Programme

23/5/2006	**Acquiring and Representing Multilingual, Specialized Lexicons: the Case of Biomedicine**
9:30	***Multilingual Lexicon Alignment***
9:30	**Invited Talk:** *Developing a multilingual medical dictionary in the "Semantic Mining" European Network of Excellence* Stefan Schulz, Kornél Markó, Patrick Ruch, Pierre Zweigenbaum
10:00	*Subword Approach For Acquiring and Cross-Linking Multilingual Specialized Lexicons* Philipp Daumke, Stefan Schulz, Kornél Markó
10:30	*Cross-Lingual Alignment of Medical Lexicons* Kornél Markó, Robert Baud, Pierre Zweigenbaum, Magnus Merkel, Maria Toporowska-Gronostaj, Dimitrios Kokkinakis, Stefan Schulz
11h00	Coffee break
11:30	*Using Word Alignment to Extend Multilingual Medical Terminologies* Louise Deléger, Magnus Merkel, Pierre Zweigenbaum
12:00	***Distributed Development***
12:00	**Invited Talk:** *Papillon project: Retrospective and Perspectives* Mathieu Mangeot
13:00	*Staff vs. Volunteer: Two Approaches to Building a Multilingual Lexical Resource* Alan Melby, Marc Carmen, Steve Asher, Gerard Meijssen
13h30	Lunch break
14:30	***Normalization and Terminology***
14:30	**Invited Talk:** *LMF for multilingual, specialized lexicons* Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, Claudia Soria
15:30	*Unification of multi-lingual scientific terminological resources using the ISO 16642 standard. The TermSciences initiative* Majid Khayari, Stéphane Schneider, Isabelle Kramer, Laurent Romary
16:00	*The Development of a MeSH-based Biomedical Termbase at Hogeschool Gent* Joost Buysschaert
16h30	Coffee break
17:00	***Multilingual Lexicon and Ontology***
17:00	*Designing an Ontology of Dialogue Elements Modeling Doctor-Patient Exchanges* Leslie Barrett
17:30	*Populating ontologies in biomedicine and presenting their content using multilingual generation* Vangelis Karkaletsis, Alexandros G. Valarakos, Constantine D. Spyropoulos
18:00	**Panel:** *Methods for the acquisition and representation of multilingual, specialized lexicons*
19:00	End of workshop

Organisers

Pierre Zweigenbaum	AP-HP; Inserm; Inalco, Paris, France
Stefan Schulz	Freiburg University Hospital, Freiburg, Germany
Patrick Ruch	University Hospital of Geneva, Switzerland

Programme Committee

Sophia Ananiadou	National Centre for Text Mining, Manchester, UK
Lars Borin	Göteborg University, Göteborg, Sweden
Stefan Darmoni	University Hospital of Rouen, France
Gil Francopoulo	Tagmatica ; Inria, Paris, France
Vangelis Karkaletsis	NCSR "Demokritos" - IIT, Athens, Greece
Philippe Langlais	University of Montreal, Montreal, Canada
Christian Lovis	University Hospital of Geneva, Switzerland
Mathieu Mangeot	University of Savoie, Annecy, France
Magnus Merkel	Linköping University, Linköping, Sweden
Alessandro Oltramari	University of Trento & LOA-CNR, Trento, Italy
Patrick Ruch	University Hospital of Geneva, Switzerland
Stefan Schulz	Freiburg University Hospital, Freiburg, Germany
Pierre Zweigenbaum	AP-HP; Inserm; Inalco, Paris, France

Additional Reviewers

Laure Vieu	IRIT-CNRS & LOA-CNR, Trento, Italy

Table of Contents

Multilingual Lexicon Alignment

Distributed Development

Normalization and Terminology

Multilingual Lexicon and Ontology

Author Index

Foreword

The development of general, multilingual lexicons and dictionaries has received much interest (see, e.g., projects EuroWordNet or Papillon). In contrast, less attention has been given to multilingual lexicons and dictionaries in specialized domains, where the focus has been put instead on terminological products such as thesauri and classifications. The need nevertheless exists for large, shared, multilingual lexicons and dictionaries which support natural language processing in specialized domains. Medicine, with terminologies ranging above a million terms, is a case in point: only few specialized lexicons exist for this domain, English being as usual the best served. This workshop aimed to examine how specialized lexicons, in particular in biomedicine and beyond English, can be acquired, represented, and linked across languages.

It is endorsed by the European Network of Excellence "Semantic Interoperability and Data Mining in Biomedicine" (NoE 507505), Workpackage 20, "Multilingual Medical Lexicon".

Topics for submission included the following, non-exhaustive list:

- building specialized lexicons;
- standards for sharing multilingual lexicons;
- methods for acquiring and cross-linking specialized lexicons of different languages;
- specific issues in representing specialized, technical lexicons (e.g., neoclassical compounds, Latin words, etc.).

While medicine was expected to be the main focus of the workshop, work on other specialized domains was welcome too.

A first group of papers address the topic of *Multilingual Lexicon Alignment.* Daumke *et al.* show how "subwords" can help to link specialized words and terms across languages. Markó *et al.* apply this method to link a series of monolingual, medical lexicons. Instead of relying on morphology, Deléger *et al.* exploit parallel corpora to identify existing term translations.

A second group of papers focus on the *Distributed Development* of large, multilingual lexicons. In his invited presentation, Mangeot looks back at the Papillon experience after five years of distributed, international collaborative work on a multilingual dictionary. Melby *et al.* propose to adopt a Wiki-based approach to volunteer-based development of multilingual, medical lexical resources.

A third group of papers deal with *Normalization and Terminology.* The invited talk by Francopoulo *et al.* presents the Lexical Markup Framework (LMF), the result of normalization work within ISO. Khayari *et al.* instantiate the Terminological Markup Framework (TMF) in a multilingual term base which includes medical sources such as the MeSH thesaurus. Buysschaert reports on the development of a Dutch translation of the MeSH thesaurus and related issues, including that of standards for representing such a term base.

The last group of papers, *Multilingual Lexicon and Ontology*, consider ontologies as a pivot representation to which lexical information can be attached for multiple languages. Barrett applies this principle to English-Arabic medical terms for doctor-patient communication. Karkaletsis *et al.* propose a corpus-based method to populate such an ontology.

Pierre Zweigenbaum, Stefan Schulz, Patrick Ruch

Subword Approach For Acquiring and Cross-Linking Multilingual Specialized Lexicons

Philipp Daumke, Stefan Schulz, Kornél Markó

Freiburg University Hospital, Department of Medical Informatics, Freiburg, Germany

Abstract

We present a new subword-based approach to automatically translate biomedical terms from one language to another. The approach may support the creation of new multilingual biomedical lexicons and make the cross-linking between different languages possible. Using subwords, i.e. morphologically meaningful units, instead of full words significantly reduces the number of lexical entries to sufficiently cover a specific language and domain. The language transfer between queries and documents is based on these subwords, as well as on lists of word-n-grams that are generated from large monolingual corpora and serve as look-up tables for different target languages. First tests were done for the translation of German terms into English.

1. Introduction

The automatic translation of biomedical terms between different languages using some sort of aligned word lists poses a big challenge whenever the coverage of terms or the linkage between these terms is not comprehensive. Particularly in languages such as German, Finnish or Swedish that are characterized by a high frequency of compounds an exhaustive list of cross-linked biomedical terms is not yet available.

This paper presents a new approach to automatically translate biomedical terms from one language to another. It combines a lexicon- and corpus-based approach that is able to translate both dictionary and out-of-dictionary biomedical terms. At its core lies a multilingual subword lexicon that contains semantically minimal, morpheme-style units called subwords. Language-specific subwords are linked by intralingual as well as interlingual synonymy and grouped into language-independent equivalence classes. Using an interlingua significantly reduces the number of entries that are needed to sufficently cover the biomedical domain. Our approach additionally exploits large monolingual word lists that are easily acquired from the web for many languages. These lists are analyzed with regard to term frequencies and correspondences of word orders.

2. Morpho-Semantic Indexing

The MorphoSaurus system is based on the assumption that neither fully inflected nor automatically stemmed words constitute the appropriate granularity level for lexicalized content description. Especially in scientific sublanguages, we observe a high frequency of complex word forms such as in *'pseudo⊕hypo⊕para⊕thyroid⊕ism'*. To properly account for particularities of 'medical' morphology , the notion of subwords was introduced as self-contained, semantically minimal units.

Subwords are assembled in a multilingual dictionary and thesaurus, which contain their entries, special attributes and semantic relations between them. Subwords are listed as entries together with their attributes such as language and subword type (stem, prefix, suffix, invariant). Each lexicon entry is assigned to one or more morpho-semantic identifier(s) representing the corresponding synonymy classes (MIDs). Intra- and interlingual semantic equivalence are judged within the context of medicine only.

Figure 1 depicts how source documents (top-left) are converted into an interlingual representation by a three-step morpho-semantic indexing procedure. First, each input word is orthographically normalized (top-right). Next, words are segmented into sequences of subwords or left unaffected when no subwords can be decomposed (bottom-right). Finally, each meaning-bearing

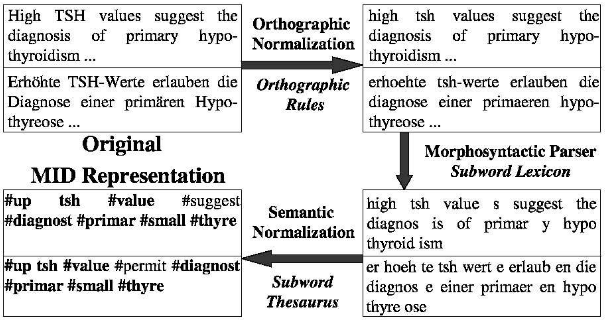

High TSH values suggest the diagnosis of primary hypothyroidism ...
Erhöhte TSH-Werte erlauben die Diagnose einer primären Hypothyreose ...
Orthographic Normalization
Orthographic Rules
high tsh values suggest the diagnosis of primary hypothyroidism ...
erhoehte tsh-werte erlauben die diagnose einer primaeren hypothyreose ...
Original
Morphosyntactic Parser
Subword Lexicon
MID Representation
#up tsh #value #suggest #diagnost #primar #small #thyre
Semantic Normalization
high tsh value s suggest the diagnos is of primar y hypo thyroid ism
#up tsh #value #permit #diagnost #primar #small #thyre
Subword Thesaurus
er hoeh te tsh wert e erlaub en die diagnos e einer primaer en hypo thyre ose

Figure 1: Morpho-Semantic Indexing Pipeline

subword is replaced by a language-independent semantic identifier, its MID, thus producing the interlingual output representation of the system (bottom-left). MIDs which co-occur in both document fragments appear in bold face.

3. Term Translation

3.1. Creating Subword Lists

In a preparation phase we acquired large (medical) domain specific corpora in different languages from the Web including abstracts from medical journals indexed in Medline[1] as well as different online health portals such as *Mayo Clinic*[2] or *Netdoctor*[3]. These corpora are normalized by removing HTML tags and stop words, transforming characters with diacritics into 7-bit ASCII by applying language specific transliteration rules and removing all non 7-bit ASCII tokens. Subsequently, these normalized corpora are tokenized into word-n-grams (henceforth, *target words*). We limited n to values between 1 and 3 resulting in lists of surface words, word bigrams and trigrams. These temporary lists are uniquely sorted counting the number of occurrences. Table 1 lists the number of generated word-n-grams for ENGlish, GERman, PORtuguese, SPAnish, FREnch and SWEdish.

The target words are now sent to the morpho-semantic normalization routine which assigns a sequence of MIDs to this input (cf. Figure 1). The resulting language specific *target lists* contain triples of the form (*target words*, *frequency*, *MIDs*). Due to the frequent occurrence of subword permutations between languages (e.g. German *"Bluthochdruck"* (literally *"blood high pressure"*) vs. English *"high blood pressure"*), bigrams and trigrams on the interlingual MID layer are ordered alphabetically. Table 2 shows a small subset of the English *target list*.

Language	Unigrams	Bigrams	Trigrams
ENG	528k	30,257k	97,673k
GER	467k	4,101k	5,530k
POR	138k	3,899k	7,058k
SPA	125k	2,382k	3,746k
FRE	85k	1,129k	1,796k
SWE	47k	423k	782k

Table 1: Number of Generated Target Queries in Different Languages (k = 1000)

Target Words	Freq	MIDs
...	...	...
side	111k	#side
side effects	76k	#effect #side
pancreatitis	9k	#itis #pancreas
heparin	574	#heparin
...	...	...

Table 2: Extract of the English Target List

3.2. Producing Translations

When a term T_{orig} is sent to our translation tool (with specified term language and desired target language), T_{orig} is transformed to its MID representation T_{MID}. Subsequently, T_{MID} is iteratively matched against the MIDs in the target list of the desired language starting with the MID sequence that correspond to the first n words of T_{orig} ($n <= 3$ depending on the length of T_{orig}). Note that the MIDs of each sequence are, again, ordered alphabetically. In case of no match, the number of tokens in this MID sequence is reduced by one and the matching process is repeated. Once a match of MID sequences is successful the translation term is returned and the matching process reiterates using the MID sequence that correspond to the next three words of T_{orig}. This procedure is repeated until all MIDs in T_{MID} are processed.

Let's take the German T_{orig} *"Nebenwirkungen von Heparin"* (English: *"side effects of heparin"*) as an example that shall be translated into English, as depicted in table 3. Firstly, T_{orig} is transformed to the MID representation T_{MID}

[1] http://www.ncbi.nlm.nih.gov/entrez/

[2] http://www.mayoclinic.com/

[3] http://www.netdoctor.co.uk/

”*#side #effect #heparin*”.

Q_{orig}	*Nebenwirkungen von Heparin*
Q_{MID}	#side #effect #heparin
$Matching\ Process$	#effect #heparin #side → <-> #effect #side → <side effects> #heparin → <heparin>
$Translation$	*side effects heparin*

Table 3: Translation Process for T_{orig}

As the number of words in T_{orig} is smaller than 3, the matching process against the english target list starts with all MIDs. As no translation can be found for ”*#effect #heparin #side*”, the MID sequence is reduced by one MID and ”*#effect #side*” is matched against the target list which returns ”*side effects*”. Next, ”*#heparin*” is looked up in the target list and ”*heparin*” is returned.

4. Experiments

In a preliminary evaluation we evaluated our approach on a list of 200 German UMLS[4] entries that were translated into English using our translation approach. When creating the term list, we firstly excluded all entries from the German UMLS list that contained symbols (such as hyphens, commas etc.) since our subword lexicon is not yet particularly adapted to medical abbreviations or chemical expressions. From the resulting list we randomly chose a set of 200 entries. The average number of words per entry is 1,35. *Zahnverletzungen (engl. Tooth Injuries)* or *Dapson (engl. dapsone)* are typical entries.
After applying our algorithm, the resulting translation list was manually evaluated by a medical expert. For each term we offered check boxes with three alternatives: *Exact Translation* (EX) for an exact translation of the UMLS term, *Related Translation* (REL) for translations at which the textual meaning of a translation is right, but the grammatical number, the word order or the part of speech varied, or stop words (e.g. *of*) are missing, respectively, and *Wrong Translation* (WRO) for all erroneous translations. As a baseline we determined the number of identical cognates, i.e. the number of terms that have their source terms as its own translation.

5. Results

Table 4 shows the results of our evaluation. The German-English translation achieves 78% exact or related translations, i.e. 156 of 200 translations are correct or related, respectively. 22% (44) translations are erroneous. Our baseline, i.e. the number of identical German-English cognate pairs achieved 34,5% (70) correct translations.

LANG	EX	REL	WRO	BASE
GER	55	23	22	34,5

Table 4: Results for the translation of 200 German UMLS terms into English (in %)

6. Discussion

While our first test runs are quite promising the number of 22% wrong tranlations still show the need of a detailed failure analysis to optimize our approach.
1. Incomplete Coverage (89%): Most of the erroneous translations are due to incomplete coverage of our subword lexicon. This indicates that with additional lexicographic effort our approach may achieve results over 90%.
2. Ambiguity (11%): In four cases our approach produced an erroneous translation due to ambibguity. Here, either a subword that is linked to two other subwords, lead to the wrong translation alternative (e.g. ”*#steuer*” is linked to the two english MIDs ”*#control, #tax*” and ”*control*” was returned instead of ”*taxes*”) or our approach found the right translation of each subword in their textual meaning, but in the context of the other translation units this was erroneous. E.g. the German expression ”*Foetale Gefaehrdung*” (engl. ”*fetal distress*”) was translated into ”*dangerous fetuses*”. Here, both (*Foetale → fetuses*) and (*Gefaehrdung → danger*) are correct/related when viewed separately.

7. Related Work

Term translation approaches recur - with different focus - in several different research contexts including Cross Language Information Retrieval (Levow et al., 2005), Corpora Alignment (Resnik, 1998), Word Sense Disambiguation (Markó et al., 2005a) or Automatic Lexicon

[4] http://www.nlm.nih.gov/research/umls/

Acquisition (Markó et al., 2005b). In all these contexts the translation of unknown, so called out-of-vocabulary terms are a major challenge. Usually existing bilingual word lists are used as seed lexicons, or parallel, related or even unrelated corpora are exploited.

Baud et al. (1998) applied a multilevel method to automatically create a bilingual English-French dictionary of nearly 10.000 word pairs exploiting co-occurrences of words in the ICD-10 classification. Similar to our subword based approach they transform compounds or derivational words to their to underlying concepts using a dictionary with 8.000 entries. The resulting word pairs proved correct in 98% of the cases.

Schulz et al. (2004) introduced a method of directly translating terms from Portuguese to Spanish using simple string transformation rules. These translations are then validated in the local context of language-specific corpora resulting in a list of biomedical cognate pairs.

Claveau and Zweigenbaum (2005) propose an algorithm that infers transducers from examples of bilingual word pairs. They achieve up to 85% of correct translations for translations between French and English. This approach, again, counts for biomedical simple terms (composed of one word) only and may be less effective in languages in which word compounding is used extensively (such as German, Dutch or Swedish).

In a previous work Chiao and Zweigenbaum (2002) identified translational equivalents of out-of-dictionary words from French to English in the medical domain relying on non-parallel, comparable corpora and an initial bilingual medical lexicon. They achieved about 60% of correct translations in the top ten candidates.

8. Acknowledgments

This work was supported by the European Network of Excellence "Semantic Mining" (NoE 507505).

9. References

RH Baud, C Lovis, AM Rassinoux, PA Michel, and JR Scherrer. 1998. Automatic extraction of linguistic knowledge from an international classification. In *MEDINFO'98 – Proceedings of the 9th World Congress on Medical Informatics. Vol. 1*, pages 581–585. Seoul, Korea, August 1998.

Y.C. Chiao and P. Zweigenbaum. 2002. Looking for french-english translations in comparable medical corpora. In Isaac S. Kohane, editor, *AMIA 2002 – Proceedings of the Annual Symposium of the American Medical Informatics Association. Biomedical Informatics: One Discipline*, pages 150–154. San Antonio, TX, November 9-13, 2002.

Vincent Claveau and Pierre Zweigenbaum. 2005. Translating biomedical terms by inferring transducers. In *Artificial Intelligence in Medicine. Proceedings of the 10th Conference on Artificial Intelligence in Medicine in Europe – AIME 2005*, volume 3581 of *Lecture Notes in Artificial Intelligence*, pages 236–240. Aberdeen, Scotland, July 23 - 27, 2005.

Gina-Anne Levow, Douglas W. Oard, and Philip Resnik. 2005. Dictionary-based techniques for cross-language information retrieval. *Information Processing and Management: an International Journal*, 41(3):523–547.

Kornél Markó, Stefan Schulz, and Udo Hahn. 2005a. Unsupervised multilingual word sense disambiguation via an interlingua. In *AAAI 2005 – Proceedings of the 20th National Conference on Artificial Intelligence & IAAI'05 – Proceedings of the 17th Innovative Applications of Artificial Intelligence Conference*, pages 1075–1080. Pittsburgh, PA, USA, July 9-13, 2005.

Kornél Markó, Stefan Schulz, Alyona Medelyan, and Udo Hahn. 2005b. Bootstrapping dictionaries for cross-language information retrieval. In *SIGIR 2005 – Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 528–535. Salvador, Brazil, August 15-19, 2005.

Philip Resnik. 1998. Mining the web for bilingual text. In *ACL'99 – Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 527–534. College Park, MD, USA, 20-26 June 1999.

Stefan Schulz, Kornél Markó, Eduardo Sbrissia, Percy Nohama, and Udo Hahn. 2004. Cognate mapping: A heuristic strategy for the semi-supervised acquisition of a Spanish lexicon from a Portuguese seed lexicon. In *COLING Geneva 2004 – Proceedings of the 20th International Conference on Computational Linguistics*, volume 2, pages 813–819. Geneva, Switzerland, August 23-27, 2004.

