Interchanging Lexical Information for a Multilingual Dictionary

RH Baud[1], M Nyström[2], L Borin[3], R Evans[4], S Schulz[5], P Zweigenbaum[6].

[1] Service of Medical Informatics, University Hospitals of Geneva, Switzerland
[2] Department of Biomedical Engineering / Medical Informatics, Linköping University, Sweden
[3] NLP Section, Department of Swedish, Göteborg University, Sweden
[4] Information Technology Research Institute, University of Brighton, UK
[5] Department of Medical Informatics, Freiburg University Hospital, Germany
[6] Assistance Publique – Paris Hospitals, STIM; INSERM, U729; INALCO, TIM, Paris, France

Objective*: To facilitate the interchange of lexical information for multiple languages in the medical domain. To pave the way for the emergence of a generally available truly multilingual electronic dictionary in the medical domain.* ***Methods****: An interchange format has to be neutral relative to the target languages. It has to be consistent with current needs of lexicon authors, present and future. An active interaction between six potential authors aimed to determine a common denominator striking the right balance between richness of content and ease of use for lexicon providers.* ***Results****: A simple list of relevant attributes has been established and published. The format has the potential for collecting relevant parts of a future multilingual dictionary. An XML version is available.* ***Conclusion****: This effort makes feasible the exchange of lexical information between research groups. Interchange files are made available in a public repository. This procedure opens the door to a true multilingual dictionary, in the awareness that the exchange of lexical information is (only) a necessary first step, before structuring the corresponding entries in different languages.*

Present context

There is currently no large electronic dictionary in the medical domain, with a true multilingual dimension, say up to 10 languages with relevant coverage and substantial lexical information. The multilingual dimension means at least that the corresponding entries in different languages are connected, which is a difficult and never finished task. However, the situation is not so bad in a specific subdomain like gross anatomy, where international consensus has been reached [1] and where modeling has been successfully achieved [2].

The contribution of UMLS has to be mentioned here. The NLM has continuously made efforts for the integration of foreign languages in the metathesaurus. For each concept, there may be a list of representative terms in several languages (though with incomplete coverage), one of them being flagged in each language as the preferred term. However, lexical information such as part-of-speech is missing and character representation was insufficient until recently, even for most Latin-based orthographies, let alone other writing systems. The UMLS editors presently consider these problems, but numerous languages are outside of their main priorities.

Outside of the medical domain, there have been numerous authors working on multilingual dictionary development. But most of them present poor or inconsistent coverage of the medical domain because it obviously was not their target [3, 4]. The WordNet system, despite not being especially developed for the medical domain, has a good coverage in English [5]. In addition, its EuroWordNet versions tend towards a multilingual system, but they have diverse levels of coverage of medicine [6].

Where a medical domain dictionary has been developed for multiple languages, it lacks convenient coverage or has been developed as a demonstrative prototype [7, 8]. Whatever the situation today, at any time when building a multilingual dictionary it is a good idea to look for existing bilingual mappings, even with partial coverage of the domain.

Such issues are being addressed within the framework of the European Network of Excellence "Semantic Interoperability and Data Mining in Biomedicine". A multinational team of researchers from the fields of linguistics, computational linguistics and medical informatics, including the authors, gathered in a serie of meetings with the goal of building a European multilingual dictionary.

The need for an electronic multilingual dictionary

As a first argument, electronic patient records are most often stored in the language of the patient. If in North America this means at least 3 languages (English, Spanish and French), in the European Union this means at least 20 languages, and many

more in the rest of the world. Increasingly, new tools for automatic treatment of electronic patient records are emerging, like indexing, literature search, patient encoding, etc. In each situation there is a need for: 1) a medical dictionary in local language; 2) access to a corresponding English dictionary for literature search; 3) access to other languages for bilateral exchanges. The last two points are basically multilingual aspects, not to be solved by a monolingual dictionary.

Second, many workers in the discipline of Natural Language Processing are active in their own language and have already built monolingual medical dictionaries, which may be of considerable size. These researchers are ready to interchange their data and their experience. It is usually agreed that the methodology for building a medical dictionary is language independent and mainly content-driven, even if language idiosyncrasies are quite common and should be resolved. Collecting the information in a specific language is largely facilitated by the presence of another dictionary in a related language or in English, which is the language of the literature. Latin and Greek roots largely inspired the medical vocabulary. Quite often corresponding words in different languages only differ in their ending, the choice of characters and the use of accents. A medical expert may often recognize a word in a language where (s)he is not fluent at all.

