse=A01.047#MeSH.


TBX - TermBase eXchange. http://www.lisa.org/standards/tbx/

The specialist lexicon.