Cross-Lingual Alignment of Medical Lexicons

Kornél Markó[1], Robert Baud[2], Pierre Zweigenbaum[3], Magnus Merkel[4], Maria Toporowska-Gronostaj[5], Dimitrios Kokkinakis[5], Stefan Schulz[1]

[1]Freiburg University Hospital, Department of Medical Informatics, Freiburg, Germany
[2]University Hospitals of Geneva, Service of Medical Informatics, Geneva, Switzerland
[3]Inserm, U729; Assistance Publique – Paris Hospitals, STIM; Inalco, CRIM, Paris, France
[4]Linköping University, Department of Computer and Information Science, Linköping, Sweden
[5]Göteborg University, NLP Section, Department of Swedish, Göteborg, Sweden

Abstract

We present an approach for the creation of a multilingual medical dictionary for the biomedical domain. In a first step, available monolingual lexical resources are compiled into a common interchange format. Secondly, according to a linking format defined by the authors, the cross-lingual mappings of lexical entries are added. We show how these mappings can be generated using a morpho-semantic term normalization engine, which captures intra- as well as interlingual synonymy relationships on the level of subwords.

1. Introduction

There is currently no large electronic dictionary in the medical domain which is characterized by a true multilingual dimension, relevant coverage, and substantial lexical information. Multilinguality means at least that the corresponding entries in different languages are connected, which is a difficult task and raises simple questions and concerns open issues, like e.g., in which cases a translation relationship truly holds for lexical entities. Therefore, syntactical as well as semantic criteria have to be developed, or, at least, a consensus of different lexical input providers has to be found.

Within the European Network of Excellence "Semantic Interoperability and Data Mining in Biomedicine", a multinational team of researchers from (computational) linguistics, medicine, and medical informatics, including the authors, gathered in a series of meetings with the goal of building a European multilingual medical lexicon with high coverage and the integration of complete morpho-syntactic information.

Of course, monolingual resources exist for different languages, so the first task to merge them is to create a common framework for the integration of lexical entities from different languages, with respect to their intrinsic peculiarities.

2. Interchange Format Definition

The Interchange Format is a convention about the way to exchange linguistic information entering in the building process of a medical multilingual lexicon (Baud et al., 2005). The basic idea is that the exchange of information is performed through the Interchange Format only, and each contributor of lexical resources is converting his or her data into that representation.

Table 1 lists the fields of the interchange format. The most important ones are the following:

- **Typ** The *basic entry* (B) encodes single words. The *subword entry* (S) is a marker for parts of words entering in the composition of a *compound entry* (C). Finally, a *term entry* (T) describes a sequence of words.
- **Lem** The lemma is the representation of the entry in its basic form. It is supposed to be recoverable from any occurring form by an inflectional morphology process.
- **Mul** The code for encoding morphological and syntactic information is defined as in the open standard MULTEXT.[1]

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

Field	Description	Definition
Lng	Language	the language to which pertains the present entry
Id	Multilingual Identifier	the unique identifier of this entry
Typ	Entry Type	one of the 4 allowed types of entry (B,C,S,T)
Lem	Lemma	the entry in its basic form
Mul	Morpho-syntactic Features	the MULTEXT morpho-syntactic tag of the lemma
Frm	Inflected Form	any inflected form
Mfr	Features of Inflected Form	the MULTEXT morpho-syntactic tag of the inflected form
Inf	Inflection Model	language specific information
Mis	Language Specific Argument	to be used freely by provider of entries
Prt	Decomposition	the decomposition of a compound entry into its parts
Str	Head	the head word of the term
Ref	Reference Lemma	ID of its lemma's entry (if inflection form)

Table 1: Fields of the Lexicon Interchange Format

- **Frm** Inflected form that is linked to an entry for its lemma through the **Ref** field.
- **Mfr** The morpho-syntactic features of the inflected form using MULTEXT exactly as for the **Mul** field.

Table 2 shows an excerpt of different lexicons encoded in the Interchange Format. One obvious shortcoming is that the different lexical resources provide different amounts of information.

3. Monolingual Resources

After agreeing upon the Interchange Format, partners from five different institutions collected their monolingual lexical resources.[2] These are:

- the French UMLF lexicon from different French health-related organizations and the University Hospitals of Geneva, Switzerland (33,718 entries) (Zweigenbaum et al., 2004)
- an English lexicon from Linköping University, Sweden (22,686 entries)
- a Swedish lexicon from Linköping University (23,223 entries)
- a Swedish lexicon from Göteborg University, Sweden (6,786 entries)
- the German Specialist Lexicon from Freiburg University Hospital, Germany (41,316 entries) (Weske-Heck et al., 2002)

Up until now, 127,730 lexical entries for the biomedical domain, fully encoded with morpho-syntactical features, were collected covering four European languages (cf. Table 2 for a sample[3]).

4. Linking Format Definition

The cross-lingual grouping of corresponding entries is the essence of a multilingual dictionary.

Unfortunately, this is not a straightforward process and a couple of cross-lingual phenomena are problematic to capture, especially regarding the different characteristics of case, gender and number in different languages, as well as multiple derivations, e.g. for adjectives, dependent on whether a definite or indefinite object follows or whether their use is attributive or predicative.

Consider the German words *Schere* and *Hose* (both noun, singular) and the English equivalents, *scissors* and *trousers* (both noun, plural). Singular forms of the latter examples do not exist, whilst for other examples, of course, singular forms can be translated to a corresponding singular form in the other language.

Different languages also make different use of grammatical gender or noun classes. Whilst in German, Greek or Latin, three grammatical gen-

[2] The English Specialist Lexicon, which is part of the Unified Medical Language System (UMLS, 2005), will be included in future work.

[3] The first character of the *Mul* field encodes the part-of-speech: N (noun), A (adjective). In case of nouns, c denotes common nouns, m masculine, s singular, n neuter or nominative, depending on the position. For adjectives, f stands for qualitative, p positive. The character " $-$ " indicates that a particular feature does not fit into the language given (e.g. gender in English) or is unspecified for this entry.

Lng	Id	Typ	Lem	Mul	Frm	Mfr	Prt	Str
FR	UMLF:10081	B	doigt	Ncms				
EN	LIU:EN8427	T	finger nail	Nc-sn				nail
SV	LIU:SV6663	B	digital	Afp-sn				
SV	UGOT:3373	C	fingeravtryck	Nc-sn			finger–avtryck	avtryck
DE	UKLFR:39556	B	Fingerpanaritium	Ncnsn	Fingerpanaritien	Ncnpa		

Table 2: Sample of Compiled Lexical Resources (some fields omitted)

Field	Description	Definition
Src	Source Entry ID	The Id of the source entry to be linked to a target entry
Tar	Target Entry ID	The Id of the target entry linked from the source entry
Typ	Link Type	Type of relation

Table 3: Fields of the Linking Format

ders are distinguished (masculine, feminine and neuter), French and Italian only use two (masculine, feminine). Swedish and Danish discriminate the classes *common* and *neuter*. Finally, English does not account for any of these features at all.
In a first version, in order to find an agreement on the question, in which cases two lexical items, *A* and *B*, can be regarded as translations (or, within one language, synonyms) of each other, we defined the following ”levels” of relationships:

1. **Rel1:** *A* and *B* share the same part of speech (POS) and all MULTEXT features

2. **Rel2:** *A* and *B* share the same POS, but at least one MULTEXT feature differs

3. **Rel3:** *A* and *B* do not share the same POS

Having these types of relations in mind, we created a simple Linking Format, which is depicted in Table 3.
So far, the meaning of words and their possible translations have not been discussed. In the following section, we show how lexical entities can be aligned on the semantic level.

5. Cross-Lingual Alignment

For the medical domain, methods for the automatic search for translation candidates have already been explored. One promising idea is to use already existing translations at a subword level in order to support the acquisition of translations at a term level (Namer and Baud, 2005). For the linkage of lexemes on the semantic level, we make use of the MORPHOSAURUS system (Markó et al., 2005), a text normalization engine using subword lexicons for different languages, as well as a multilingual thesaurus.

5.1. Morpho-Semantic Indexing

The MORPHOSAURUS system is based on the assumption that neither fully inflected nor automatically stemmed words constitute the appropriate granularity level for lexicalized content description. Especially in scientific sublanguages, we observe a high frequency of complex word forms such as in *‘pseudo⊕hypo⊕para⊕thyroid⊕ism’*. To properly account for particularities of ‘medical’ morphology , the notion of subwords was introduced as self-contained, semantically minimal units.
Subwords are assembled in a multilingual dictionary and thesaurus, which contain their entries, special attributes and semantic relations between them. Subwords are listed as entries together with their attributes such as language and subword type (stem, prefix, suffix, invariant). Each lexicon entry is assigned to one or more morpho-semantic identifier(s) representing the corresponding synonymy classes (MIDs). Intra- and interlingual semantic equivalence are judged within the context of medicine only.
Figure 1 depicts how source documents (top-left) are converted into an interlingual representation by a three-step morpho-semantic indexing procedure. First, each input word is orthographically normalized (top-right). Next, words are segmented into sequences of subwords or left unaffected when no subwords can be decomposed (bottom-right). Finally, each meaning-bearing

Src	Tar	Typ	Lng1	Lem1	Mul1	Lng2	Lem2	Mul2
LIU:EN147	LIU:SV151	REL1	EN	abdominal hernia	Nc-sn	SV	bukbråck	Nc-sn
LIU:EN143	UKLFR:34985	REL2	EN	abdominal aorta	Nc-sn	DE	Bauchaorten	Ncfpn
LIU:EN947	UMLF:1123	REL3	EN	alveolar	Afp–n	FR	alvéole	Ncfs

Table 4: Sample Links between Lexical Items (some fields omitted). Additionally, the MULTEXT values of the corresponding items are depicted in Column four to nine. Also cf. Footnote 3 for the explanation of *Mul* values.

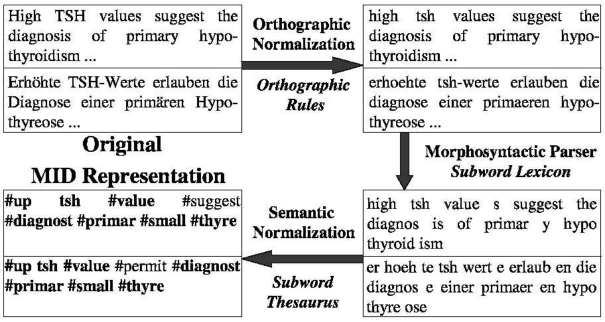

High TSH values suggest the diagnosis of primary hypothyroidism ...
Orthographic Normalization
high tsh values suggest the diagnosis of primary hypothyroidism ...
Erhöhte TSH-Werte erlauben die Diagnose einer primären Hypothyreose ...
Orthographic Rules
erhoehte tsh-werte erlauben die diagnose einer primaeren hypothyreose ...
Original
MID Representation
Morphosyntactic Parser
Subword Lexicon
#up tsh #value #suggest #diagnost #primar #small #thyre
Semantic Normalization
high tsh value s suggest the diagnos is of primar y hypo thyroid ism
#up tsh #value #permit #diagnost #primar #small #thyre
Subword Thesaurus
er hoeh te tsh wert e erlaub en die diagnos e einer primaer en hypo thyre ose

Figure 1: Morpho-Semantic Indexing Pipeline

subword is replaced by a language-independent semantic identifier, its MID, thus producing the interlingual output representation of the system (bottom-left). MIDs which co-occur in both document fragments appear in bold face.

5.2. Linking Algorithm

In a first step, all lexical entries are processed with the MORPHOSAURUS system. Afterwards, a quite simple algorithm was used to perform the mappings between all entries: Every lexeme i and its attributes is compared to any other lexeme j in the list. If their representations in the interlingua format are identical, they are considered as potential translations or synonyms and linked. Then the relation type (REL1, REL2 or REL3, cf. previous section) is determined, by comparing the lexical attributes.

6. Results and Conclusion

Using the algorithm introduced, we obtained a list of 300,894 bi-directional relations between lexemes, some of which are depicted in Table 4. For English-German, 30,716 translations have been generated, for English-French 16,123 and for English-Swedish 27,441. Furthermore, 21,619 relations have been extracted for French-Swedish, 32,805 for French-German and finally, 41,966 for German-Swedish. All ther relations (130,224) cover intralingual synonymy. The distribution of different types of relations is 32,805 occurrences for REL1 (11%), 147,145 for REL2 (49%) and 120,944 for REL3 (40%). First examinations of the data proved many alignments to be valid. Of course, an extensive evaluation of the multilingual medical lexicon is still due. Further work will also examine relations with ongoing lexicon standardization efforts such as the Lexical Markup Framework of ISO/TC 37/SC 4.[4]

Acknowledgments: This work was supported by the European Network of Excellence “Semantic Mining” (NoE 507505).

7. References

Robert Baud, Mikaël Nyström, Lars Borin, Robert Evans, Stefan Schulz, and Pierre Zweigenbaum. 2005. Interchanging lexical information for a multilingual dictionary. In *Proceedings of the Annual Symposium of the American Medical Informatics Association*, pages 31–35.

Kornél Markó, Stefan Schulz, and Udo Hahn. 2005. Morphosaurus - design and evaluation of an interlingua-based, cross-language document retrieval engine for the medical domain. *Methods of Information in Medicine*, 44(4):537–545.

Fiammetta Namer and Robert Baud. 2005. Guessing lexical relations between biomedical terms: towards a multilingual morphosemantics-based system. In *Proceedings of the 19th International Congress of the European Federation of Medical Informatics.*

UMLS. 2005. *Unified Medical Language System.* Bethesda, MD: National Library of Medicine.

Gesa Weske-Heck, Albrecht Zaiss, Stefan Schulz, Wolfgang Giere, Michael Schopen, and Rüdiger Klar. 2002. The German Specialist Lexicon. In Isaac S. Kohane, editor, *AMIA 2002 – Proceedings of the Annual Symposium of the American Medical Informatics Association.*, pages 884–888.

Pierre Zweigenbaum, Robert Baud, Anita Burgun, Fiammetta Namer, Éric Jarrousse, Natalia Grabar, Patrick Ruch, Franck Le Duff, Jean-François Forget, Magaly Douyère, and Stéfan Darmoni. 2004. A unified medical lexicon for French. *International Journal of Medical Informatics*, 74(2):119–124.

[4] http://tagmatica.fr/doc/ISO24613cdRev7.pdf

Using Word Alignment to Extend Multilingual Medical Terminologies

Louise Deléger[*], **Magnus Merkel**[†], **Pierre Zweigenbaum**[*‡]

[*]Inserm U729, Paris, France
[†]Dept of Computer and Information Science, Linköping University, Sweden
[‡]Assistance Publique – Paris Hospitals, STIM; Inalco, CRIM, Paris, France
louise.deleger@spim.jussieu.fr, mme@ida.liu.se, pz@biomath.jussieu.fr

Abstract

Medical terminologies such as those provided in the UMLS are never exhaustive and there is a constant need to enrich them, especially in terms of multilinguality. We present a methodology to acquire new French translations of English medical terms based on word alignment in a parallel corpus — i.e. pairing of corresponding words. We automatically collected a 27.7-million-word parallel, English-French corpus. Based on a first 1.3-million-word extract of this corpus, we detected 3,255 French translations of English MeSH terms, among which 1,956 are new translations.

1. Introduction

The UMLS Metathesaurus is an extensive vocabulary database that gathers and provides a link between different existing biomedical terminologies. But despite being a multilingual resource, it is mostly composed of English vocabulary, and other languages such as French are under-represented in comparison to English. There is therefore a need to enrich the terminologies of the UMLS. The acquisition of new translations of English terms is required. This is the purpose of the VUMeF[1] project which aims at extending the French part in the UMLS and which provides the background for this work.

Plenty of multilingual texts can be found as regards a specific domain but exhaustive terminologies and dictionaries are far less numerous — as can be seen in the case of the UMLS. Hence the idea of using parallel corpora (collections of multilingual texts) to enlarge terminologies. So instead of employing a human translator, we can make use of existing translated texts from which translations at the term level can be extracted.

We present a methodology to acquire medical terms based on word alignment in a parallel corpus. Word alignment is a natural language processing technique and is used in several applications such as terminology development (which is the case here), automatic translation and cross-lingual information retrieval. It consists in pairing words that are translations of each other in a parallel corpus.

Previous work has addressed the issue of multilingual medical terminologies. Chiao and Zweigenbaum (2002) collect translations from comparable corpora. Baud et al. (1998) make use of already parallel medical vocabularies to derive word translations. Widdows et al. (2002) use a statistical vector model on a corpus aligned at the document level. Névéol and Ozdowska (2005) have an approach similar to ours in that it deals with word alignment in parallel medical corpora to extract French translations of English terms. However, we deal with a larger corpus and process all kinds of alignments.

Our task involves issues such as dealing with errors in the alignment process that will spread from step to step, and detecting multi-word units — a term being either a single word or a multi-word expression.

This work is outlined in the following way: based on a French-English corpus (2.1), we align sentences (2.2) and words (2.3, 2.4). Medical terms are then selected (2.5) through the projection of a list of English terms from the MeSH. We obtain a list of bilingual English-French medical terms that we review. We extract samples for evaluation purposes (2.6) and expose results (3.). We discuss (4.) and conclude (5.) on the method.

2. Material and Method

2.1. Corpus

The corpus used for this experiment is collected from the web. The web is indeed a powerful resource for building corpora, both in terms of quantity and multilinguality. The quality of such a corpus can nevertheless be questioned and this might account for a proportion of the noise detected in the results.

Our corpus is gathered from a bilingual (French-English) Canadian health web site[2]. It is intended for the general public as well as for specialists and the proportion of specialized terms might therefore be lower than in resources dedicated to medical specialists.

Several techniques exist for building a parallel corpus from the web (Resnik and Smith, 2003; Patry and Langlais, 2005). We generate pairs of parallel documents (i.e., documents that are translations of each other) using information contained in the document structure — namely, HTML links to corresponding documents in the other language. Indeed, after a study of the documents, we noticed that each document provided access to its translation page through an image or a text tag labelled in the corresponding language (specified in the "alt" attribute of the HTML tag). This gives us 11,041 pairs of parallel documents and a total of 27.7 million words.

Documents from the web usually come in either HTML or PDF format, and need to be converted to text format. As

[1] Project VUMeF (French Unified Medical Vocabulary), led by Stéfan Darmoni (Darmoni et al., 2003), is partially funded by the French Ministry of Research (National Network of Health Technologies).

[2] http://www.hc-sc.gc.ca/

for us, we have HTML documents and thus have access to structural information that may be useful to keep even in a text file. After cleaning the HTML files and converting them to XML format, we use a XSLT stylesheet to transform them to text format while keeping a number of information — title, paragraph and link tags which will be used as correspondence points for sentence alignment. Indeed, we assume that a title in a source language corresponds to a title in a target language, a link to a link, and in most cases a paragraph to a paragraph. The resulting texts are segmented into sentences to prepare the way for further processing.

2.2. Sentence Alignment

The first step towards word alignment is to align the corpus at sentence-level. Sentence alignment is a mandatory task since there is not a full one-to-one correspondence between the sentences of two parallel texts. Although it is most common that one sentence in a source language corresponds to one sentence in a target language, there are instances where one sentence is translated with two — or sometimes even three or more — sentences, and this needs to be determined before working at the word level.

To do so, we use Dan Melamed's GMA [3] (Geometric Mapping and Alignment) (Melamed, 2000), a robust tool which performs sentence alignment of parallel texts using both statistical and linguistic techniques. It is based, among other things, on length measurements, bilingual lexicons and cognates (words sharing similar spelling and meaning). Though sentence aligners in general and GMA in particular achieve high-quality performances, any mistake at this level will be reflected at the next one — i.e., word alignment — and will make things even harder for this already complex process. So, in order to work on cleaner data, we attempt to automatically detect and remove incorrect sentence alignments as well as bad document pairing (documents that are not parallel) using criteria such as sentence length and quality evaluation of sentence alignment.

2.3. Word Alignment

Once sentences are aligned, we can proceed to word alignment. This task is far more problematic than sentence alignment. There is no true word-to-word correspondence between the words of two sentences. A word is often translated with several ones, or can be omitted in the corresponding sentence (this is typically the case for grammatical words that are specific to a language). Parallel sentences, though being translations of each other, can differ considerably in terms of structure. In that case even a human has trouble determining which words should be paired together. The results we expect are therefore on a lower level than from the previous sentence alignment task.

The issue of the type of word alignment should be raised. That is, are we satisfied with a word-to-word alignment? The objective of this work is to obtain medical terms. A term can be either a single word or a multi-word unit. A common approach is to first extract candidate terms using a separate tool — a term extractor — and then to proceed to their alignment (Daille et al., 1994; Gaussier, 1998). The originality of this work is that we do not separate the detection of candidate terms from the alignment process. In other words we use a tool that is able to detect multi-word units and to align both single words and multi-word expressions.

Word alignment systems usually derive from either statistical approaches or linguistic ones, or a combination of both. Statistical methods (Brown et al., 1993) involve co-occurrence measures and probability scores, and are especially effective on large corpora with high-frequency words but performances decrease with low-frequency occurrences. Linguistic ones (Wu, 2000) make use of information such as syntactic parsing. They are less robust despite being able to deal with low-frequency words. Hybrid approaches (Ahrenberg et al., 2000; Barbu, 2004) seem to be a good compromise.