A third economical reason has to be invoked now: limited manpower resources for dictionary development. It is well known that dictionary construction is very labor-intensive. This aspect becomes a major obstacle when considering “minor” languages, spoken by less than 20 million people: Slovenian, Catalan, Norwegian, etc. There is presently little opportunity to find an electronic medical dictionary in these languages. The only reason is the lack of adequate resource: there is absolutely no technical reason.

At fourth should be considered the potential benefit of such an investment. The advent of a true multilingual dictionary will offer to the community a new terrain for exchange. The exchange of scientific literature has been a success for many years, in particular thanks to the efforts of Medline. But the exchange of medical records about patients or at least related information to take into account the confidentiality constraints, is rather limited at an international level. There are at least two fields with a clear interest to break this barrier: 1) the fight against epidemics, where the example of SARS is not so far in the past; 2) the search for orphan diseases necessitates the collection of multiple cases not to be found in a single country or region.

Last but not least, Natural Language Generation (NLG) is an emerging discipline in the medical domain [9]. There is no doubt that the existence of a multilingual dictionary is a prerequisite for widespread NLG tools.

Constraints for interchange

The basic rule for successful interchange is that each partner finds some benefits and advantages.

The cost for the author of a dictionary in a single language is the transformation from the actual format to the interchange format. Therefore the interchange format should be straightforward to use and easily accessible.

The benefit is the possible availability of interchange files in the same language, seen as enrichment of his or her dictionary. Feedback from other users may act as a validation process and an improvement. The link to other languages (see below) is another expected benefit.

In addition, there is a more pragmatic constraint to consider: the format of interchange should be kept as simple as possible. The reason is that the existing candidate dictionaries are developed at considerably different granularity of information. Any author may be unwilling to work with a complicated format, where (s)he will be unable to feed many fields. On the contrary (s)he will be pleased to make a rich contribution when being able to feed all the defined fields.

What are the attributes for interchange?

This section is about the introduction of the selected interchange attributes and the nature of their content. It does not pretend to be exhaustive; only the official documentation meets such an expectation [10].

The main attribute is the lemma or basic form of a dictionary entry. Due to the existence of morphological variants of many words in most languages, it is necessary to decide about an arbitrary representative form. For each language, there is normally an agreed convention by lexicon editors and by linguists about which form should be used to represent a word of a particular part of speech in a dictionary. Using this basic or citation form and language-specific morphological rules, it should be possible to generate any other form.

The type of the dictionary entry is an important attribute. Possible values are the followings: single words, parts of words, terms and compound entries. Single words are lemma without blank character;

parts of words are for instance roots, prefixes and endings, which enter into the composition of numerous words in the medical domain; terms are multiword expressions, where the meaning of the expression cannot be simply deduced from the senses of its constituent words[1]; compound entries are made by composition of parts of words according to language-specific rules[2].

For each compound entry, there is a dedicated attribute, where the decomposition is presented. Each part of the word is separated from the others by a double underline character. Any intermediate character like dash would appear between the two underline characters. The value of knowing the parts of a compound term is the fact that they carry an explicit information about the term content.

The part of speech is another attribute of importance. The problem with this attribute is to set up a convention, which is valid with most languages[3]. Hopefully, the MulText representation [11] acts as a standard de facto to this respect, has been adapted to numerous languages and exists with a detailed documentation.

The next attribute is for the creation of a unique identifier attribute, valid now and forever. The author of an interchange file should be able to generate it immediately without referring to a central source. This is theoretically impossible, but practically each author is required to define a unique 3-letter identifier for his group or his institution and to check manually its uniqueness. There are not so many expected contributors and making a list of actual contributors is not so difficult. Then the unique identifier of any dictionary entry is obtained by juxtaposition of the author short name and a unique identifier within the set of all interchange files provided by this author.

Another attribute recognizes the quality of the present entry, if it is a canonical form in the language or not. Mistyped words, jargon, orthographic errors are the possible values. This entry should normally have a pointer to a referent entry, which is the canonical form itself[4]. Such a reference attribute is indeed part of the interchange under the form of the lemma of another entry. This is not a satisfactory solution but again it is a pragmatic one. The unique identifier of another entry possibly in another interchange file prepared by another author is generally not available. Using the lemma is not perfect because the same lemma may have multiple meanings and there is no way to find which one is the good one at the level of the interchange file. Such soft references will necessitate further validation at appropriate time.

Other attributes are present like an additional field for an inflected form, which does not follow any rule of the language, together with a MulText corresponding additional argument. Another one makes it possible to specify the inflection class of the word, interpreted in a language-specific way.

Attributes are also defined for comments and examples, for documentary purpose only. A catchall entry is at disposal of the author, but any usage of its content is not guaranteed.