2.4. Aligning Words with the I*Tools

We use the I*Tools suite (developed at Linköping University, Sweden) to perform word alignment. We chose these tools partly because they are based on a hybrid approach, using both statistical and linguistic techniques. They also align multi-word units which suits our terminological purpose. They make use of resources such as co-occurrence measures, bilingual dictionaries, POS tagging and syntactical analysis.

A pre-processing step is required: the corpus is tagged and lemmatized (using Treetagger [4]) and syntactically annotated (with the syntactic analyzer SYNTEX (Bourigault et al., 2005)). The files are transformed into XML format encoding this information.

The alignment process with the I*Tools can be divided into three steps: training, automatic alignment and review of the results; each one corresponding to a specific tool of the suite. Training and review are both done with graphical, interactive tools that are fast to work with.

Training of the system is manually done using a special tool of the suite — the I*Link [5] interactive aligner (Merkel et al., 2003). This tool proposes word pairings to the user who accepts or rejects them. The user's decisions are stored into the resources of the system and by learning from them, the performances become increasingly accurate. These resources provide training data for the automatic word aligner.

The corpus is then automatically aligned by I*Trix, the automatic aligner of the suite, using the resources created with I*Link. We obtain a list of word alignments — i.e., source words paired with target words. The system can also exploit data created during the next step (the reviewing phase). In that case, the automatic alignment is repeated after a first run and takes into account the review made by the user. This is useful if the results first obtained are not as good as expected.

Results are reviewed with the I*View tool which enables the user to confirm, reject, or simply remove an alignment. An alignment is « removed » when it is neither an error nor a correct alignment, meaning it is a partial alignment (some

[3] http://nlp.cs.nyu.edu/GMA/

[4] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

[5] http://www.ida.liu.se/~nlplab/ILink/

parts are correct). This tool also indicates for each alignment a quality score, which enables the user to rank the alignments. The quality score used in the I*Tools is based on the mutual information formula (Stolz, 1965). Mutual information has been used in several works, including (Church and Hanks, 1990) which derives a new measure for estimating word associations and (Fabry et al., 2005) which uses mutual information for term extraction to build a terminology. In our case, the measure is expressed in terms of the frequency of the words as a pair, and the frequency of each source and target word of the pair independently in other pairs. This means that for a proposed word pair which occurs with a high frequency and where the source word and the target word only occur in this pair and not in any other suggested word pairs, we have good reasons to assume that the quality is high. On the other hand if the word pair has a low frequency and the source and target words of the pair are found in several other suggested pairs, then there is reason to be more doubtful to the suggestion. The formula is Q=f(st)/n(s)+n(t), where f(st) is the frequency of the word pair and n(s) and n(t) are the number of different word pairs in which the source and target words occur respectively.

2.5. Term Selection and review

In practice, there is no need to review all of the results since we are only looking for medical terms. We thus retain only those likely to be of interest. We select them using an English medical terminology — namely MeSH (as extracted from the 2005AC version of the UMLS). We project this list of terms onto the English entries of our alignment pairs and select those present in the pairs. Only then do we review the alignments. These alignments constitute French candidate translations of English MeSH terms. The review can be done by a linguist engineer. Afterwards, we can determine the proportion of new translations retrieved and submit these translations to medical experts.

In order to restrain manual interaction, we also tested a different solution, that is, not reviewing the results directly after the term selection, but only after some filtering — elimination of duplicates, verbs, translations already existing in the French version of the MeSH, and alignments with a poor quality score. Indeed, we consider that MeSH terms are mainly nouns and that verbs are not needed in the selection — they introduce too much noise. Low quality score alignments are also likely to be errors and might not be worth reviewing. This filtering phase will reduce the amount of terms to be reviewed.

2.6. Implementation and Evaluation

The methodology described above was implemented as follows:

1. conversion of the corpus into text format;
2. sentence alignment;
3. training of the automatic word aligner on a set of 600 sentences randomly taken from our corpus, by interacting manually with I*Link;
4. automatic word alignment with I*Trix. If results seem poor, a first review may also be done followed by a second run of I*Trix.
5. selection of medical terms (see section 2.5);

Implementation 1:

6. review, with I*View, of the alignments for the terms selected.

Implementation 2:

6. filtering of the results:
   - elimination of duplicates
   - elimination of verbs (the MESH entries are considered to be nouns)
   - elimination of terms already registered in the French MeSH
   - elimination of pairs with a poor quality score (equal to 0); however, there may be correct alignments among the low quality score ones. In that case, a small proportion of translations will be lost. We have tested the implementation both with this step and without it.
7. review of the results.

Evaluation was performed at several points of the implementation. First, we performed an evaluation of the quality of the alignment at step 2 by checking 100 sentences randomly taken from the corpus and measuring the percentage of correct alignments (precision measure).

The quality of word alignment was evaluated at step 4 by measuring precision on two samples: sample 1 consists of 100 word pairs randomly taken from the whole resulting pairing, and sample 2 of 100 word pairs taken from the best word alignments (alignments with a frequency higher than 1 and a good quality score — equal to or higher than 1).

Last, step 6 of implementation 1 allowed us to have a gold standard to evaluate word alignment for the medical terms. We evaluated the performances using information retrieval evaluation techniques, namely precision-recall measures. Other teams have also used these measures for evaluating tasks aside from information retrieval — text categorization for instance (Larkey and Croft, 1996). In information retrieval, precision is computed at 11 recall points from a list of retrieved documents. In our case, we used a ranked list of alignments instead of documents, considering that an alignment being correct is similar to a document being relevant for a query. We used trec_eval[6] to compute these recall points and obtained a precision-recall curve. These measures are calculated on the basis of the alignments ranked by frequency and quality score, meaning that the first alignments are expected to be the best ones. These recall-precision points also allowed us to measure the mean average precision.

This step was also useful to determine the proportion of errors and correct alignments in the filtered results at step 6 of implementation 2, thereby allowing us to experiment with the setting of a threshold for the quality score of the alignments to be filtered out.

3. Results

3.1. Valid Alignments (Language Engineer)

We completed steps 1 and 2 on the whole corpus, thus obtaining 1.1 million sentence pairs. As the corpus is huge,

[6] http://trec.nist.gov/trec_eval/trec_eval.8.0.tar.gz

	Sample 1	Sample 2	Set of medical terms
Precision	50%	92.2%	52%
Errors	19.6%	4.9%	30.3%
Partial alignments	30.4%	2.9%	17.7%

Table 1: Evaluation figures for word alignment

we have currently processed only part of it from step 3 to the end — a set of 540 pairs of documents (1.3 million words) gathered in two corresponding files. From this set, we obtained 91,171 word alignments and selected 10,392 pairs of medical terms.

Among these pairs, there are 2,567 different source terms (a term can have several translations), so we have a mean value of 4.05 French translations per English term. 5,403 alignments were confirmed as correct ones — by a language engineer (LD) — which gives a precision of 52% (see table 1) and a mean value of 2.1 correct translations per term. We count 2,159 different source terms in the confirmed alignments, meaning that 408 terms only had incorrect translations.

Table 1 shows that evaluation results for the overall quality of word alignments (step 4 of 2.6) are very good for the top alignments (sample 2, taken from the 7,366 best alignments as described in 2.6) and average for the whole aligned corpus (sample 1). As for sentence alignment (step 2 of 2.6), we achieved a precision of 95%, which is excellent.

Precision-recall figures for the evaluation of the set of medical terms (as described in 2.6, step 6) with trec_eval are detailed in the table on figure 1 and emphasize the previous statement that precision is excellent for the first alignments and decrease afterwards. To be more accurate, these recall points were computed on a scale of 10,000 alignments instead of the standard of 1,000 documents used in information retrieval. The increasingly descending slope of the curve on figure 2 shows that the ranking algorithm does push the majority of incorrect alignments towards the end of the list, with an inflection around 60% recall, obtaining more than 80% precision. The mean average precision measured is indeed 82%.

A proportion of the noise in word alignment can be attributed to errors in the sentence alignment process: 17% of the incorrect alignments are due to bad sentence pairing. Other factors include errors in POS tagging, bad document pairing (in our case we observed some English-English document pairs) and low quality of the data — misspelling of words, insertion of spaces inside a word, missing spaces between words.

3.2. Useful Medical Translations (Medical Specialist)

If we take a look at the resulting list of 5,403 confirmed medical term alignments, we notice 306 pairs that are not real translations but merely pairs of English words — i.e. the English words have not been translated. These are considered correct alignments but are of no use for our purpose, so we simply ignore them. Among the remaining 5,097, we eliminate a number of duplicates (pairs that are the same but were not considered as such by the alignment tools due to case differences) and obtain 3,607 word pairs. As stated in 2.5, we do not consider verbs as valid candidate translations and we eliminate them, thus lowering down the number of translations to 3,255. In this set, we look at the number of different concepts (CUI in the UMLS), terms already present in the French version of the MeSH and new translations (see table 3). The translations include morphological variants — for instance adjectives instead of prepositional phrases — and synonyms. However we do not consider plural/singular and masculine/feminine morphological variants as new translations.

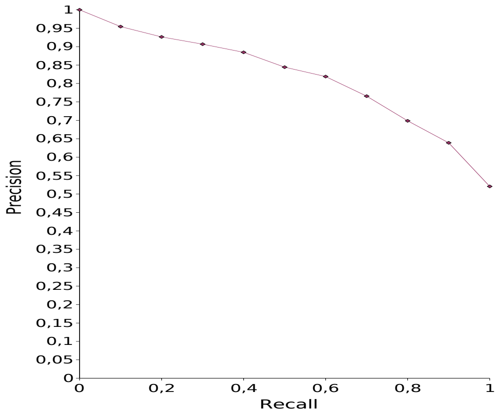


Table 2: Precision at 11 recall levels, measured with trec_eval on a scale of 10,000 alignments

A sample of 145 MeSH terms (see figures in table 3) was also extracted for validation purposes. 79 terms had new translations which were submitted to expert validation. 64 have been validated. Examples of translations are given in table 4.

	Complete set of MeSH terms	Validation sample
Translation pairs	3,255	145
Concepts	1,868	138
Already registered	1,299	66
New	1,956	79
New and valid		64 (81%)

Table 3: Figures for the MeSH translations

English	French	Valid
bone cancer	cancer des os	Yes
breast milk	lait maternel	Yes
reproduction rights	droits de reproduction	No

Table 4: Translation examples

The second implementation tested — i.e., reviewing only the filtered results — gives the following figures. From the 10,392 pairs of selected terms, we eliminate duplicates (8,699 resulting pairs), verbs (7,985 resulting pairs), and select only the new translations (not registered in the French

MeSH), which gives us 6,670 candidate translations to be reviewed. Thanks to the first implementation, we can easily determine the proportion of noise. Since we expect 1,956 new translations, 4,714 should be eliminated. Among these, there are incorrect alignments (4,452) and English-English pairs (262). We can see that the precision is very low, but that was expected. Since we are only looking at new translations pairs, incorrect word alignments are bound to be considered as new. Most of the noise is therefore selected and we do have the same balance as in implementation 1. Evaluation of the alignment is here meaningless and should only be performed on the non-filtered results. But this implementation enables us to considerably reduce the amount of word pairs to be reviewed: we have to review 6,670 alignments instead of 10,392, that is, 35.8% less.

We also tested filtering out alignments with a low quality score. We set the threshold to 0: all alignments with a score equal to 0 are removed. With this criterion, we lower down the selection to 4,493 alignments, which means that we now have 56.8% less to review and 2,766 noisy alignments. Among the 2,177 removed, 238 were actually correct. So there is a loss of information, but the proportion remains small (12%). Alignments with a score equal to 0 have an 89% chance of being incorrect, which can justify for their being removed.

4. Discussion

Our approach presents a number of advantages as well as some drawbacks. It allows us to acquire medical terms which are actually used in French documents for certain MeSH descriptors but are not registered in the current French version of the MeSH. It does not require a human translator and makes the best of existing resources. In terms of word alignment, we are able to process noisy data quite efficiently. We do not use monolingual term extractors and we align single words and multiword expressions with a uniform approach unlike other methods which concentrate on 1-1 and 2-2 word alignments (Névéol and Ozdowska, 2005).

Though being an automatic approach, it still needs human help in the process (training and validation). The success of this method is also heavily dependent on the efficiency of word alignment which is a complex task. However, the remaining processing of the rest of the medical web corpus, if done incrementally, could steadily increase the quality of word alignment. Using the techniques outlined in this paper to minimize the reviewing process it should be possible to rapidly include verified data in each step and include this as positive training data for each new iteration. If the corpus is divided into roughly 25 sets containing just over a million words per set, and these subcorpora were processed one by one, with a short reviewing process included, the confirmed entries of each run could be fed into the training data for the next run of the automatic alignment. We also assume that interactive training using I*Link will not be necessary for each subsequent iteration, which means that the manual time spent on each new iteration will decrease and the precision will likely increase due to new training data.

The quality of the corpus is an important feature and its choice is a major issue. In our case, we used the Web as a resource and processed a whole website. A study of the documents would have been useful in order to best characterize the type of data acquired — which documents are intended for medical specialists, which ones are for the general public and which ones have no medical content (index pages for instance). Interesting developments of this method will include the specific search for patient-oriented translations (consumer vocabulary) which are even more lacking in medical terminologies. This can be achieved, for instance, to look for candidate translations of Medline Plus vocabulary.

5. Conclusion

We described a methodology to acquire new translations of English medical terms in order to enrich existing medical terminologies. We argued that a natural language processing technique such as word alignment is an efficient way to do so. Indeed, we were able to find a number of new translations of English MeSH terms. Moreover, it is an automatic process which only requires limited human intervention. Finally, this method raises interesting prospects such as the acquisition of patient vocabulary, and more generally its application to other parallel corpora.

Acknowledgements

We thank Stéfan Darmoni for reviewing the candidate MeSH translations, and Didier Bourigault for providing SYNTEX.

6. References

Lars Ahrenberg, M Andersson, and Magnus Merkel. 2000. A knowledge-lite approach to word alignment. In Jean Véronis, editor, *Parallel Text Processing: Alignment and Use of Translation Corpora*. Springer.

Ana-Maria Barbu. 2004. Simple linguistic methods for improving a word alignment algorithm. In *Proceedings 7th International Conference on the Statistical Analysis of Textual Data*, pages 88–98, Louvain-la-Neuve, Belgium. Presses universitaires de Louvain.

Robert H Baud, C Lovis, AM Rassinoux, PA Michel, and Scherrer JR. 1998. Automatic extraction of linguistic knowledge from an international classification. In C Safran B Cesnik and P Degoulet, editors, *Proc 9th World Congress on Medical Informatics*, pages 581–5.

Didier Bourigault, Cécile Fabre, Cécile Fréerot, Marie-Paule Jacques, and Sylwia Ozdowska. 2005. Syntex, analyseur syntaxique de corpus. In *Proceedings Traitement automatique des langues naturelles (Traitement automatique des langues naturelles)*, Dourdan.

PF Brown, SAD Pietra, VJD Pietra, and RL Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2002. Looking for French-English translations in comparable medical corpora. *Journal of the American Medical Informatics Association*, 8(suppl):150–154.

K Church and P Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16:22–29.

Béatrice Daille, Éric Gaussier, and Jean-Marie Langé. 1994. Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the 15 th COLING*, pages 515–521, Kyoto, Japan, August.

Stéfan J. Darmoni, Éric Jarrousse, Pierre Zweigenbaum, Pierre Le Beux, Fiammetta Namer, Robert Baud, Michel Joubert, Huguette Vallée, Roger A. Côté, Antoine Buemi, Didier Bourigault, Gaelle Recourcé, S. Jeanneau, and Jean-Marie Rodrigues. 2003. Extending the French part of the UMLS. In Mark Musen, editor, *Proceedings AMIA Annual Fall Symposium 2003*, page 824, Washington, DC, November. AMIA. (poster).

P Fabry, R Baud, P Ruch, C Despont-Gros, and C Lovis. 2005. Methodology to ease the construction of a terminology of problems. *Int J Med Inform*.

Éric Gaussier. 1998. Flow network models for word alignment and terminology extraction from bilingual corpora. In Christian Boitet, editor, *Proceedings of the 17 th COLING*, Montréal, Canada, 10–14 August.

L S Larkey and W B Croft. 1996. Combining classifiers in text categorization. In *Proceedings of SIGIR*, pages 289–297. ACM Press, New York.

I. Dan Melamed. 2000. Bitext maps and alignments via pattern recognition. In Jean Véronis, editor, *Parallel Text Processing: Alignment and Use of Translation Corpora*. Springer.

Magnus Merkel, M Petterstedt, and Lars Ahrenberg. 2003. Interactive word alignment for corpus linguistics. In *Proceedings Corpus Linguistics*.

Aurélie Névéol and Sylwia Ozdowska. 2005. Extraction bilingue de termes médicaux dans un corpus parallèle anglais/français. In *Proceedings EGC'05*.

A Patry and Philippe Langlais. 2005. Paradocs: un systèeme d'identification automatique de documents parallèles. In Michèle Jardino, editor, *Proceedings of TALN 2005 (Traitement automatique des langues naturelles)*, pages 223–232, Dourdan, June. ATALA, LIMSI.

Philip Resnik and N.A. Smith. 2003. The Web as a parallel corpus. *Computational Linguistics*, 29:349–380. Special Issue on the Web as a Corpus.

W Stolz. 1965. A probabilistic procedure for grouping words into phrases. *Language and Speech*, 8:219–235.

D Widdows, B Dorrow, and C.K. Chan. 2002. Using parallel corpora to enrich multilingual lexical resources. In *Proceedings LREC*, pages 240–244, Las Palmas, Spain, May. ELRA.

Dekai Wu. 2000. Bracketing and aligning words and constituents in parallel text using stochastic inversion transduction grammar. In Jean Véronis, editor, *Parallel Text Processing: Alignment and Use of Translation Corpora*. Springer.

Papillon project: Retrospective and Perspectives

Mathieu Mangeot

Condillac-LISTIC
F-73376 Le Bourget du Lac Cedex
Mathieu.Mangeot@univ-savoie.fr

Abstract

This paper describes the first five years of life of the Papillon project with four main phases: the birth with the motivations of such a project; the extension with the decision to build a multilingual pivot dictionary; the implementation with the realization of "Jibiki", a generic dictionary management platform and the population with the use of semantic vectors for linking entries and an ongoing project: word games, for creating specific lexical information.

1. Introduction

This paper describes the first five years of life of the Papillon project which goal is to build a multilingual pivot dictionary with a rich microstructure. The idea is that everyone can contribute online to the dictionary. The resulting data is freely and publicly available.

The paper is divided in four sections, one for each phase of the project in historic order: the birth, the extension, the implementation and the population.

2. Phase I, birth: a French-Japanese bilingual dictionary

2.1. History

The Papillon project (first named FeJ for French-English-Japanese) (Boitet et al., 2002; Mangeot et al., 2004) was launched in early 2000 by Emmanuel Planas, François Brown de Colstoun and Mutsuko Tomokiyo. Emmanuel Planas was postdoc researcher at NTT Research Centre, located in Keihanna, Japan and François Brown de Colstoun was scientific attaché at the embassy of France in Tokyo, Japan. Mutsuko Tomokiyo was a linguist Ph.D. student in Grenoble, France.

They were confronted every day with the needs of a good French-Japanese dictionary. That was the starting point of the project.

The first institutional partners were the home institutions of the initiators: the GETA-CLIPS laboratory in Grenoble, France and the Embassy of France in Japan. The National Institute of Informatics (NII) also joined the project through contacts with NII researchers.

2.2. Motivations

The first motivations of the project were the following:

- **Few resources** The main problem is the lack of free and good French-Japanese dictionaries. The few complete French-Japanese resources are expensive, and tailored for Japanese speakers. The free lexicons available on the Web are very insufficient even for simple vocabulary (10,000 entries). Thus, the majority of French speakers have no choice but using English-Japanese dictionaries. This is also true for many other languages. Even for those with a good knowledge of English, it automatically adds confusion.

- **Lack of information** The most complete French-Japanese dictionaries were built for Japanese speakers, thus there is a lack of information necessary for French speakers: transliteration of kanji, numerical specifiers, etc.

- **High construction costs** The traditional way of building a dictionary needs lots of money and time. As an example, the construction of the EDR English-Japanese dictionary cost 1,200 human-year for about 300,000 entries in each language. The public price, 14,3 millions of yens (̃100,000 €) is so expensive that only companies can afford it. Furthermore, it does not even reflect the construction costs. The initiators had no choice but finding another way to build their dictionary.

- **Collaborative projects** An interesting way seems to launch a collaborative project like the LINUX construction paradigm. People contribute at their level. The result is free of rights and free so that every can benefit from it. At that time, there were already dictionaries building projects that were using this method, like the Edict Japanese-English dictionary project launched and still managed by Jim Breen for more than ten years. Now, the success of the Wikipedia project confirms our idea.

2.3. Meetings

The initiators had a user point of view of the dictionary. They were not specialists of computational lexicography. They decided to ask other researchers (mainly from GETA-CLIPS) to join the project and the decision was taken to hold the first Papillon meeting (Tomokiyo et al., 2000) at the National Institute of Informatics, Tokyo, Japan in August 2000.

Since then, we decided to organize a meeting every year. The 2001 meeting took place in July in Grenoble. The dictionary structures (Sérasset and Mangeot, 2001) were adopted during this meeting.

The 2002 meeting took place in July in Tokyo. We took there important decisions concerning the data built in the framework of the project: it is free of rights and freely and publicly available. In order to ensure a long life to the Papillon project, we organized our way of working in a way that it would not depend on any specific founds. The scientific leaders are university researchers with a full time position. The project advances also thanks to Ph.D. fellows or post-doctorate researchers whose subject integrates a scientific issue of the project. We decided also to organize every meeting as a workshop with scientific reviewing committee in parallel with an international conference so that it would be easier for researchers to obtain founds for coming.

The 2003 meeting took place in July in Sapporo. We discussed mainly about the platform used for building the dictionary. The 2004 meeting took place in August in Grenoble, France. The 2006 meeting took place in Chiang Rai, Thailand.

Every meeting gathers about roughly 50 people from all parts of the world. Nowadays, the main actors are Christian Boitet, Gilles Sérasset and Mutsuko Tomokiyo from GETA-CLIPS, Grenoble, France; Mathieu Lafourcade from LIRMM, Montpellier, France, Michael Zock from LIF, Marseille, France; Yves Lepage from ATR, Keihanna, Japan; Asanee KAwtrakul from Kasetsart U., Bangkok, Thailand; Jim Breen from Monash U., Melbourne Australia and myself ;-).

3. Phase II, extension: a multilingual pivot dictionary

3.1. History

The first idea was to build a multi-target French to English and Japanese dictionary following the model of the FeM French-English-Malay dictionary. But then, the research conducted at GETA-CLIPS on pivot structures and the opportunity to open the project to many languages led us to decide to build a multilingual pivot dictionary. In the same way, we decided to use an entry microstructure based on the word sense level and very detailed in order for the dictionary to be used both by humans and by machines.

3.2. Macrostructure

The multilingual pivot macrostructure with interlingual links is based on Gilles Sérasset's Ph.D. Thesis(Sérasset, 1994b; Sérasset, 1994a; Sérasset, 1994c) and has been experimented at a small scale by Etienne Blanc (Blanc, 1995) with the PARAX database.

This structure consists in one monolingual volume for every language of the dictionary and one pivot volume in the middle see Figure 1.

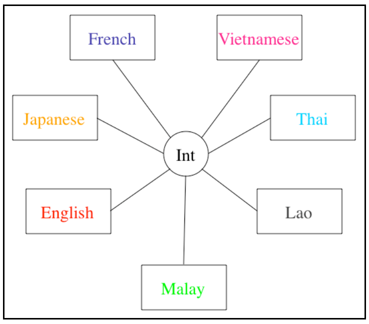


Figure 1: Multilingual Pivot Macrostructure

The monolingual volumes gathers monolingual entries at a word sense level, i.e. monolingual acceptions (called lexies). The entries of different languages are then linked between each others via interlingual acceptions (called axies) that can be seen as complex translation links. These acceptions may also be linked together by refinement links in order to cope with the semantic discrepancies between languages.

Each sense or meaning of each entry of a monolingual volume is linked to one or more acceptions of the pivot volume. For example, like in figure 2 in French “ affection ” has two meanings: “affection” and “disease”. The vocable “affection” will consequently be linked to two "lexies" (corresponding to two word senses) in the French monolingual dictionary, which in turn will be linked to two interlingual acception or "axies" in the pivot volume.

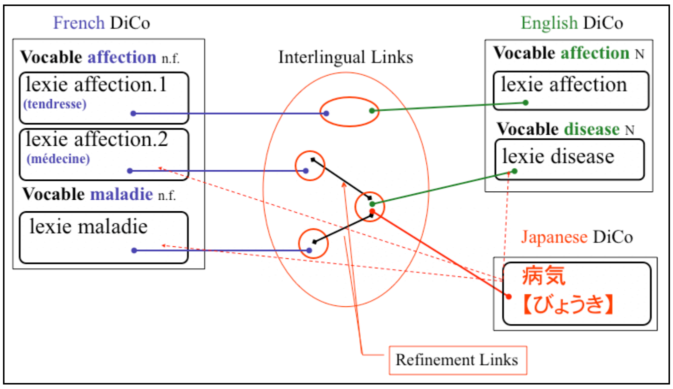


Figure 2: Macrostructure in Detail with Interlingual Links

3.3. Microstructure

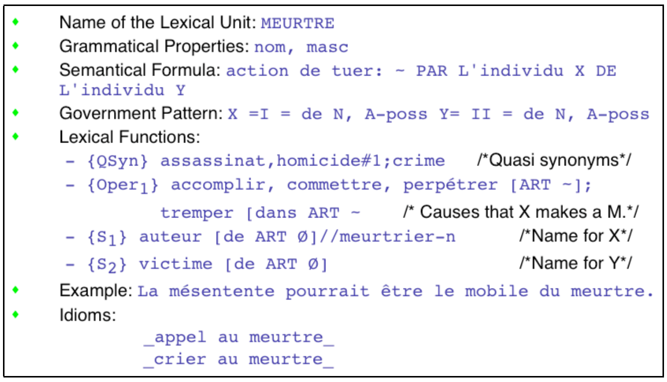
The structure of the entries or microstructure of the monolingual volumes is based on the structure used for the formal lexical database DiCo (Polguère, 2000) of the OLST laboratory in Université de Montréal. The encoding methodology is directly borrowed from the Explanatory and Combinatorial Lexicology (ECL)(Mel'čuk et al., 1995), which is part of the Meaning-Text Theory elaborated by Igor Melčuk and his colleagues first in Moscow, Russia and then in Montreal, Canada.

This structure, rather complex (see Figure 3) has been chosen for mainly two reasons:

1. It has been proven language independent and thus, appropriate for any of our languages present in our dictionary. Of course, there are some parts that are language dependent such as the grammatical properties or the language levels, but the main part remains the same.

2. It has been elaborated to be theoretically used both by humans or machines.

Each lexie or lexical unit is made of a name, grammatical properties (mainly a part of speech), a semantic formula which can be seen as a formal definition. In the case of a predicative lexie, it describes the entire predicate and its



- Name of the Lexical Unit: MEURTRE
- Grammatical Properties: nom, masc
- Semantical Formula: action de tuer: ~ PAR L'individu X DE L'individu Y
- Government Pattern: X =I = de N, A-poss Y= II = de N, A-poss
- Lexical Functions:

```
 - {QSyn} assassinat,homicide#1;crime     /*Quasi synonyms*/
 - {Oper₁} accomplir, commettre, perpétrer [ART ~];
          tremper [dans ART ~       /* Causes that X makes a M.*/
 - {S₁} auteur [de ART Ø]//meurtrier-n          /*Name for X*/
 - {S₂} victime [de ART Ø]                      /*Name for Y*/
```

- Example: La mésentente pourrait être le mobile du meurtre.
- Idioms:

```
        _appel au meurtre_
        _crier au meurtre_
```

Figure 3: Microstructure of the French lexie "MEURTRE"

arguments, a government pattern which describes the syntactic realization of the arguments of the predicate, a list of lexico-semantic functions. There is a fixed number of 56 basic functions that can be applied in any language. These functions can be combined to create more elaborated ones; a list of examples; a list of full idioms.

4. Phase III, implementation: an online generic dictionary management platform

4.1. History

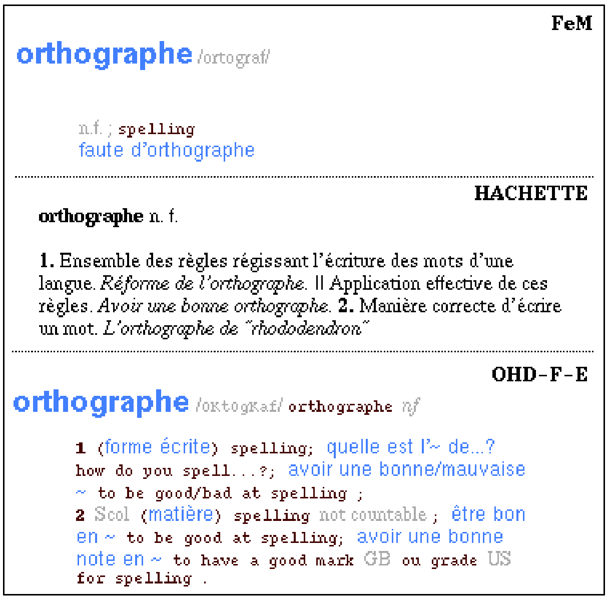
I began my Ph. D. (Mangeot, 2001) in 1998 taking the results of Gilles Sérasset's Ph.D. (Sérasset, 1994c) Ph.D. as a starting point and having the goal to implement a demonstrator. I implemented a first prototype in Perl called DicoWeb. It was able to query several dictionaries with different structures and display the results in the same window (this tool is still used daily at XRCE laboratory).
After the first Papillon meeting in July 2000, Gilles Sérasset and I began to implement a more robust prototype in Java based on the specifications described in my Ph.D. thesis with the goal to obtain a generic platform for managing (querying, editing, importing, exporting) dictionaries in any structure.
In order to follow the LINUX construction paradigm not only for the data but also for the software, we chose to use only free open-source software for building the platform. Furthermore, we plan to release it in the future as a sourceforge project.

4.2. The Jibiki Platform

The Jibiki platform[1] (Mangeot and Sérasset, 2002), (Sérasset, 2004) is a community web site primarily developed for the Papillon project. This platform is entirely written in Java using the "Enhydra[2]" web development Framework. All XML data is stored in a standard relational database (Postgres). This community web site proposes several services:

- a unified interface to simultaneously access the Papillon MLDB and several other monolingual and bilingual dictionaries;
- a specific edition interface to contribute to the dictionaries stored on the platform,
- an open document repository where registered users may share writings related to the project; among these documents, one may find all the papers presented in the different Papillon workshops organized each year by the project partners;
- a mailing list archive,

To encourage volunteers, we think that it is important to give a real service to attract as many Internet users as possible. As a result, we began our development with a service to allow users to access to many dictionaries with different structures but in a unified way (see Figure 4). This service currently gives access to thirteen (13) multilingual, bilingual and monolingual dictionaries, representing more than one million entries.



FeM

orthographe /ortograf/

n.f. ; spelling
faute d'orthographe

HACHETTE

orthographe n. f.

1. Ensemble des règles régissant l'écriture des mots d'une langue. *Réforme de l'orthographe.* || Application effective de ces règles. *Avoir une bonne orthographe.* **2.** Manière correcte d'écrire un mot. *L'orthographe de "rhododendron"*

OHD-F-E

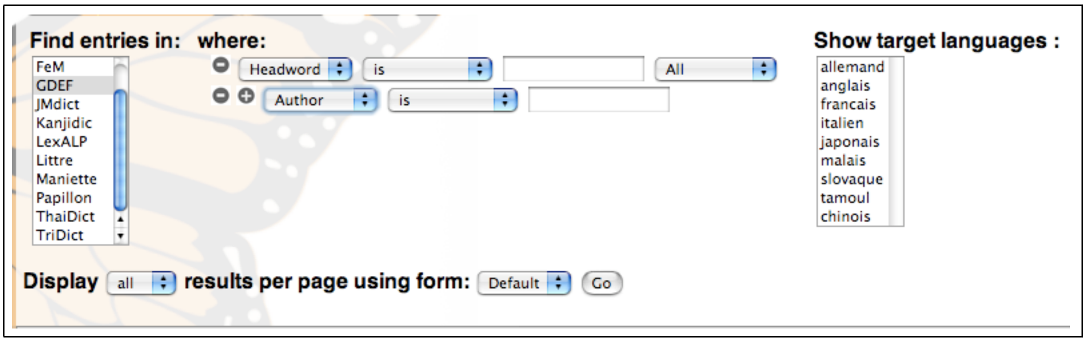
orthographe /ɔʀtɔgʀaf/ orthographe nf

1 (forme écrite) spelling; quelle est l'~ de...? how do you spell...?; avoir une bonne/mauvaise ~ to be good/bad at spelling ;
2 Scol (matière) spelling not countable ; être bon en ~ to be good at spelling; avoir une bonne note en ~ to have a good mark GB ou grade US for spelling .

Figure 4: Query of "Orthographe" in three dictionaries

Every available dictionary will be queried according to its own structure from a multi-criteria search interface (see 4.2.). Moreover, all results will be displayed in a form that fits the structure. Any monolingual, bilingual or multilingual dictionary may be added in this collection, provided that it is available in XML format. With the Jibiki platform, giving access to a new, unknown, dictionary is a matter of writing two XML files: a dictionary description and an XSL stylesheet. For currently available dictionaries, this took an average of about one hour per dictionary.
The description file gathers dictionary meta-information and a minimum set of information in the dictionary's XML structure. The Jibiki platform defines a standard structure of an abstract dictionary containing the most frequent subset of information found in most dictionaries. This abstract structure is called the Common Dictionary Markup (Mangeot, 2002). To describe a new dictionary, one has to write an XML file that associate CDM elements to pointers in the original dictionary structure.

[1] see http://jibiki.univ-savoie.fr/jibiki
[2] see http://www.enhydra.org/



Find entries in: FeM, GDEF, JMdict, Kanjidic, LexALP, Littre, Maniette, Papillon, ThaiDict, TriDict

where:
Headword | is | | All
Author | is |

Show target languages : allemand, anglais, francais, italien, japonais, malais, slovaque, tamoul, chinois

Display all **results per page using form:** Default Go

Figure 5: Multicriteria Advanced Search Interface in Several Dictionaries

Along with this description, one has to define an XSL style sheet that will be applied on requested dictionary elements to produce the HTML code that defines the final form of the result. If such a style sheet is not provided, the Jibiki platform will itself transform the dictionary structure into a CDM structure and apply a generic style sheet on this structure.

4.3. The key feature: an online generic editor

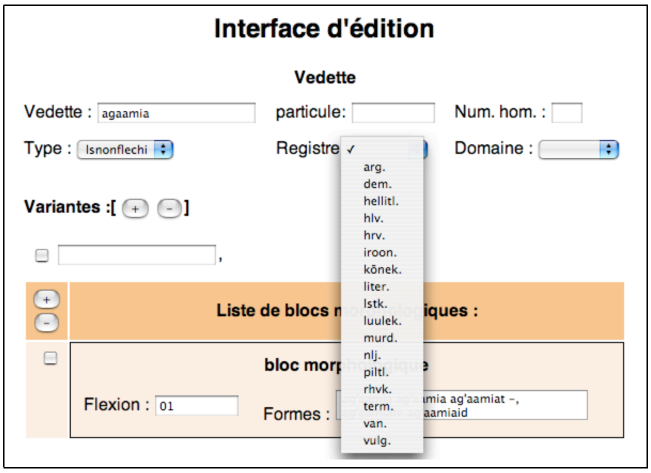
The main purpose of the Jibiki platform is to gather a community around the development of one or several dictionaries. Thus, the crucial challenge that we faced was to provide a way to edit the dictionary entries directly on the platform. It was specifically difficult because we wanted to be able to edit any kind of dictionary entry (the editor had to adapt itself to the structure of the entries) and to edit them online with a simple browser (it had to be bult only with a combination of HTML forms and simple javascripts). We did not even want to use java applets because of compatibility problems.

A preliminary version of the editor (Mangeot and Thevenin, 2004) was developed in collaboration with David Thevenin with his tool called ARTStudio for the development of adaptative plastic user interfaces. It was fragile and very difficult to handle. Furthermore, some parts of the code were not open source. Thus, a new simplified version has been recoded from scratch afterwards.

The new editor works with a template XHTML interface that is instanciated with the entry that the user wants to edit. This template can be generated automatically from a description of the entry structure in XML schema. It can be modified afterwards for improving the rendering on the screen. Thus, the only data needed to edit a dictionary entry on the jibiki platform (apart from the dictionary metadata described previously) is the XML schema of the structure of the entry and furthermore, any type of dictionary entry as long as it is encoded in XML.

We chose to use XML schema because it allows for a finer description compared to DTDs (for instance, we may define the set of valid values of the textual content of an XML element). Moreover XML schemata provides a simple inheritance mechanism that is useful for the definition of a dictionary.

HTML forms are very limited. The available interactors are text fields, radio buttons, check boxes and pop up menus. It was not enough to be able to edit complex entries. Thus, we had to build more complex interactors from the combination of the previous ones in order to handle lists (adding,deleting, moving an item on a list) and links (links to entries in the same volume or other ones). These elements can be themselves complex objects containing lists of other objects, etc. Any user, who is registered and logged in to the Papillon web site, may contribute to the Papillon dictionary by creating or editing an entry. Moreover, when a user asks for an unknown word, he is encouraged to contribute it to the dictionary. Contribution is made through a standard HTML interface (see Figure 6).



Interface d'édition

Vedette

Vedette : agaamia
particule:
Num. hom. :
Type : Isnonflechi
Registre ✓ arg. dem. hellitl. hlv. hrv. iroon. kõnek. liter. lstk. luulek. murd. nlj. piltl. rhvk. term. van. vulg.
Domaine :

Variantes :[+ -]

□ ,

Liste de blocs morphologiques :

□ **bloc morphologique**

Flexion : 01
Formes : ag'aamia ag'aamia ag'aamiat –, ag'aamiate ag'aamiaid

Figure 6: Interface for Writing an Entry

Every change made in the entry is stored in a history. It is then possible to come back to any previous version of the entry just like the usual "undo" commands. The writing process is divided in several steps depending on the project. The GDEF is the most complete with three steps:

- A contributor writes an entry;
- It is next revised by a reviewer;
- It is then validated by a validator;

4.4. Conclusion

The platform is now used by four different projects:

- the Papillon[3] project,
- the GDEF[4] project (Chalvin and Mangeot, 2006), about a bilingual Estonian-French dictionary,
- the LexALP[5] project (Sérasset, 2005), about a multilingual (English, French, German, Italian and Slovene) terminological database on the legal terms of the alpine convention,
- the TriDict trilingual (Sinhala, Tamil, English) dictionary.

There are still lots of ongoing developments on the platform with still a perspective of genericity in the different resources handled.
For those who want to use the platform for their projects, we are open to any collaboration. The only condition is that all the data produced with the platform must be publicly available and free of rights.

5. Phase IV, population: semantic vectors and word games

In order to facilitate the construction of Papillon dictionary, we decided to reuse existing data. The hypothesis are that it is easier to correct existing data than to build new data from scratch and that the public users prefer to have slightly incorrect data that no data at all when they lookup words in a dictionary.
The population faces two serious issues: the building of interlingual links between the lexies and the specific lexical information that is not available in any dictionary. We decided to use semantic vectors for the first issues and word games for the second one.

5.1. Semantic Vectors

The first problem is augmented by the fact that we chose to work at the word sense level, not at the vocable level. There is no unique way to divide a word into senses. In two dictionaries of the same language, for many entries, the division into word senses will be different. Thus, when one wants to merge the entries of two different dictionaries at the word sense level, s/he has to find a way to cope with this problem.
The solution we found uses semantic vectors in order to calculate the semantic distance between two lexies of two different dictionaries we want to merge and to determine if they can be merged or not.
The conceptual vectors model has been presented in (Lafourcade, January 2001; Lafourcade et al., 2002). Each textual segment (word, phrase, text) is linked with a thematic association that is represented by a vector of concepts. The set of concepts is predefined and constitutes a multidimensional vector space on which the word senses can be projected.
In this vector framework, it is possible to use the notion of *similarity* (usually used in information retrieval) and *angular distance* between two vectors. It will be used as an evaluation of the *thematic distance* beween word senses.

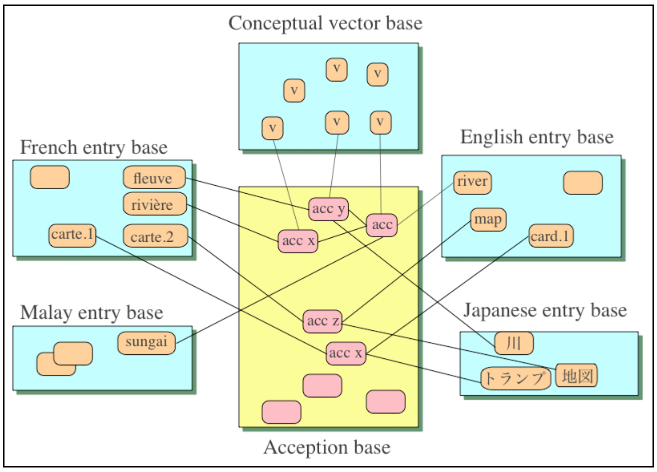


Figure 7: Linking Acceptions with Vectors

In order to merge lexies coming from different dictionaries, the first step is to calculate the conceptual vector that is linked to each of the lexies. For example, in French, the set of concepts is predefined with the 873 concepts of the Larousse thesaurus.
The second step is to bootstrap the computation by manually indexing 5,000 terms in each language.
Then the definition of each lexie is analyzed with a morphological analyzer. Then, using the manual indexed vectors of known words and the resulting analysis tree, we compute the vectors associated to each lexie and word-form. The process is reiterated until a stability is reached.
Once the process is finished, the dictionary is "vectorized". It is then possible to merge two dictionaries of the same language by looking at the thematic distance of the conceptual vectors of each lexie.
We consider that two conceptual vectors are close enough if their thematic distance is less than a threshold t. The more the threshold is low, the more the lexies can be considered as being merged. Nevertheless, it may be difficult to merge completely automatically the lexies. An acceptable value for the threshold is $\pi/4$.