Interchange format

The interchange format is officially XML, but an alternative solution exists in the form of a pipe-delimited record. A simple utility, knowing the list and order of attributes can convert one form to the other at any time.

The 7-bit ASCII representation of characters is too limited for European languages, and its 8-bit "Latin" extensions (among which Latin-1 or ISO-8859-1, used in Western Europe) each cover regionally limited requirements. The extensive Universal Character Set (ISO 10646) was therefore chosen, using its UTF-8 encoding which conveniently falls back to ASCII for the basic Latin characters used, e.g., in English..

The exchange procedure has three steps: any author would first register[5] and obtain a unique author identifier of 3 letters. Then he decides to work on a given language (one language per file) and he prepares the conversion from his own sources. Thirdly, he makes the file available in a central public repository.

Then the interchange procedure is ended, but in fact additional steps are feasible and are waiting for further initiatives. They are the following:

[1] The *femoral artery* is not the artery of the femur, but the artery of the whole leg.

[2] *Acidemia* is recognized as *acid__emia* meaning excess of acid in blood.

[3] For present time, we limit ourselves to a number of Western European languages, simply because this is where our competence lies.

[4] Typically the entry *flegmon* is incorrect and should point to the canonical form *phlegmon*.

[5] Actually the SemanticMining Network of Excellence conducts the work. Later a webpage for federating the interchanges will be developed with a moderator.

- Merging two files of the same language with deletion of duplicates;
- Validation of a file with updates and production of a new file;

In all cases, the old files are not deleted unless they contain trivial errors or wrong duplications. A short electronic notice giving administrative information about the content will accompany each file.

Building the multilingual dictionary

The advantage of the interchange procedure is its simplicity. Multiple providers from several languages are expected to participate after an initial roundup by the authors of this paper and adjustment of the rules. The result will be a number of interchange files acting as raw material for a future multilingual dictionary. It is now time to examine what further steps are necessary.

Further steps are dependent on individual initiatives, which may be coordinated or not. There is no preconception to nominate a coordinator and we prefer this space of freedom to a guided solution, which is at risk of being unsatisfactory for many people and many languages.

Imagine first the initiative of an expert of a given language, who will merge all files for this language. The main problem to resolve is evidently the detection of duplicates, probably based on identical lemma and part-of-speech argument. In addition, automatic or manual validation is welcome. Such a task can be performed separately for each language.

But the main phase towards a true multilingual dictionary is the detection of "corresponding" words between different languages[6]. Let us define what these words are. Considering an object of the domain under scrutiny, there is generally in each language at least one dictionary entry as a noun, one dictionary entry as an adjective and possibly one dictionary entry as a root or part of word. A typical triple is *liver/hepatic/hepato*, which are separate entries in the final version of the dictionary. Of course at any step one entry may be absent for a given language. In the case of two languages, corresponding words are two nouns, adjectives or roots being the expression of the very same object. With the pair of languages English and French, corresponding words are in this example: *liver/foie*, *hepatic/hépatique* and *hepato/hépato*.

[6] Additional syntactic information may be useful for such a task, but this point is left open for today.

Automatic search of candidates for grouping has been explored. The idea of taking advantage of the Latin and Greek roots valid in several languages has been developed by one of the authors [12] and is applied to the present situation. The underlying idea is to use already existing translations at a subword level in order to support the acquisition of translations at a term level. Another author is associated with a work based on morphological analysis going in the same direction [13]. However, attention must be paid to erroneous mappings (so called false cognates) when constructing translation groups automatically [14].

The grouping of the corresponding entries is the essence of a multilingual dictionary: this operation transforms a set of monolingual lexicons into a multilingual dictionary. Before this operation, the dictionary entries are independent; afterwards, they are organized as clusters of entries. The first grouping of corresponding entries may be followed by other links between all the groups related to the same object expressed by a noun, an adjective or a root (or ending). In practice, it is not uncommon to find true synonyms; this means we may have two or more groups of nouns, adjectives or even roots[7] for the same object.

This two-step grouping is fundamental and represents the expected added value of a multilingual dictionary. It introduces the basic links between languages, which any multilingual application is seeking for. Additional links to existing nomenclatures or ontologies are then possible, but this is another development not to be considered here. It should be mentioned that in the present situation, any added semantic link is valid for all languages: it is no more necessary to set up such links in all languages. In other words, the additional burden of producing the multilingual dictionary is potentially compensated by this heavy benefit, at least in the long term.

Discussion

The initial result is clearly the set up of an agreement on an interchange format. Six different linguists representing 8 languages[8], active in the medical domain, came to a practical solution and are

[7] For example *spondylo-* and *vertebro-* are synonyms.