5.2. Word Games

The issue is to find methods for building some particular crucial lexical data which is furthermore not available in existing dictionaries. It is the case for collocations coded in our entries through lexical functions.
For example, in English, the notion of "fever" is intensified by the adjective "strong, the notion of "smoker" by the adjective "heavy", etc. or, more particularly for asian languages, special counters must be used for specific types of objects. In Japanese, "wa" is the counter for the rabbits (usagi san wa, 3 rabbits) and "hiki" is the counter for cats (neko ni hiki, 2 cats).
The goal of the project "jeu de mots" (word game) is to experiment and study the use of "word games" for building or

[3] see http://www.papillon-dictionary.org
[4] see http://www.estfra.ee
[5] LexALP: Legal Language Harmonisation System for Environment and Spatial Planning within the Multilingual Alps

collecting precise lexical information. The idea is to generate automatically or semi-automatically word games tat can take the shape of a multiple-choice test (e.g. Is it possible to say ... in English ?) or fill-In-the-blank exercises (complete "strong fever, heavy smoker, ______ rain")

Each generated exercise will be used to complete or validate the information available in the dictionaries. Te targeted languages in this project framework are Chinese, French, Japanese, Malay and Thai. The exercises will be submitted to students and web surfers, (via the Papillon website) who will work on their mother tongue. The answers collected will be analyzed and the method will be tested and evaluated on each language. The gathered information will be publicly available on the Papillon website.

This project has been accepted and funded by the French government under the STIC-Asia program driven by INRIA research organization.

6. Conclusion

We presented a very challenging project that is already six years old and has already produced interesting results theoretically with research on multilingual pivot structures and practically through "jibiki", an online generic dictionary management platform.

We are welcoming anybody who is motivated by the project and wants to join the project. It is mainly based on voluntary work and aims to build a reference lexical resource.

7. References

Etienne Blanc. 1995. Une maquette de base lexicale multilingue à pivot lexical : Parax. In *Lexicomatique et Dictionnairique, Actes du colloque LTT*, pages 43--58. Universités Francophones, Actualités scientifiques, AUPELF-UREF.

Christian Boitet, Mathieu Mangeot, and Gilles Sérasset. 2002. The papillon project: cooperatively building a multilingual lexical data-base to derive open source dictionaries and lexicons. In Graham Wilcock, Nancy Ide, and Laurent Romary, editors, *Proc. of the 2nd Workshop NLPXML 2002, Post COLING 2002 Workshop*, pages 93--96, Taipei, Taiwan, 1 September.

Antoine Chalvin and Mathieu Mangeot. 2006. Méthodes et outils pour la lexicographie bilingue en ligne : le cas du grand dictionnaire estonien-français. In *EURALEX 2006, à paraître*, Turin, Italie, 6-9 septembre.

Mathieu Lafourcade, Violaine Prince, and Didier Schwab. 2002. Vecteurs conceptuels et structuration émergente de terminologies. *TAL*, 43(1):43--72.

Mathieu Lafourcade. January 2001. Lexical sorting and lexical transfer by conceptual vectors. In *First International Workshop on MultiMedia Annotation (MMA'2001)*, page 6, Tokyo.

Mathieu Mangeot and Gilles Sérasset. 2002. Frameworks, implementation and open problems for the collaborative building of a multilingual lexical database. In Grace Ngai, Pascale Fung, and Kenneth W. Church, editors, *Proc. of SEMANET Workshop, Post COLING 2002 Workshop*, pages 9--15, Taipei, Taiwan, 31 August.

Mathieu Mangeot and David Thevenin. 2004. Online generic editing of heterogeneous dictionary entries in papillon project. In *Proc. of the COLING 2004 conference*, volume 2, pages 1029--1035, Geneva, Switzerland, 26 August.

Mathieu Mangeot, Gilles Sérasset, and Mathieu Lafourcade. 2004. Construction collaborative d'une base lexicale multilingue. *Traitement Automatique des Langues*, 44(2):151--176, February.

Mathieu Mangeot. 2001. *Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue*. Thèse de nouveau doctorat, spécialité informatique, Université Joseph Fourier Grenoble I, Septembre.

Mathieu Mangeot. 2002. An xml markup language framework for lexical databases environments: the dictionary markup language. In *LREC Workshop on International Standards of Terminology and Language Resources Management*, pages 37--44, Las Palmas, Spain, 28 May.

Igor Mel'čuk, Andre Clas, and Alain Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*. Universites francophones et champs linguistiques. AUPELF-UREF et Duculot, Louvain-la Neuve.

Alain Polguère. 2000. Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for french. In *Proceeding of EURALEX'2000, Stuttgart*, pages 517--527.

Gilles Sérasset and Mathieu Mangeot. 2001. Papillon lexical database project: Monolingual dictionaries and interlingual links. In *NLPRS-2001*, pages 119--125, Tokyo, 27-30 November.

Gilles Sérasset. 1994a. An interlingual lexical organisation based on acceptions, from the parax mock-up to the nadia system. In *ICLA-94*, pages 21--33, July.

Gilles Sérasset. 1994b. Interlingual lexical organisation for multilingual lexical databases in nadia. In Makoto Nagao, editor, *COLING-94*, volume 1, pages 278--282, August.

Gilles Sérasset. 1994c. *Sublim : un système universel de bases lexicales multilingues et Nadia : sa spécialisation aux bases lexicales interlingues par acceptions*. Thèse nouveau doctorat, Université Joseph Fourier-Grenoble 1, Décembre.

Gilles Sérasset. 2004. A generic collaborative platform for multilingual lexical database development. In Gilles Sérasset, editor, *COLING 2004 Multilingual Linguistic Resources Workshop*, pages 73--79, Geneva, Switzerland, 28 August.

Gilles Sérasset. 2005. Multilingual legal terminology on the jibiki platform: The lexalp project. In Mathieu Lafourcade, editor, *Proc. of Papillon 2005 Workshop*, pages 64--73, Chiang Rai, Thailand, 11-13 December.

Mutsuko Tomokiyo, Mathieu Mangeot, and Emmanuel Planas. 2000. Papillon: a project of lexical database for english, french and japanese, using interlingual links. In *JST'00 Journées Science et Technologie*, page 3, National Olympic Memorial Youth Center, Tokyo, Japon, 13-14 novembre.

Staff vs. Volunteer: Two Approaches to Building a Multilingual Lexical Resource

Alan Melby, Marc Carmen, Steve Asher, and Gerard Meijssen

Brigham Young University, Brigham Young University, Mediphrase Project, and WiktionaryZ Project
Dept. of Linguistics and English Language, BYU campus, Provo, Utah 84602
Email: akmtrg@byu.edu, marc.carmen@byu.edu, steve@mediphrase.org, and gerard.meijssen@gmail.com

Abstract

Face-to-face encounters occur around the world on a daily basis between people who do not share a native language, and many of them deal with non-emergency medical needs. However, existing methods used to alleviate the frustration that results from these encounters are not always readily available, and they do not always meet the needs of the situation. One approach is a freely available online terminology database (termbase). Based on the success of Wikipedia, the authors propose that the use of a world-wide volunteer community could produce an accurate termbase. Medical Terminology (MediTerm) is a project that uses a similar community-based approach as Wikipedia but with the goal of alleviating linguistic frustrations within the domain of non-emergency medical needs. In possible collaboration with WiktionaryZ, a sister of Wikipedia, and Mediphrase, a project with goals similar to MediTerm, along with possible assistance from professional translators and interpreters, MediTerm hopes to develop a comprehensive multilingual medical database of terms and phrases.

1. Introduction

In the world in which we live, multilingual face-to-face transactions occur daily. A tourist attempting to barter for a trinket, business executives striking deals, and immigrants trying to survive day to day, these events are occurring around the world in many languages. We will categorize all face-to-face transactions as micro-global events; as opposed to macro-global events like the Olympic Games. An important sub-section of micro-global events are those dealing with medical needs; like communication needed to find medical care when there is not a life threatening emergency or an attempt to find a pharmacy or communicate with the pharmacist while trying to find non-prescription medication. These medical micro-global events are especially abundant at any macro-global event but they also occur separately from macro-global events. No matter where these micro-global events originate, the language barrier is often very difficult to overcome.

To overcome the language barrier there are several options including, but not limited to, (1) a physically present human interpreter, (2) an over-the-phone interpreter, or (3) a multilingual lexical resource. Having access to a human interpreter by phone or in person is not an option for most travelers because of the cost involved. Moreover, most hospitals and primary care physicians already have resources, that often include an interpreter over-the-phone or in person, to help with this type of special situation. We are not proposing to replace the system currently in use by medical professionals; but rather, we are hoping to provide a resource that helps solve this type of communication frustration for the average traveler that needs help with simple medical needs. For example, a traveler who needs a specific non-prescription medication enters a pharmacy but cannot communicate their symptoms or the medication they want. Or a second example is when a traveler is in need of medical attention from a physician but they are nowhere near a hospital or clinic and need to describe their symptoms to the person who is helping them. Attempting to communicate with police officers, store and hotel clerks, and local residents can be very frustrating and sometimes nearly impossible. So what does a traveler in a situation like this do? When travelers do not have access to personal interpreters or an over-the-phone interpreter, where do they currently turn for help? They must turn to available lexical resources which traditionally takes the form of a phrasebook or a bilingual dictionary on paper. However, both of these lexical resources typically lack in two ways: 1) they can never contain all of the possible terms that a traveler may need and 2) the only way to update them is to buy a new book. Therefore, the most useful type of lexical resource for travelers will be one that is available on-demand and has a large number of multilingual entries that are also up-to-date. The rest of this paper will discuss the possibility of an online multilingual lexical resource and different approaches to the construction of such a lexical resource.

The basic design we propose for a multilingual lexical resource is a terminology database (termbase) with multiple access modes. The overall format of a termbase is a concept-related hierarchy. Instead of the typical dictionary approach where each entry contains a headword and one or more senses, each entry is based on just one sense/concept. Then each entry has one or more terms, often in different languages, which are used to designate the concept. How could a termbase provide help in situations of medical need? A termbase can be useful assuming that the information it contains is accurate and accessible. With the advent of the Internet, accessibility has become less of an issue; however, quantity and quality of the data is still an issue.

It is important to note that there exist other multilingual lexical resources. One of the most commonly known is Papillon (http://www.papillon-dictionary.org/Home.po). Papillon is a freely available and accessible lexical database that includes English, French, Japanese, Lao, Thai, and Chinese. The database consists of entire resources, or dictionaries, that can be contributed to by users or by editing existing entries and adding new entries. The primary access method for Papillon is via the web interface. However, there is a test interface for mobile phones and a user can use a dictionary client like OmniDictionary (Papillon, 2005; OmniDictionary, 2006). Papillon is just one example of the resources that are currently available via the internet.

Although there are existing online lexical resources the question truly is, how is the best way to create an online multilingual lexical resource? In the following sections this paper will discuss the standard terminologist-centric approach and the wiki approach.

2. Standard Staff Approach

The standard approach to creating a termbase is taken on by one or more paid terminologists on the staff of an organization and consists of four steps: (1) organization of the domains, (2) monolingual corpus construction and analysis by hand or via computer, (3) construction of the termbase, and (4) population of multilingual terms to their respective concept entries (Canadian Translation Bureau, 2006).

2.1. Organization of Domains

According to the Translation Bureau of the Canadian government, domain classification is necessary to help organize terminology research. The domain names may be adopted from an existing system or created as needed by the terminologist (Subject Field, 2006). The first step a terminologist will take is to determine a tentative hierarchy of domains. By doing this, they narrow their focus as they prepare to gather terms to include in the termbase.

2.2. Corpus Construction and Analysis

An important part of the termbase creation is corpus construction. Terminologists will gather together texts that relate to the domain hierarchy they have organized. Then they analyze these texts to determine which terms pertain to their domains and should be included in the termbase. The analysis process can be done several ways. First, the terminologist could read through every document of the corpus and highlight terms they feel should be included. Second, they could use computer software to parse the texts and determine words and phrases that should be included. Unfortunately most term mining software is still not very accurate; and as a result, the terminologist does a lot of hand selection; but, the computer often decreases the total amount of work required by the terminologist.

2.3. Termbase Construction

Once the terminologist has determined words that will be included in the database, they begin database construction. Database construction can take on several media such as flat file, XML file, or using specific software such as MultiTerm or SDLX. No matter the manner of creation, the same terminological information can be stored. Each concept entry has a domain to denote its specialization and it contains at least one, and most likely more, language sections that specify information about the concept in that language. Then the term that designates the concept in that language is found within each language section.

2.4. Multilingual Population

After terminologists have entered all of the terms in the source language they will begin to enter in language data for each concept. Depending on the breadth of the project and the number of languages needed, the terminologists may do it themselves or they may send it to specialists of the desired language and domain. The end result is that each concept entry has a domain, concept definition, and one or more language sets that contain language specific information. More information can be stored for each concept but the domain, language, and term are considered to be the bare minimum.

3. Wiki Approach

In recent years a new phenomenon has emerged on the World Wide Web. Wiki sites have sprung up in many forms such as Wikipedia, Wiktionary, and Wikiquote. These wiki sites are all part of a larger project run by the Wikimedia Foundation., which is an umbrella organization. Wikipedia (http://www.wikipedia.org/), the encyclopedia branch of the Wikimedia Foundation, currently has over 936,334 articles in English, over 346,016 articles in German, and over 229,183 articles in French. These statistics are from just three of the 201 languages that have at least one article in the Wikipedia database (List of Wikipedias, 2006). Where has all of the information been gleaned from? Information has been contributed from around the world by users in many languages.

Note that the Wiktionary project has not produced termbases. Each Wiktionary is basically monolingual and headword oriented. Translations into another language are not tied to one word sense but rather to the headword.

We will now discuss the same four steps of termbase creation but in terms of a wiki approach.

3.1. Organization of Domains

At the creation of a community-based effort there are two options for the domain organization: (1) the first administrators of the project can determine a specific domain or set of domains or (2) a more flexible approach can be taken which allows for the creation of domains as information is submitted. In both cases, the growth and population of the domain hierarchy are dependent on community submissions. After a submission is made, one of the volunteer administrators will decide if there is an existing domain for the submission or if a new one should be created. A new domain must find a place in the hierarchy of domains. A term is not directly linked to a domain. Instead, a concept entry is linked to a domain.

3.2. Corpus Construction and Analysis

The typical submission for a community-based lexical resource will be based on personal experience which can be considered an element of a dynamic corpus. As users come across terms that they need to communicate in a foreign language they will contribute the term. To expand an existing concept entry, users and administrators will use the concept definition and context sentence to assign new translations. However, similarly to the traditional approach a corpus could be used to populate the database. For example, a collection of emails, transcripts, or news stories could be used as a corpus and using the same process described in the standard approach the user can contribute terms from the corpus.

3.3. Termbase Construction

A community-based termbase is in constant construction from its inception, and could be created using an open source XML application, for example TermBase Exchange (TBX-http://www.lisa.org/tbx), a relational database, or a hybrid approach. The designer should work directly with terminologists to determine what needs to be included in the termbase and how to best represent that information.

Once the structure of the termbase is determined then the initial monolingual information will be entered into the database. One important factor to note is that all contributors will be assigned a level of expertise according to their background and participation in the community. Any user at any level of expertise will be allowed to create new entries and update existing entries. Any changes that a user makes will be tied to their user information which will allow a searcher to specify that they only want results that have been entered by someone with a certain level of expertise.

3.4. Multilingual Population

Once the initial concepts have been entered then users will continue contributing by 1) creating more entries, 2) editing existing entries, and 3) enhancing existing entries by adding multimedia, including images and recordings, and more importantly by adding terms in other languages. All users, no matter their level of expertise, will be allowed to enhance existing entries and when this is done administrators will be notified. By allowing all users to contribute freely the website enables a constant flow of information. But notifying administrators also allows for a level of verification and validation. This combined effort can produce large amounts of accurate information and it will allow users to specify exactly what results they are looking for.

4. Methodology

The traditional approach to building lexical resources requires a paid staff and follows the steps described above. However, in the technical era in which we live people can communicate, share ideas, and work together on projects with ease. The real question is whether a volunteer community-administered approach, hereafter called the wiki approach, is viable? The following analysis will attempt to determine whether or not a valid, i.e. accurate, lexical resource can be created using a wiki approach, by analyzing three questions. (1) Can the wiki approach avoid chaos? (2) Even if the approach can avoid chaos in principle, has there been a real life example? (3) Can the wiki approach deal with intentional abuse? If, during the analysis, the wiki approach can provide a positive response to all three of these questions then the wiki approach will have been shown to be a viable alternative to the traditional approach.

5. Analysis

We will now discuss the three stated questions.

5.1. Avoiding Chaos

Over the years many collaborative efforts have been started and have quickly ended in chaos, that is, inconsistent and inaccurate information. So why is the wiki approach different? Unlike other collaborative projects, the wiki framework provides for different levels of administration that help to prevent chaos.

In a recent interview on the television program Q&A, which aired on C-SPAN (Lamb, 2005), Jimmy Wales, the founder of Wikipedia and the Wikimedia Foundation, was asked about wiki administration. He explained that there are various levels of administrators that have the ability to “lock articles and block IP numbers”. The administrators can also specifically block a user “from editing temporarily”. These administrators are volunteers from the Wikipedia community that are elected. The administrators have “special powers to enforce good behavior” and for issues that can’t be resolved by the administrators there is a “judicial committee.” (About Wikipedia, 2006) Through the framework that has been setup it appears that Wikipedia can theoretically avoid chaos.

5.2. Has Wikipedia Avoided Chaos

In principle, Wikipedia can handle the burden of ensuring accurate information, but what about real life. How can Wikipedia ever hope to gain the same level of quality as commercial encyclopedias? How can the community contribute a quality and quantity of information comparable to that produced by a group of paid research professionals?

Recently a scientific journal prepared a peer reviewed study to determine the answers to these questions. The study was conducted by Nature, and used science experts to analyze fifty articles from Wikipedia and Encyclopedia Britannica in several scientific areas (Giles, 2005). The experts were not informed of the origin of the article, and they were asked to look for major errors, including “factual errors, critical omissions, and misleading statements”, in the articles (Nature Supplementary, 2005). The results from the study were actually different than what many scholars would expect. Using 42 reviews that were returned, the average number of errors in the Wikipedia articles was four and in the articles from Encyclopedia Britannica there was an average of three errors. How can a group that has such varied membership as the wiki community produce a work that is nearly as accurate as its commercial counterpart? The answer is a simple but common cliché, “Two heads are better than one.” In the case of Wikipedia 867,692 heads (contributors) and 807 administrators are greater than a small group of researchers and editors (Statistics, 2006). So not only can Wikipedia avoid chaos in theory, but there is real life proof to show that they can avoid chaos and provide accurate information.

5.3. What About Intentional Abuse

Because wiki projects are open for contribution by anyone, over the past few years there have been several cases of intentional abuse. To some this is a pock mark for the wiki effort. However, the wiki framework has organized methods to deal with this type of problem. In fact, recently it was reported that members of the United States Congress and their staffers were repeatedly editing politician's biographies to include only information approved by the politician (Lehmann, 2006). Detractors of the wiki approach immediately jumped on the opportunity to discredit the online encyclopedia and its information. However, they did not look at whole story. In response to the abuse, member administrators have temporarily blocked all offending U.S. Congress Internet addresses. Moreover, there is a request for comment that has been created and is currently under discussion within the community for a longer penalty including a long-term block (Congress RFC, 2006). Because of its strong community and the framework that has existed since the beginning the Wikipedia community has proven that it has the ability to deal with cases of intentional abuse.

5.4. Analysis Result

Wikipedia, an example wiki project, has proven that it can avoid chaos in principle and in real life. The community has also recently proved that it can deal with intentional abuse. Because these three points have been satisfactorily addressed, in the case of Wikipedia, it is reasonable to assume that wiki-style termbase development projects are also viable.

6. Conclusion

There are many traditionalists that may never accept the idea of a wiki approach. However, since its inception in 2001, Wikipedia has been winning converts and has proven that a wiki approach can be successfully used to create accurate resources.

Although Wikipedia is not a lexical resource, a lexical resource built using the wiki approach and benefiting from the success of Wikipedia will likely result in a comprehensive useful termbase.

In fact, there are several projects currently being developed using a wiki approach. The GEvTerm Initiative (http://www.gevterm.org) began software development in 2005 to create a general purpose multilingual termbase. Recently several projects within the initiative have received special attention.

Medical Terminology (MediTerm) is a GEvTerm project to make available a multilingual termbase with the goal of assisting in communication in non-emergency medical situations such as attempting to find a specific non-prescription medication or attempting to get access to professional help by describing symptoms to a non-professional such as a hotel clerk or police officer. The hope of the authors is that MediTerm can help those individual travelers with some of the more trivial tasks that are important as they travel in foreign countries and need medical attention.

7. Future Work

Despite the progress of the GEvTerm initiative and the MediTerm project there is still a lot of work to be done. Collaboration discussions are underway with the Mediphrase project (http://www.mediphrase.org), represented by a co-author. Mediphrase "is an advanced multidisciplinary project to develop a real-time medical translation system for use by doctors and other healthcare workers during examinations and other live medical situations." (Mediphrase, 2004) Similarly to the goals of the MediTerm project, project Rosetta, part of Mediphrase, is attempting to "collate, codify and translate a substantial part of the whole body of medical diagnostic Q&A" or create a lexical resource containing useful medical terminology. Also the GEvTerm Initiative and its projects are looking for momentum and support by collaborating with Wikimedia Foundation's WiktionaryZ project (http://www.wiktionaryz.org), also represented by a co-author and possible endorsement from translator and interpreter associations.

The relationship among MediTerm, Mediphrase, and WiktionaryZ is not competitive. Mediphrase intends to represent terminology that would be used in a hospital, an emergency room, or a primary care facility, and they are hoping to gain an audience of professional care givers including physicians, emergency room personnel, and other hospital and clinic personnel. An example of the types of situations that Mediphrase hopes to help with is occasions when there is no human interpreter available in a hospital or clinic, or the human interpreter available is not efficient for the specific situation. For example, a girl was taken to an emergency room for abdominal pain, and after being asked about being sexually active she responded that she had been. However, her interpreter, who was her father, responded to the physician that she had not been sexually active, because in their culture that behavior was inappropriate for someone single and her age. The girl later returned to the emergency room to be treated for the complications of an ectopic pregnancy (Boschert, 2004). In another instance, the patient spoke Fukienese and the interpreter spoke Mandarin. As a result, they were attempting to communicate in Cantonese. The patient resisted inpatient care after several invitations, but once she had an interpreter who spoke Fukienese the inpatient care was explained completely and she was admitted without further resistance (Gussman, 2002).

On the other hand, MediTerm is meant to be a resource for travelers that are seeking minor medical help; such as, searching for a specific non-prescription medicine or travelers that need to describe symptoms in non-emergency medical situations. This will be especially useful to the traveler suffering from a case of traveler's diarrhea, headaches, or many of the other common ailments that travelers often suffer from.

The last of the projects is WiktionaryZ; it is the next generation of the Wiktionary projects. Like the Wiktionary projects it aims to include information on expressions from all languages. It is however a departure from the first generation as it is based on relational data. WiktionaryZ includes relations, terminology and lexicology, the content is structured and multilingual, the user interface will include the labels that indicate particular content and almost as importantly there will be only one database. It is however still a Wiki; to be considered as such it must allow for "talk pages", changes made by identified or anonymous contributors. The challenge will be to maintain quality; one scheme that is considered is a so called "regime". The idea is that for specific types of information people can submit to a protocol that is there to give extra confidence about the validity of what is said; for instance a relation like "Substance treats Disease" is a dangerous assertion. A regime could be anything like a requirement for documentation to prove the point to a procedure with an external organization. Given the growing "Open Access" movement, an increasing amount of medical information and thus terminology will become available under compatible licenses. We are presently preparing the inclusion of data from the NLM Unified Medical Language System (UMLS) meta-thesaurus. Given our requirement but also the necessity for attribution, it will be apparent what can be trusted to be true, and updated thesauri will still be a valuable asset. Organizations responsible for thesauri may draw from the WiktionaryZ and lift information to a 'validated' level.

Although we use the very challenging field of Biomedicine as an experimental setting, WiktionaryZ can be used for any given discipline.

As apparent by the combination of authors of this paper, representatives from each of these projects are working together. And although each project has a different intended audience and domain, the three projects hope to co-exist and work together. For example, a user could come to MediTerm and search for a term, but they may not be satisfied with the results they get because they may not be specific enough. So then MediTerm could also link the traveler to Mediphrase for more specific information on emergency medical needs. However, if the results that MediTerm provides are too specific then the traveler will also be able to link to WiktionaryZ which will have a much large selection of domains. One important factor that has not been discussed about this relationship is how it will occur. Data exchange is one of the most important aspects of this relationship because without an efficient way of passing information between each of the projects many users will often become frustrated with the speed and efficiency of the search they perform. For this reason, each of the projects is attempting to implement the same exchange format within their individual projects. The exchange format chosen is called TermBase Exchange or TBX (http://www.lisa.org/tbx). TBX "is an open XML-based standard format for terminological data." (TBX, 2005) By using TBX each of the projects will be able to exchange data that is submitted and gathered by its users. As a result, each project will become richer terminologically, which would not be as easily accomplished with different exchange formats and less cooperation.

As mentioned previously the authors are hoping for an endorsement from professional associations so that translators and interpreters from around the world will be encouraged to contribute information to the MediTerm database. Moreover, the authors are currently in negotiations with LA Care, which "is a non-profit, community accountable health maintenance organization (HMO) that serves more than 750,000 Los Angeles County residents". (LA Care, 2005) LA Care has compiled medical glossaries in English, Spanish, and Armenian. The MediTerm project hopes to use the information compiled by LA Care in the MediTerm database. The authors also plan on contacting Papillon to compare underlying data structures and discuss the possibility of data exchange. The most important source of data that will be compiled is from users around the world that will contribute information.

Although data collection is very important to any project like MediTerm, it is not the only goal. In fact, one of the most important goals that MediTerm has is to make the data accessible and useful. For this reason, there are currently several ways of accessing the information: (1) on desktop computers in hotel lobbies or locations that have an Internet connection for laptop access, (2) on wireless enabled handheld devices, like cellular phones and PDAs, for mobile access, and (3) printed copies for travelers that do not have wireless enabled devices. Not only is the mode of transmission important for MediTerm but also the quality of the information. The software for MediTerm and GEvTerm is currently under development and one of the emphases is to include more than just textual data including context sentences, example sentences, sound recordings, phonetic transcriptions, and images. By providing all of this information MediTerm hopes to help increase the effectiveness with which travelers are able to communicate as they run into the medical needs that are all too common during travels in a foreign country.

8. References

About Wikipedia. (2006). Wikipedia: About. http://en.wikipedia.org/wiki/Wikipedia:About#Who_writes_Wikipedia.3F. (Accessed on February 14, 2005)

Boschert, Sherry. (2004). Language Barriers a Problem in Emergency Room. *Internal Medicine News*. 37:82.

Canadian Translation Bureau. (2005). Methodology for Creating Terminology Records. http://www.termium.gc.ca/didacticiel_tutorial/english/lesson3/index_e.html. (Accessed on February 14, 2006).

Congress RFC. (2006). Wikipedia:Requests for Comment/United States Congress. http://en.wikipedia.org/wiki/Wikipedia:Requests_for_comment/United_States_Congress (Accessed on January 30, 2006)

Giles, Jim. (2005). Internet Encyclopaedias Go Head to Head. *Nature*. 438:900-901

Gussman, Debra. (2002). From Fingersticks to Chopsticks. *Medical Economics*. 11:61

LA Care. (2005). LA Care: About L.A. Care. http://www.lacare.org/opencms/opencms/en/about/index.html. (Accessed on April 4, 2006);

Lamb, Brian. (2005). Jimmy Wales Wikipedia Founder. Q&A. National Cable Satellite Corporation. Aired on September 25, 2005.

Lehmann, Evan. (2006). Rewriting History Under the Dome. Lowell Sun. http://www.lowellsun.com/ci_3444567. (Accessed on February 2, 2006).

List of Wikipedias. (2006). http://meta.wikimedia.org/wiki/List_of_Wikipedias. (Accessed on January 24, 2006).

Mediphrase. (2004). Mediphrase Project. http://mediphrase.org/. (Accessed on March 30, 2006).

Nature Supplementary. (2005). Supplementary information to accompany Nature news article "Internet Encyclopaedias Go Head to Head" http://www.nature.com/nature/journal/v438/n7070/extref/438900a-s1.doc. (Accessed on February 1, 2006).

OmniDictionary. (2006). OmniDictionary 2. http://www.omnigroup.com/applications/omnidictionary/. (Accessed on April 4, 2006).

Papillon. (2003). Information on Papillon Web Server. http://www.papillon-dictionary.org/Information.po. (Accessed on April 4, 2006).

Statistics. (2006). http://en.wikipedia.org/wiki/Special:Statistics. (Accessed on February 2, 2006).

Subject Field. (2005). Delimiting and Structuring the Subject Field. http://www.termium.gc.ca/didacticiel_tutorial/english/lesson3/page3_2_2_e.html. (Accessed on January 26, 2006).

TBX. (2005). TermBase Exchange. http://www.lisa.org/standards/tbx/. (Accessed on April 4, 2006).

9. Acknowledgements

All of the software development for the MediTerm project and its parent the GEvTerm initiative has been funded by the Translation Research Group at Brigham Young University and the Linguistics Department at Brigham Young University.

10. Appendix A: Software Development

This appendix will discuss some of the technical aspects of the data structure and design of MediTerm. The MediTerm site has been designed and developed using several technologies including PostgreSQL, PHP, and XML. Because of the choice of database and its basic design, which doesn't utilize any PostgreSQL proprietary features, the database design could easily be adapted within any other relational database by someone with adequate knowledge. XML is an integral part of the overall design. In fact it takes on several roles including exchange of terminological data via TBX and information exchange via web services using SOAP. SOAP (http://www.w3.org/TR/soap/) is an XML protocol that facilitates the exchange of information between applications over the web and is commonly used in web services. TBX (http://lisa.org/standards/tbx/) is an XML format for terminological data that can be used in conjunction with SOAP. All of these technologies are open technologies that are freely available which facilitates development with other projects and easier adaptation of the MediTerm data structure and design.

The abstract data model used for MediTerm is taken from TMF, which is ISO standard 16642. Figure 1 provides a simplified representation of the formal data model in TMF:

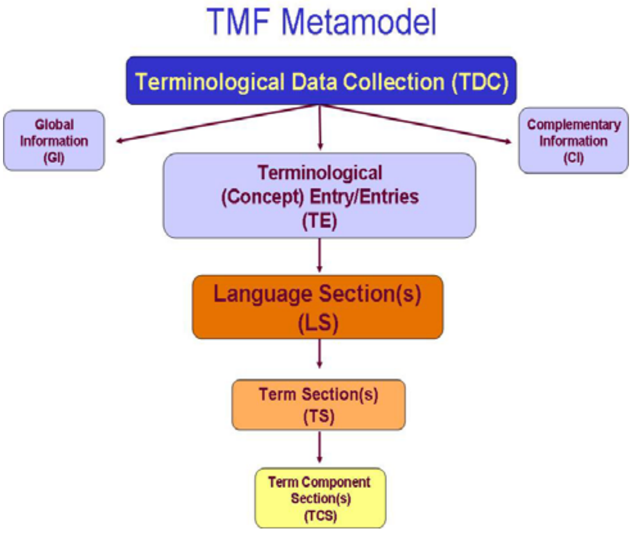

TMF Metamodel
Terminological Data Collection (TDC)
Global Information (GI)
Complementary Information (CI)
Terminological (Concept) Entry/Entries (TE)
Language Section(s) (LS)
Term Section(s) (TS)
Term Component Section(s) (TCS)

Figure 1: TMF Metamodel

The diagram shows that a termbase (called a Terminological Data Collection in TMF), consists of global information (GI, i.e., meta-metadata) about the termbase, such as who maintains it and who can use it under what conditions, complementary information (CI), such as a set of full bibliographic references that can be referenced from entries and a hierarchy of domains that entries can be linked to, and a set of terminological concept entries. Each entry (a TE), documents one concept and consists of one or more language sections (LS) (potentially hundreds in MediTerm and other multilingual termbases), and each language section consists of one or more term sections (TS). Each term section contains a term in the object language of the language section that designates the concept of the concept entry. A term section can also include information about a term, such as a sentence using the term and pronunciations pre-recorded by native speakers. Links are made between concept entries and from concept entries to complementary information. The term component section (TCS) provides an optional refinement that can be used to document information such as gender, inflectional form, etc., associated with the individual words making up multiword terms.

LMF for multilingual, specialized lexicons

Gil Francopoulo[1], Monte George[2], Nicoletta Calzolari[3],
Monica Monachini[4], Nuria Bel[5], Mandy Pet[6], Claudia Soria[7]

[1]INRIA-Loria: gil.francopoulo@wanadoo.fr
[2]ANSI: dracalpha@earthlink.net
[3]CNR-ILC: glottolo@ilc.cnr.it
[4]CNR-ILC: monica.monachini@ilc.cnr.it
[5]UPF: nuria.bel@upf.edu
[6]MITRE: mpet@mitre.org
[7]CNR-ILC: claudia.soria@ilc.cnr.it

Abstract

Optimizing the production, maintenance and extension of lexical resources is one the crucial aspects impacting Natural Language Processing (NLP). A second aspect involves optimizing the process leading to their integration in applications. With this respect, we believe that the production of a consensual specification on lexicons can be a useful aid for the various NLP actors. Within ISO, the purpose of LMF (ISO-24613) is to define a standard for lexicons that covers multilingual and specialized data.

1. Introduction

Lexical Markup Framework (LMF) is a model that provides a common standardized framework for the construction of Natural Language Processing (NLP) lexicons. The goals of LMF are to provide a common model for the creation and use of lexical resources, to manage the exchange of data between and among these resources, and to enable the merging of a large number of individual electronic resources to form extensive global electronic resources.

Types of individual instantiations of LMF can include monolingual, bilingual or multilingual lexical resources. The same specifications are to be used for both small and large lexicons. The description range from morphology, syntax, semantic to translation information organized as different extensions of an obligatory core package. The model is being developed to cover all natural languages. The range of targeted NLP applications is not restricted. LMF is also used to model machine readable dictionaries (MRD), which are not within the scope of this paper.

2. History and current context

In the past, this subject has been studied and developed by a series of projects like GENELEX [Antoni-Lay], EAGLES, MULTEXT, PAROLE, SIMPLE , ISLE and MILE [Bertagna]. More recently within ISO[1] the standard for terminology management has been successfully elaborated by the sub-committee ISO-TC37 and published under the name "Terminology Markup Framework" (TMF) with the ISO-16642 reference. Afterwards, the ISO-TC37 National delegations decided to address standards dedicated to NLP. These standards are currently elaborated as high level specifications and deal with word segmentation (ISO 24614), annotations (ISO 24611, 24612 and 24615), feature structures (ISO 24610), and lexicons (ISO 24613) with this latest one being the focus of the current paper. These standards are based on low level specifications dedicated to constants, namely data categories (revision of ISO 12620), language codes (ISO 639), scripts codes (ISO 15924), country codes (ISO 3166), dates (ISO 8601) and Unicode (ISO 10646).

This work is in progress. The two level organization will form a coherent family of standards with the following simple rules:

1) **low level specifications** provide standardized constants;

2) **high level specifications** provide structural elements that are adorned by the standardized constants.

3. Scope and challenges

The task of designing a lexicon model that satisfies every user is not an easy task. But all the efforts are directed to elaborate a proposal that fits the major needs of most existing models.

In order to summarise the objectives, let's see what is in the scope and what is not.

LMF addresses the following difficult challenges:

1. Represent words in languages where multiple orthographies (native or transliterations) are possible, e.g. some Asian languages.
2. Represent the morphology of languages where a description in extension of all inflected forms is not manageable (e.g. Hungarian). In this case, representation in intension is the only manageable issue.
3. Easily associate written forms and spoken forms for all languages.
4. Represent complex compound words (like in German, Dutch among other languages)
5. Represent fixed, semi-fixed and flexible multiword expressions.
6. Represent specific syntactic behaviors (as recommended in Eagles).
7. Allow complex argument mapping between syntactic and semantic descriptions (as recommended in Eagles).
8. Allow a semantic organization based on SynSets (like in WordNet) or on semantic predicates (like in FrameNet).

[1] www.iso.org

9. Represent large scale multilingual resources based on interlingual pivots or on transfer linking.

LMF does not address the following topics:

1. General sentence grammar of a language
2. World knowledge representation

In other terms, LMF is mainly focused on lexical linguistic information representation.

4. Key standards used by LMF

LMF utilizes Unicode in order to represent the scripts and orthographies used in lexical entries regardless of language.

Linguistic constants, like /feminine/ or /transitive/, are not defined within LMF but are specified in the Data Category Registry (DCR) that is maintained as a global resource by ISO TC37 in compliance with ISO/IEC 11179-3:2003.

The LMF specification complies with the modeling principles of Unified Modeling Language (UML) as defined by OMG[2] [Rumbaugh]. A model is specified by a UML class diagram within a UML package: the class name is not underlined. The various examples of word description are represented by UML instance diagrams: the class name is underlined.

5. Structure and core package

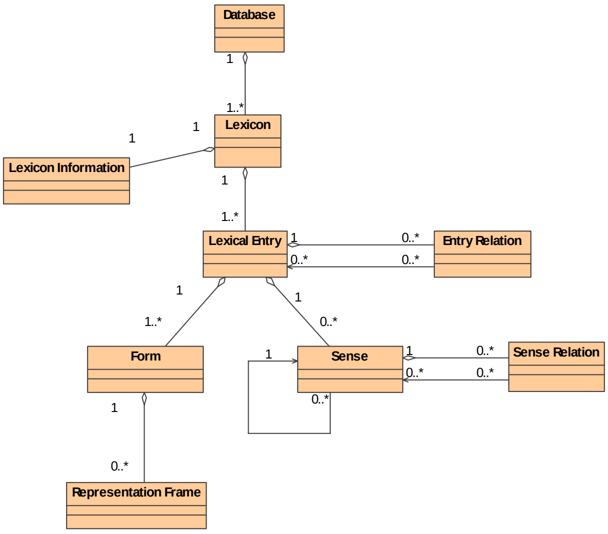
LMF is comprised of two components:

1) **The core package** which is the structural skeleton which describes the basic hierarchy of information in a lexical entry.

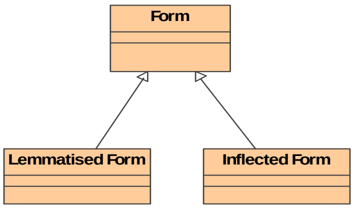
2) **Extensions to the core package**, which are expressed in a framework that describes the re-use of the core components in conjunction with these additional components required for the description of the contents of a specific lexical resource.

In the core package, one class called *Database* represents the entire resource and is a container for one or more lexicons. The *Lexicon* class is the container for all the lexical entries of the same language within the database. The *Lexicon Information* class contains administrative information and other general attributes. The *Lexical Entry* class is a container for managing the top level language components. As a consequence, the number of representatives of single words, multiword expressions and affixes of the lexicon is equal to the number of lexical entries in a given lexicon. The *Form* and *Sense* classes are parts of the *Lexical Entry*. Form consists of a text string that represents the word. *Sense* specifies or identifies the meaning and context of the related form. Therefore, the *Lexical Entry* manages the relationship between sets of related forms and their senses. If there is more than one orthography for the word form (e.g. transliteration) the *Form* class may be associated with one to many *Representation Frames*, each of which contains a specific orthography and one to many data categories that describe the attributes of that orthography.

The core package classes are linked by the relations as defined in the following UML class diagram:



Form class can be subclassed into *Lemmatised Form* and *Inflected Form* class as follows:



A subset of the core package classes are extended to cover different kinds of linguistic data. All extensions conform to the LMF core package and cannot be used to represent lexical data independently of the core package. From the point of view of UML, an extension is a UML package. Current extensions for NLP dictionaries are: NLP Morphology, NLP inflectional paradigm, NLP Multiword Expression pattern, NLP Syntax, NLP Semantic and Multilingual notations, which is the focus of this paper. Extensions for Morphology, Syntax and Semantic extensions are described in [Francopoulo]. All extensions are described in [LMF 2006].

6. NLP Multilingual extension

The NLP multilingual notation extension is dedicated to the description of the mapping between two or more languages in a LMF database. The model is based on the notion of *Axis* that links the notions of *Sense*, *Syntactic Behavior* and *Example* pertaining to different languages. "Axis" is a term taken from the Papillon project[3] [Sérasset]. Axis can be organized at the lexicon manager convenience in order to link directly or indirectly objects of different languages.

6.1. Considerations for standardizing multilingual data

The simplest configuration of multilingual data is a bilingual lexicon where a single link is used to represent

[2] www.omg.org

[3] www.papillon-dictionary.org

the translation of a given form/sense pair from one language into another. But a survey of actual practices clearly reveals other requirements that make the model more complex. Consequently, LMF has focused on the following ones:

(i) Cases where the relation 1-to-1 is impossible because of lexical differences among languages. An example is the case of English word “river” that relates to French words “rivière” and “fleuve”, where this last one is used for specifying that the referent is a river that flows into the sea. The bilingual lexicon should specify how these units relate.

(ii) The bilingual lexicon approach should be optimized to allow the easiest management of large databases for real multilingual scenarios. In order to reduce the explosion of links in a multibilingual scenario, translation equivalence can be managed through an intermediate "Axis". This object can be shared in order to contain the number of links in manageable proportions.