[8] The languages are: English, Swedish, German, French, Latin, Spanish, Portuguese and Italian. The initial expected coverage may largely vary from one to the other. This is the current status before discovering new partners willing to work on their language.

currently implementing it. The initial, fairly mechanical step, the transfer of raw lexical data, will take place between the involved research groups during the year of publication of this paper. However, the question open for discussion is about the reality of the follow-up steps, when building the true multilingual dictionary. Links with ongoing lexicon standardization efforts such as the Lexical Markup Framework of ISO/TC37/SC4 [15] will also be sought.

The authors are firmly convinced that no authority can be successfully exercised for the federation of languages. I may be possible tomorrow under the responsibility of an international leadership, but it is not an open statement today. For this reason, a public repository of raw lexical data, with a sensible but not directive action of a moderator, is a federative temporary solution, which will be compatible with both future individual and group initiatives. The existence of a corpus of raw data is an invitation to any scientist or SME to develop the multilingual aspects. Multiple initiatives are expected; certainly a few of them will be recognized by the scientific community for their value and their services. At this moment, and not before, it would be time to put in place a more structured entity or to require the services of an existing one (like the NLM), in order to collect adequate resources and to insure the continuity of services.

Conclusion

The need to develop a multilingual dictionary is evident today and five main arguments in this direction have been presented. There is no huge technical obstacle. Only the lack of adequate manpower resources explains the slowness of current progress.

This paper has presented an ongoing initiative for the collection of raw lexical data in the medical domain. An agreement for an interchange format has been set up. Dictionary construction for 8 languages is underway.

This initial collection is seen as a trigger for further actions, and especially as a starting point for the construction of a true multilingual dictionary, having links between corresponding words in different languages. This unauthoritative approach has the potential of preparing the ground for a federative solution developed on the long term.

Acknowledgement

The development and dissemination of the Interchange Format as described in this paper has been supported by the Network of Excellence Semantic Interoperability and Data Mining in Biomedicine, Working Group 20, project funded by the European Union.

References

[1] Federative Committee on Anatomical Terminology. Terminologia Anatomica: International Anatomical Terminology. Thieme Ed. 1998.

[2] Rosse C, Mejino JL. A reference Ontology for Biomedical Informatics: the Fundational Model of Anatomy. J Biomed Inform 2003 ;36(6) :478-500.

[3] JW Brent. Practical Issues and Problems in Building a Multilingual Lexicon. www.csse.monash.edu.au/~jwb/ws2002_paper.html

[4] Basic Multilingual Lexicon – MEMODATA www.elda.org/catalogue/en/text/doc/lexmult.html

[5] Bodenreider O, Burgun A, Mitchell JA. Evaluation of WordNet as a source of lay knowledge for molecular biology and genetic diseases: a feasibility study. Stud Health Technol Inform. 2003;95:379-84.

[6] Vossen, Piek (ed.) 1998. EuroWordNet: a multilingual database with lexical semantic networks. Dordrecht : Kluwer. ("Reprinted from Computers and the humanities, volume 32, nos. 2-3, 1998")

[7] Eurodicautom http://www.translatum.gr/dics/multi.htm

[8] Chiao YC, Zweigenbaum P. Looking for French-English translations in comparable medical corpora. Proc AMIA Symp. 2002:150-4.

[9] Binsted K, Cawsey A, Jones R. Generating personalized patient information using the medical record. In AIME'95, Pavia (Italy).

[10] Technical Specification for Multilingual Medical Dictionaries. Deliverable 20.3, NoE SemanticMining.

[11] MULTEXT: Common Specifications and Notation for Lexicon Encoding and Preliminary Proposal for the Tagsets. http://nl.ijs.si/ME/V3/msd/related/msd-multext/

[12] Schulz S, Hahn U. Morpheme-based, cross-lingual indexing for medical document retrieval. Int J Med Inform. 2000 Sep;58-59:87-99.

[13] Fiammetta Namer, Robert Baud. Guessing Lexical Relations between Biomedical Terms: towards a Multilingual Morphosemantics-based system. Proceedings of MIE2005, Geneva, IOS Press.

[14] Schulz S, Markó K, Sbrissia E, Nohama P, Hahn U. Cognate Mapping - A Heuristic Strategy for the Semi-Supervised Acquisition of a Spanish Lexicon from a Portuguese Seed Lexicon. 2004 (The 20th International Conference on Computational Linguistics. Geneva, Aug 2004: http://www.issco.unige.ch/coling2004/

[15] The TC37 SC4 website is http://www.tc37sc4.org/ and further information can be found in http://tagmatica.fr/doc/ISO24613wd.pdf