(iii) The model should cover both *transfer* and *pivot* approaches to translation, taking also into account hybrid approaches. In LMF, the pivot approach is implemented by a “Sense Axis”. The transfer approach is implemented by a “Transfer Axis”.

(iv) A situation that is not very easy to deal with is how to represent translations to languages that are similar. The problem arises for instance when the task is to represent translations from English to European Portuguese and Brazilian. The difference between the two last languages is not very important: a certain number of words are different and the syntax of pronouns is different. Instead of managing two distinct copies, it is more effective to distinguish variations through a limited number of specific Axis, the vast majority of Axis being shared.

(v) The model should allow for representing the information that restricts or conditions the translations. The representation of tests that combine logical operations upon syntactic and semantic features must be covered.

6.2. Structure

The model is based on the notion of Axis that link Senses, Syntactic Behavior and examples pertaining to different languages. Axis can be organized at the lexicon manager convenience in order to link directly or indirectly objects of different languages. A direct link is implemented by a single axis. An indirect link is implemented by several axis and one or several relations.

The model is based on three main classes: Sense Axis, Transfer Axis, Example Axis.

6.3. Sense Axis

Sense Axis is used to link closely related senses in different languages, under the same assumptions of the interlingual pivot approach, and, optionally, it can also be used to refer to one or several external knowledge representation systems.

The use of the Sense Axis facilitates the representation of the translation of words that do not necessarily have the same valence or morphological form in one language than in another. For example, in a language, we can have a single word that will be translated by a compound word into another language: English “wheelchair” to Spanish “silla de ruedas”. Sense Axis may have the following attributes: a label, the name of an external descriptive system, a reference to a specific node inside an external description.

6.4. Sense Axis Relation

Sense Axis Relation permits to describe the linking between two different Sense Axis. The element may have attributes like label, view, etc.

The label enables the coding of simple interlingual relations like the specialization of “fleuve” compared to “rivière” and “river”. It is not, however, the goal of this strategy to code a complex system for knowledge representation, which ideally should be structured as a complete coherent system designed specifically for that purpose.

6.5. Transfer Axis

Transfer Axis is designed to represent multilingual transfer approach. Here, linkage refers to information contained in syntax. For example, this approach enables the representation of syntactic actants involving inversion, such as (1):

(1) fra:“elle me manque” => eng:“I miss her”

Due to the fact that a lexical entry can be a support verb, it is possible to represent translations that start from a plain verb to a support verb like (2):

(2) fra:“Marie rêve” => jpn:"Marie wa yume wo miru"
(Mary dreams)

6.6. Transfer Axis Relation

Transfer Axis Relation links two Transfer Axis. The element may have attributes like: label, variation.

6.7. Source Test and Target Test

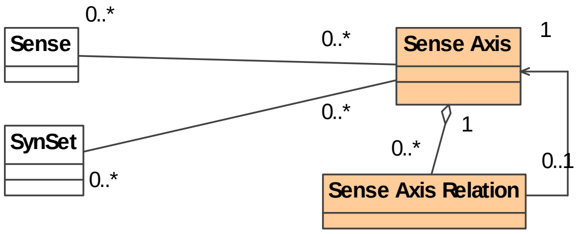
Source Test permits to express a condition on the translation on the source language side while Target Test does it on the target language side. Both elements may have attributes like: text and comment.

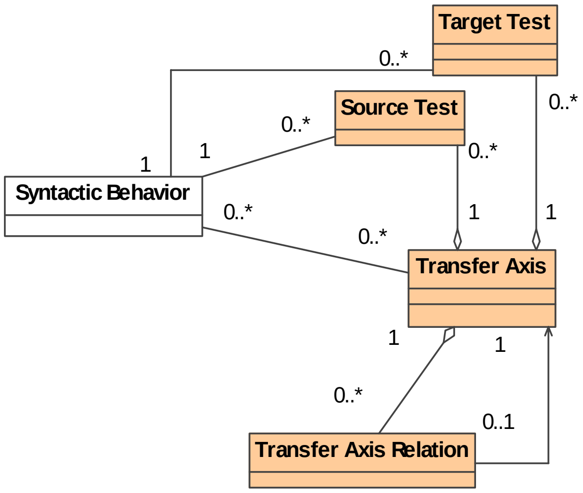
6.8. Example Axis

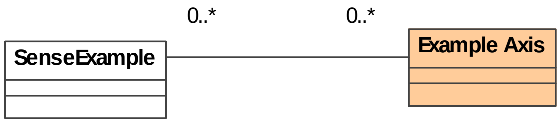
Example Axis supplies documentation for sample translations. The purpose is not to record large scale multilingual corpora. The goal is to link a Lexical Entry with a typical example of translation. The element may have attributes like: comment, source.

6.9. Class Model Diagram

The UML class model diagram for multilingual notations is as follows:

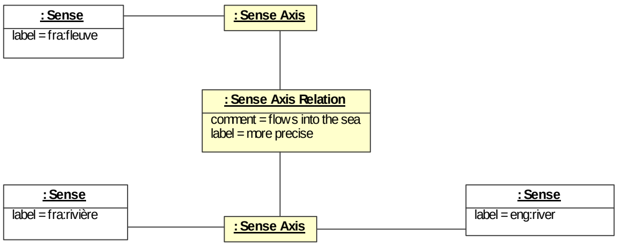




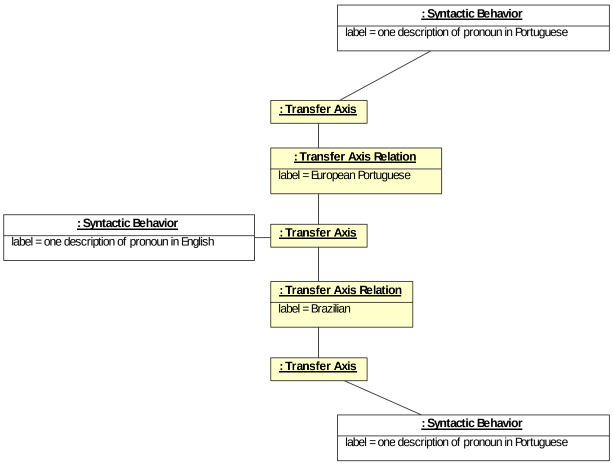


7. Three examples

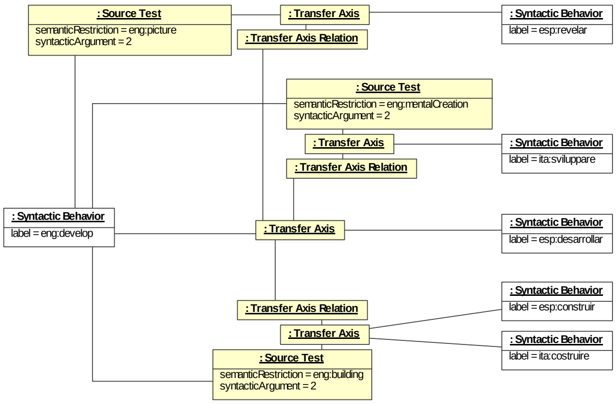
The first example is about the interlingual approach with two axis to represent a near match between "fleuve" in French and "river" in English. The axis on the top is not linked directly to any English sense because this notion does not exist in English. In the diagram, French is located on the left side and English on the right side.



Let's see now an example about the transfer approach about slight variations between similar languages. The example is about English on one side and European Portuguese and Brazilian on the other side. Due to the fact that these two last languages have a very similar syntax, but with some local exceptions, the goal is to avoid a full and dummy duplication in order to ease maintenance of both languages. The transfer axis relations hold a label to distinguish which axis to use depending on the target language.



A third example shows how to use the Transfer Axis relation to relate different information in a multilingual transfer lexicon. It represents the translation of the English “develop” into Italian and Spanish. Recall that the more general sense links “eng:develop” and “esp:desarrollar”. Both Spanish and Italian have restrictions that should be tested in the source language: if the second argument of the construction refers to certain elements (picture, mentalCreation, building) it should be translated into specific verbs.



8. LMF for specialized lexicons

LMF, that has not specially been conceived and tested on specialized lexicons, can be used for all kinds of lexicons included the specialized ones.

Compared to general NLP lexicons, specialized lexicons have the following properties:

1. High number of multiword expressions
2. High number of orthographic variants including abbreviations and acronyms
3. Inclusion of domain specific information: terminological definitions, particular codes (like in UMLS).
4. Domain (and sub-domain) marks are needed in the two following situations:
   - when the domain is subdivided into several subdomains
   - when the lexicon is a mix of general and specialized words.

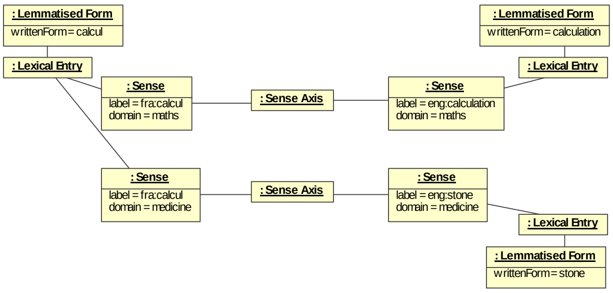
LMF offers for these cases different solutions which are mostly in line with the recommendations for general language lexica [LMF 2006].

The first case is for the encoding of multiword expressions which can be referred to as a unique element because of, for instance, translation equivalences. This is the case for Italian “cervello terminale” which must be translated into English as “cerebrum” and into Spanish as “encéfalo”.

The second case: variation can take the form of orthographic variation, as in the case of “gonadotropin” vs. “gonadotrophin”. But it can also be two entries linked by a synonym relation: take the case of the English medical terms “hypophysis” and “pituitary gland”.

Concerning the two last cases (i.e. domain specific information and domain marks), every LMF element can be adorned by an attribute/value pair. In a multilingual perspective, these marks can be used to condition a translation.

Let's see for instance, the translation of the French word "calcul" into English. There are two senses in French: one in Maths and the other one in Medicine. The translations into English give two different senses and two different lexical entries, as follows:



9. LMF in XML

During the last three years, the ISO group focused on the conceptual model by the mean of a UML specification. In the last version of the LMF document [LMF 2006] a DTD has been provided as an informative annex. Concerning UML to XML conversion, the following conventions are adopted:

1. each UML attribute is transcoded as a DC element
2. each UML class is transcoded as an XML element
3. UML aggregations are transcoded as content inclusion
4. UML shared associations (i.e. associations that are not aggregations) are transcoded as IDREF(S)

An example of entries is the following XML tag structure, where three senses are shown: a French entry "gonadotrophine" is linked both to a Spanish entry "gonadotrofina" and to an English entry "gonadotropin". The Spanish fragment shows two orthographic variants "gonadotrofina" and "gonadotropina". The English fragment shows also two variants.

```
<Database languageCode="ISO-639-2">
<!—                              French section -->
<Lexicon>
<LexiconInformation>
        <DC att="name" val=”French Extract”/>
        <DC att="language" val="fra"/>
</LexiconInformation>
<LexicalEntry
        <DC att=”partOfSpeech” val="noun"/>
        <LemmatisedForm>
                <DC att=”writtenForm” val="gonadotrophine"/>
        </LemmatisedForm>
        <Sense id="fra#gonadotrophine">
                <DC att="domain" val="medicine"/>
        <SemanticDefinition>
                <DC att=”text” val="Lycoprotéine d'un poids moléculaire
d'environ 43 000 daltons produite par le syncytiotrophoblaste"/>
                <DC att=”source” val="Wikipedia"/>
        </SemanticDefinition>
        </Sense>
</LexicalEntry>
</Lexicon>
<!—                              Spanish section -->
<LexiconInformation>
      <DC att="name" val=”Spanish Extract”/>
      <DC att="language" val="esp"/>
</LexiconInformation>
<LexicalEntry
      <DC att=”partOfSpeech” val="noun"/>
      <LemmatisedForm>
                <DC att=”writtenForm” val="gonadotrofina"/>
      </LemmatisedForm>
      <LemmatisedForm>
                <DC att=”writtenForm” val="gonadotropina"/>
      </LemmatisedForm>
      <Sense id="esp#gonadotrofina">
                <DC att="domain" val="medicine"/>
      <SemanticDefinition>
                <DC att=”text” val="Cada una de las hormonas secretadas
mayoritariamente por la hipófisis"/>
                <DC att=”source” val="UPF-Term"/>
       </SemanticDefinition>
       </Sense>
</LexicalEntry>
</Lexicon>
<!—                                Multilingual section -->
<SenseAxis id="A1" senses="fra#gonadotrophine esp#gonadotrofina
eng#gonadotropin">
</SenseAxis>
<!—-                         English section -->
<LexiconInformation>
      <DC att="name" val=”English Extract”/>
      <DC att="language" val="eng"/>
</LexiconInformation>
<LexicalEntry
      <DC att=”partOfSpeech” val="noun"/>
      <LemmatisedForm>
                <DC att=”writtenForm” val="gonadotropin"/>
      </LemmatisedForm>
      <LemmatisedForm>
                <DC att=”writtenForm” val="gonadotrophin"/>
      </LemmatisedForm>
      <Sense id="eng#gonadotropin">
                <DC att="domain"  val="medicine"/>
      <SemanticDefinition>
```

```
            <DC att=”text” val="a hormone (eg, follicle-stimulating
hormone) that acts on the gonads to promote their growth and
function"/>
            <DC att=”source” val=”www.aegis.com"/>
            <DC att=”UMLS code”  val=”E0030121” />
        </SemanticDefinition>
        </Sense>
</LexicalEntry>
</Lexicon> </Database>
```

10. Conclusion

In this paper we presented the results of the ongoing research activity of the LMF ISO standard. The design of a common and standardized framework for multilingual lexical databases will contribute to the optimization of the use of lexical resources, specially their reusability for different applications and tasks. Interoperability is the condition of a effective deployment of usable lexical resources.

In order to reach a consensus, the work done has paid attention to the similarities and differences of existing lexicons and the models behind them.

Acknowledgements

The work presented here is partially funded by the EU eContent-22236 LIRICS project[4], partially by the French TECHNOLANGUE[5] + OUTILEX[6] programs.

References

Antoni-Lay M-H., Francopoulo G., Zaysser L. 1994 A generic model for reusable lexicons: the GENELEX project. Literary and linguistic computing 9(1) 47-54

Bertagna F., Lenci A., Monachini M., Calzolari N. 2004 Content interoperability of lexical resources, open issues and MILE perspectives LREC Lisbon

Francopoulo G., George M., Calzolari N., Monachini M., Bel N., Pet M., Soria C. 2006 Lexical Markup Framework LREC Genoa

LMF 2006 Lexical Markup Framework ISO-CD24613-revision-9, ISO Geneva

Rumbaugh J., Jacobson I., Booch G. 2004 The unified modeling language reference manual, second edition, Addison Wesley

Sérasset G., Mangeot-Lerebours M. 2001 Papillon Lexical Database project: monolingual dictionaries & interlingual links NLPRS Tokyo

[4] http://lirics.loria.fr

[5] www.technolangue.net

[6] www.at-lci.com/outilex/outilex.html

Unification of multi-lingual scientific terminological resources using the ISO 16642 standard. The TermSciences initiative

Majid Khayari[1], Stéphane Schneider[1], Isabelle Kramer[2], Laurent Romary[2]

INIST-CNRS
2, Allée du Parc de Brabois- CS 10310
54519 Vandoeuvre-Lès-Nancy

LORIA, Universities NANCY, CNRS INRIA
Campus Scientifique -BP 239
54506 Vandoeuvre-Lès-Nancy-Cedex

majid.khayari@inist.fr, stephane.schneider@inist.fr, isabelle.kramer@loria.fr, laurent.romary@loria.fr

Abstract

The TermSciences initiative aims at building a multi-purpose and multi-lingual knowledge system from different source vocabularies produced by major French research institutions and which were initially intended to be used for indexing and cataloguing scientific literature. Since the construction of language resource repositories is cost-effective and time-consuming, the producers of these vocabularies wished to both share their terminological material and develop common tools for the collaborative management of the integrated resource. Sharing terminologies poses some problems because of the heterogeneous nature of the source data (i.e., coverage, granularity and compositionality of concepts, etc.), and to the discrepancy between partner needs (i.e., simple diffusion of the terminological material, use of the shared material to enhance information engineering tasks, etc.). This paper presents the TermSciences portal[1], which deals with the implementation of a conceptual model that uses the recent ISO 16642 standard (Terminological Markup Framework). This standard turned out to be suitable for concept modeling since it allowed for organizing the original resources by concepts and to associate the various terms for a given concept. Additional structuring is produced by sharing conceptual relationships, that is, cross-linking of resource results through the introduction of semantic relations which may have initially be missing. A special emphasis is put on medical resources used in this project, i.e. the French translation by the Institut National de la Santé et de la Recherche Médicale (INSERM) of the MeSH thesaurus from the US National Library of Medicine, the public health thesaurus of the Banque de Données de Santé Publique (BDSP) and the dictionary of human and mammals reproduction biotechnology of the Institut National de la Recherche Agronomique (INRA).

1. INTRODUCTION

The development of Communication and Information Technologies, and in particular, in the field of natural language resources, including terminology, raises the crucial question of standardization. Since the construction of language resource repositories is cost-effective and time-consuming, the producers and users of specialized vocabularies may benefit from sharing their resources. Still, sharing resources implies to agree about common formats and data models. This paper presents the TermSciences initiative whose purpose is to build a common terminological reference database (Bourigault and Condamines, 1995) from terminological resources (lexicons, dictionaries, thesauri) produced and maintained by various French public research institutions. As such, it is the first public initiative to implement the recently adopted Terminological Markup Framework (TMF, ISO 16642). TMF aims at providing a platform for the interchange of computerized lexical data, as used in many kinds of applications.

In this context, an important issue is to provide a uniform way of representing such databases considering the heterogeneity of both their formats and their descriptors. This is an essential aspect of natural language processing since it allows for both reusing linguistic data such as lexicons or grammars and deploying interoperable linguistic components in complex processing lines. The TermSciences project allowed us to validate step by step different stages related to the deployment of such an infrastructure, within the context of a concrete implementation of the TMF methodology and principles: modelling (ISO 704), import, fusion, update and export of data, and modification of the model.

2. REQUIREMENTS

2.1. Need of conceptualization

A major obstacle to the sharing of terminologies is the lack of conceptual integration of terms (Gangemi and al, 1998). Since the meaning of terms may be different according to the domain in which they appear (Wüster, 1976) and to the context of use (Rastier, 1995), any successful integration relies on a conceptualization process. However, most terminologies used in this project were built according to a term-centred (i.e. a descriptor-oriented) model (Condamines, 1994). This means that the linking of terms to concepts implies firstly to find or define some abstract high level terminologies (list of concepts) or ontologies and then to clear and consensual definition of concepts, i.e. if multiple terms (synonyms) may refer to the same object, a concept is unique for a given object and there is no place for an alternate or

[1] www.termsciences.fr

complimentary concept related to the same object (Baud et al, 1998).

2.2. Documenting meta data

A major issue of the TermSciences initiative is the management of the integrated terminological database. Because the common database is being built from resources managed independently by different institutions, the conceptual model includes meta data about the sources of each element composing a terminological entry. Additionally, every native resource file is formatted in the target format and stored as is. The use of pointing mechanisms based on “xml:id” and the XPointer syntax make it possible to reach any native record in these formatted files and capture new elements such us updates made lately by the producer of a given resource.

2.3. Collaborative Update

The management of the terminological content is planned to be taken in charge by collaborators that are involved in terminological works and by others who are indexers dealing more with indexing vocabularies (i.e. artificial languages) than with terminologies. This implies that staff education is a pre-requisite to the advancement of this project. The essential difference between words and concepts, the notion of synonymy, which applies, to the first but not to the second, and the need of a natural “compositionality” of terms represent the main distinctions to be made.

3. TMF

The representation using TMF can be summarized as the description of computerized terminological data representation languages; it is based on two components: a meta-model, i.e. the underlying structural skeleton and a description of constraints of attachment of some information to the structural model, i.e. data categories as described in the ISO 12620 standard.

3.1. TMF metamodel

A meta-model does not describe one specific format, but acts as a kind of high level mechanism based on the following elementary notions: structure, information, and methodology. The structuring elements of the meta-model are called “components” and they may be “decorated” with information units, called Data Categories. A meta-model should also comprise a flexible specification platform for elementary units. This specification platform should be coupled to a reference set of descriptors that should be used to parameterize specific applications dealing with content. The terminological meta-model is based on guidelines concerning the methods and principles of terminology management involving the production of terminological entries as described in ISO 704 (ISO 704). Because a terminology always deals with special language in a particular field of knowledge, the concept shall be viewed as a unit of knowledge. The concept is a higher level of abstraction in a terminology; it links an object and its designations. The concepts contextualized in the special language of the subject field can be expressed in the various forms: terms, appellations, definitions or other linguistic forms (ISO 704). One of the most important characteristics of a terminological entry is its concept orientation: a terminological entry represents one concept which is designated by one or several terms in one or several languages.

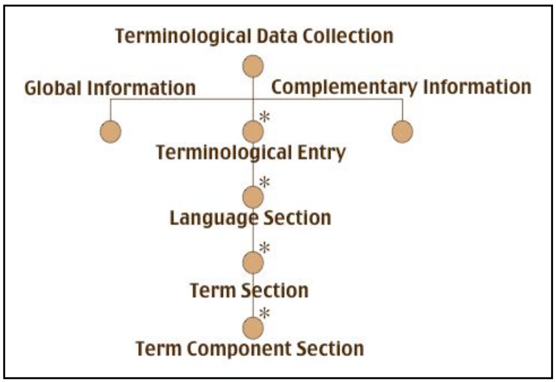


Figure 1: TMF Meta-model

Each entry can have multiple language sections, and each language section can have multiple terminological units. Each data element in an entry can be associated with various kinds of descriptive and administrative information.

3.2. Data category

A meta-model contains several information units related to a given format, which we refer to as “Data Categories”. A selection of data categories (DCS) can be derived as a subset of a Data Category Registry (DCR) (Ide and Romary, 2004) ensuring that the semantics of these data categories are well defined and accepted by community of specialists. A data category is the generic term that references a concept. For example, the data category /*originatingInstitution*/ indicates an institution (i.e. company, government agency, etc.) treated as a source of information for the purpose of bibliographic documentation. For each element in TermSciences, the originating institution is mentioned in order to document the source of the data. A Data category Selection is needed in order to define, in combination with a meta-model, the various constraints that apply to a given domain-specific information structure or interchange format. A DCS is firstly used to specify constraints on the implementation of a meta-model instantiation, and secondly to provide the necessary information for implementing filters that convert one instantiation to another and to produce a “Generic Mapping Tool” (GMT) representation.

3.3. Introduction to GMT

GMT can be considered as a XML canonical representation of the generic model. The hierarchical organization of the meta-model and the qualification of each structural level can be realized in XML by instantiating the abstract structure shown above (Figure 2) and associating information units to this structure. The meta-model can be represented by means of a generic element <struct> (for structure) which can recursively express the embedding of the various representation levels of a TMF instance. Each structural node in the meta-model shall be identified by means of a type attribute associated with the <struct> element. The possible values of the type attribute shall be the identifiers of the levels in the meta-model (i.e., Terminological Data Collection, Global Information, Terminological Entry, Language Section, Term Section, Term Component Section).

Basic information units associated with a structural skeleton can be represented using the <feat> (for feature) element. Compound information units can be represented using the <brack> (for bracket) element, which can itself contain a <feat> element followed by any combination of <feat> elements and <brack> elements. Each information unit must be qualified with a type attribute, which shall take as its value the name of a standardized data category or one user-defined data category.

4. IMPLEMENTATION

As the source vocabularies are diverse with respect to format, structure and content, they were analyzed and restructure to fit the meta-model, in order to allow for high interoperability between terminological systems. Following this, comparisons were made between all the resources and common concepts were grouped in terminological entries in which data belonging to different resources were issued with their sources. Terminological resources

4.1. Terminological resources

The terminologies used in the preliminary phase of this project are vocabularies from four French research institutes: indexing vocabularies from the Institut de l'Information Scientfique et Technique (INIST-CNRS); the MeSH thesaurus from the US National Library of Medicine including its French translation by the Institut de la Santé et de la Recherche Médicale (INSERM); the thesaurus of public health produced by the Banque de données de Santé Publique (BDSP) and the Dictionary of Human and mammals reproduction biotechnology produced by the Institut National de la Recherche Agronomique (INRA).

4.2. From descriptors to concept

Instead of simply being aggregated, these native resources were fused together. For example, the term "*Gamete Intrafallopian Transfer"* with several translations and a definition (Figure 2) was found in NLM, INRA and INIST resources and it refers to the same object.

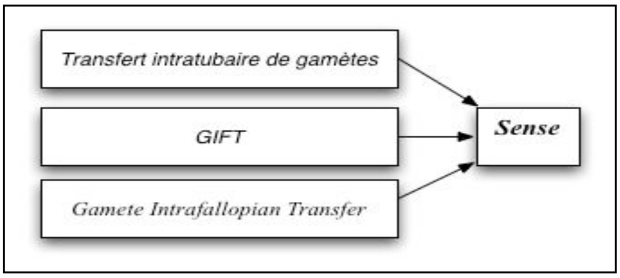

Transfert intratubaire de gamètes
GIFT
Gamete Intrafallopian Transfer
Sense

Figure 2: Semasiologic view of GIFT

Thesauri or lexicons present a semasiologic view of the world (figure 2) and are frequently arranged by alphabetic order. The main challenge of this project was to have another view of the data, no more a semasiologic view but rather an onomasiologic one (Romary andVan Campenhoudt, 2001) (Figure 3).

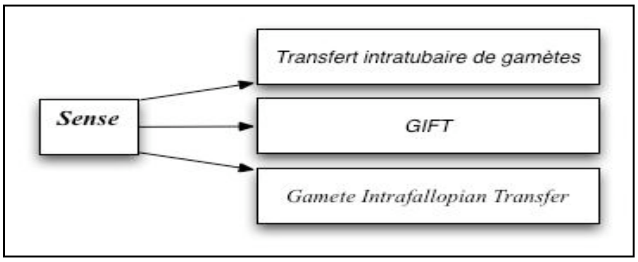

Sense
Transfert intratubaire de gamètes
GIFT
Gamete Intrafallopian Transfer

Figure 3: Onomasiologic view

4.3. Heterogeneous data

The resulting terminological record for a given concept presents terms and relationships that may be conflicting. In source terminologies such as the MeSH thesaurus or the public health thesaurus which are organized and used for library indexing, different concepts may be present in the same record under the same descriptor depending on the degree of specificity. For example, the record in BDSP thesaurus presents the term "Brain" as a descriptor for "Cortex", i.e. a "Used for" relation links the two terms in this thesaurus which presents only broader levels for anatomical terms. When this record was processed for integration in the common terminological database, the term "Cortex" was captured as a synonym of the term "Brain". In highly structured resources such as the MeSH thesaurus, entry terms which are synonyms, or closely related terms are documented as non-preferred concepts which allowed us to discard them during the integration process. Additionally, every resource comes up with its own categories and relationships. Thus, this first substrate needs major improvements in terms of smoothing of conflicts that may appear between concept categorization or semantic networking strategies.

In the integrated TermSciences terminological content it is important to document and identify the source of each element. Thus, the resulting terminological record for a given concept presents meta data for terms, relationships, definition, etc. These meta data allows for inclusion of some administrative information like the last modification date for an element. Additionally partners can update or export their own data according to their origin. Figure 4 illustrates the documentation of sources meta data for the above example on figure 2. The concept will be illustrated by a definition and a set of terms in different languages; each element being accompanied by its origin.

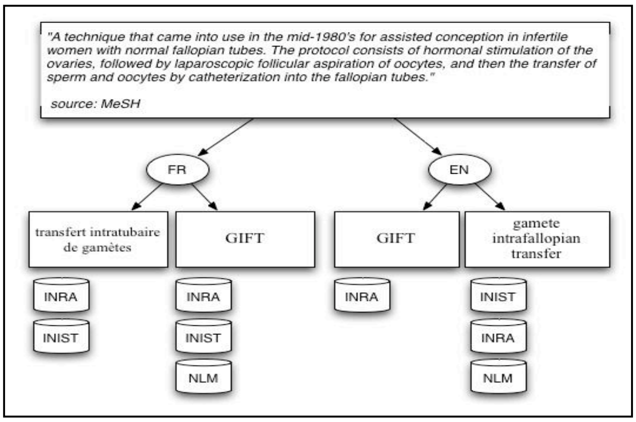
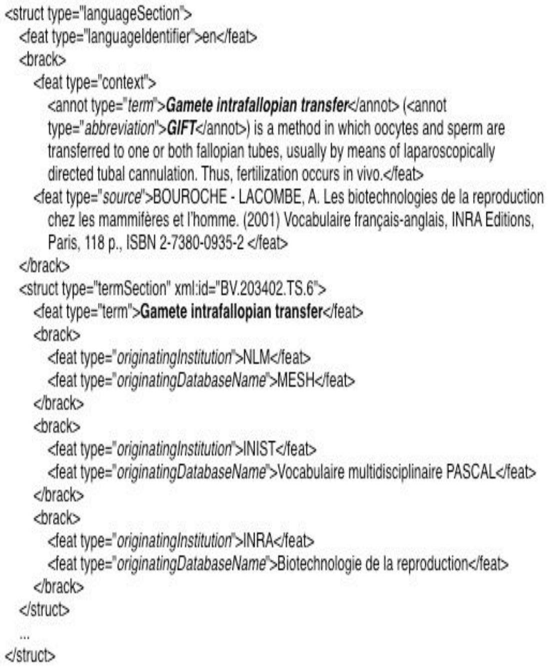

"A technique that came into use in the mid-1980's for assisted conception in infertile women with normal fallopian tubes. The protocol consists of hormonal stimulation of the ovaries, followed by laparoscopic follicular aspiration of oocytes, and then the transfer of sperm and oocytes by catheterization into the fallopian tubes."
source: MeSH
FR
EN
transfert intratubaire de gamètes
GIFT
GIFT
gamete intrafallopian transfer
INRA
INIST
INRA
INIST
NLM
INRA
INIST
INRA
NLM

Figure 4: Data sources

The meta data shows the origin institution and/or database, but can also give a bibliographical reference. For example, several partners furnished this term "Gamete Intrafallopian Transfer"; one of them published this vocabulary. It is important to be able to complete institutional information by a bibliographical source (figure 5).



```
<struct type="languageSection">
  <feat type="languageIdentifier">en</feat>
  <brack>
    <feat type="context">
      <annot type="term">Gamete intrafallopian transfer</annot> (<annot
      type="abbreviation">GIFT</annot>) is a method in which oocytes and sperm are
      transferred to one or both fallopian tubes, usually by means of laparoscopically
      directed tubal cannulation. Thus, fertilization occurs in vivo.</feat>
    <feat type="source">BOUROCHE - LACOMBE, A. Les biotechnologies de la reproduction
      chez les mammifères et l'homme. (2001) Vocabulaire français-anglais, INRA Editions,
      Paris, 118 p., ISBN 2-7380-0935-2 </feat>
  </brack>
  <struct type="termSection" xml:id="BV.203402.TS.6">
    <feat type="term">Gamete intrafallopian transfer</feat>
    <brack>
      <feat type="originatingInstitution">NLM</feat>
      <feat type="originatingDatabaseName">MESH</feat>
    </brack>
    <brack>
      <feat type="originatingInstitution">INIST</feat>
      <feat type="originatingDatabaseName">Vocabulaire multidisciplinaire PASCAL</feat>
    </brack>
    <brack>
      <feat type="originatingInstitution">INRA</feat>
      <feat type="originatingDatabaseName">Biotechnologie de la reproduction</feat>
    </brack>
  </struct>
  ...
</struct>
```

Figure 5:Gamete Intrafallopian Transfer in GMT

5. DISCUSSION

The TermSciences initiative deals with the construction of a multi-purpose and multi-lingual terminological database from various source vocabularies produced by major French research institutions. The first requirement of this work was the use of a model that allows for good modeling of data present in these source vocabularies. This was achieved using a data model based on the ISO 16642 standard which was found to be very suitable for modeling term-centred terminological resources into a concept-oriented system. Transformation of terms into concepts was accompanied by transformation of term relationships into concept relationships, i.e. hierarchical and associative relationships are no more at the term level but at the concept one. Adaptations of the traditional terminology principles (wüsterian) are necessary when dealing with specific terminological resources such as thesauri and indexing vocabularies. Thus, the representation of preferred and non-preferred concepts referring to the same descriptor was achieved by introducing a relation at the level of terms. Non-preferred concepts are introduced in the terminological database as separate records but are linked to the preferred concept by a relation occurring at the level of terms. This relation links a term which corresponds to a synonymous concept in a given thesaurus to the term corresponding to the preferred concept which is labeled as being the descriptor. The organization of concepts relevant for a particular domain varies from one source vocabulary to another depending on the degree of precision needed by each application (Rassinoux et al, 1998). Thus, the hierarchy in the MeSH thesaurus may be simple or multiple presenting a given descriptor in different positions in the hierarchy. Furthermore, hierarchies from different source vocabularies may not map correctly, resulting in conflicting positioning of some concepts in the semantic network. Dealing with this topic can be achieved by a) finding a consensual typology of concepts which is not impossible if the level of detail of the typology is not high or b) by representing multiple typologies, i.e. the hierarchies present in the different source vocabularies and additional typologies further introduced.

5.1. Reusing of the terminological database

TermSciences is already available on-line and can be used for querying a bibliographical database or helping translator or linguist in a specific subject field. We are planning to add other free bibliographical databases such as PubMed and others. Using the French and English terms contained in a terminological entry, the query is automatically composed and launched on the specified repository. In addition to the cross-language retrieval of relevant documents and citations, another great advantage of this system is the possibility to search bibliographical databases with terms from alternative thesauri and vocabularies. Indexing and cataloging activities being upstream from information retrieval, the terminological database is intended to be connected to bibliographical databases production systems. These systems are those of the TermSciences partners whose needs are about the improvement of their controlled vocabularies management processes and tools, and of the optimization of the indexing process, especially machine-aided indexing programs which performance relies on the quality of the terminological content. The multiple representations (terms) of a given concept which are documented in the terminological database and the variant forms that can be obtained using natural language processing techniques (see bellow) are expected to enhance precision of the machine-aided indexing procedures through consistent interpreting of texts and suggestion of appropriate indexing terms. Another important application is the HAL (Hyper Article en Ligne) institutional open archive of the French researchers which provides authors with an interface enabling them to deposit and index their scientific articles in this repository which is managed by the Center for Direct Scientific Communication, a service unit of the CNRS. At least, this resource will be freely available.

5.2. Adding linguistic resources

Additional resources are crucial for a) harmonising the quality and the granularity of the various linguistic descriptions of terms, and b) for purposes such as semi-automatic indexing, information retrieval, translation, etc (Cabre and al., 2005). Natural language processing using on the available lexical features of terms is needed to enhance the recognition rate and quality.

The adding of lexical features in the TermSciences terminological database is being examined from two points of view: tagging of terminological database terms or capturing of lexical features from existing lexical resources such as Morphalou (ATILF) for French terms.

Adding of lexical information is intended to meet another requirement, i.e. to increase the consistency of the set of synonym terms present in a terminological entry. That is, in controlled vocabularies such as those used to build the TermSciences terminological database, morphological variants of the same term are often present and are considered as being synonymous of the preferred term (Zweigenbaum et al. 2003). This results in an artificial inflation of permuted or inflected expressions in some terminological entries. For instance, the MeSH thesaurus presents permuted forms in records such as ‘Primary Parkinsonism’ and ‘Parkinsonism, Primary’. Term tagging or coupling with lexical resources will result in a deflation of the set of terms by discarding the terms which correspond to lexical variants differing from each other only by spelling, word order, number, etc.

In the biomedical field, a salient project, i.e. the Unified Medical Language System (UMLS; McCray et al. 1993) deals with this topic. In this project, lexical knowledge is provided as a distinct source, the SPECIALIST lexicon (McCray, 1998). Coverage of this knowledge source includes both commonly occurring English words and biomedical vocabulary. As English language part in UMLS knowledge sources is greater than that of other languages including French, two projects, i.e. the Unified Medical Lexicon for French (UMLF) which aims at providing a French equivalent for the SPECIALIST lexicon (Zweigenbaum et al. 2003, Zweigenbaum and Grabar, 2003), and the VUMeF project (French Unified Medical Vocabulary) which aims at extending the French part of the UMLS metathesaurus (Darmoni et al.2003).

5.3. Corpora

The use of selected corpora represents another important topic for the capture of additional elements in the terminological database such as contexts of use and for terminological extraction. For instance, contexts of use are very useful to translators since they reflect the actual use (or misuse) of a term. The automatic capture of contexts from bibliographical database abstracts or full-text records produced by TermSciences partners is explored as a first step toward context assignment to each term in the terminological database. As human indexers handle the terminological material during rule editing for machine-aided indexing, automatically-captured candidate contexts for terms will be suggested and then verified by human indexers for final selection before addition to the terminological database.

Concerning term extraction, corpora stored in bibliographical databases or incoming bibliographical records subjected to machine-aided indexing routines will be used to suggest candidate terms and candidate semantic relationships between terms (Jacquemin, 1997). The expression of term relationships in texts being revealed by connective words such as ‘is called’ ‘is a’, etc (Jacquemin and Bourigault, 2003), cue words and rules for different knowledge domains must be defined through linguistic studies of text samples and then used by computer prohrams to explore these texts and find semantically related terms. Other methods do not require patterns or rules and may use collocation, i.e. cohesive lexical clusters, retrieving (Smadja, 1993)

6. Conclusion

The TermSciences initiative aims at building a terminological database by integrating various vocabularies mainly used for indexing purposes. As a first step toward integration, standardization of the source vocabularies was obtained through deployment of the ISO 16642 also called TMF. Although, this standard turned on be suitable for modeling and sharing of the source vocabularies, adaptations were necessary for modeling specific relations which occur frequently in indexing controlled vocabularies, i.e. relations linking non-preferred terms (non-descriptors) to the preferred term (descriptor). Further work is also needed to improve the content of the terminological database and to introduce additional data such as linguistic features, contexts of use, etc.

7. References

Baud, A.H., Lovis, C., Rassinoux, A.-M., Sherrer, J.-R. (1998) Alternative ways for knowledge collection, indexing and robust language retrieval. Met. Inform. Med. 37(4-5). 315-326.

Bourigault, D., Condamines, A. (1995) Réflexions sur le concept de base de connaissances terminologiques. In: Actes des Cinquièmes journées nationales du PRCGDRIA, Nancy.

Cabre, T., Condamines, A., Ibekwe Sanjuan, F., (2005). “Introduction to Application-driven Terminology Engineering”. Terminology n°11-1.

Cimino, J.J. (1998). Desiderata for controlled medical vocabularies in the twenty-first century. Met. Inform. Med. 37(4-5). 394-403.

Condamines, A. (1994). Terminologie et représentation des connaissances. *La banque des mots*;6:29–44.

Darmoni, S.J., Jarrousse, E., Zweigenbaum, P., Le Beux, P., Namer, F., Baud, R., Joubert, M., Vallée, H., Cote R.A., Buemi, A., Bourigault, D., Recource, G., Jeanneau, S., Rodrigues, J.M. (2003) *VUMeF: Extending the French Involvement in the UMLS Metathesaurus*. Proc AMIA Symp.

Gangemi, A., Pisanelli, D.M., Steve, G.(1998). Some requirements and experiences in integrating terminological ontologies in Medicine (http://ksi.cpsc.ucalgary.ca/KAW/KAW98/gangemi/).

ISO 704 (1987). *Principes et méthodes de la terminologie*, Geneva, International Organization for Standardization (ISO/TC 37).

ISO 12620 (1999). Computer applications in terminology -- Data categories, Geneva, International Organization for Standardization

ISO 16642 (2003). *Computer applications in terminology -- Terminological Markup Framework,* Geneva, International Organization for Standardization (ISO/TC 37).

Jacquemin, C. (1997). Variation terminologique : Reconnaissance et acquisition automatique de termes et de leurs variantes en corpus. Mémoire d'habilitation à diriger des recherches, Université de Nantes.

Jacquemin, C. and Bourigault, D. (2003). Term extraction and automatic indexing. In The Oxford Handbook of computational Linguistics (pp. 599-615). Ed. Ruslan Mitkov, Oxford University Press.

Ide N., Romary L. (2004). A Registry of Standard Data Categories for Linguistic Annotation, in: "4th International Conference on Language Resources and Evaluation - LREC'04", May 2004.

McCray, A.T. (1998). The nature of lexical knowledge. Met. Inform. Med. 37(4-5). 353-360.

McCray, A.T., Aronson, A.R., Browne, A.C., Rindflesch, T.C., Razi, A. and Srinivasan, S. (1993). *UMLS knowledge for biomedical language processing*. Bulletin of the Medical Library Association 1:184-94.

Rassonoux, A.M., Miller, R.A., Baud, R.H., Scherrer, J.R. (1998). Modeling concepts in medicine for medical language understanding. Met. Inform. Med. 37(4-5). 361-372.

Rastier F. Le terme : entre ontologie et linguistique. *La banque des mots* 1995;7:35–65

Romary (L.) et Van Campenhoudt (M.), (2001). « Normalisation des échanges de données en terminologie : le cas des relations dites "conceptuelles" », dans *Actes des 4es Rencontres terminologie et intelligence artificielle (Nancy, 3-4 mai 2001)*, Nancy : INIST-C.N.R.S., p. 77-86.

Smadja, F. (1993) Retrieving collocations from text : Xtract. Computational Linguistics 19. 143-176.

WÜSTER, E. (1976). "La théorie générale de la terminologie -un domaine interdisciplinaire impliquant la linguistique, la logique, l'ontologie, l'informatique et les sciences des objets", dans DUPUIS (H.), éd., *Essai de définition de la terminologie. Actes du colloque international de terminologie (Québec, Manoir du lac Delage, 5-8 octobre 1975)*, Québec, Régie de la langue française, p. 49-57.

Zweigenbaum, P., Baud, R., Burgun, A., Namer, F., Jarrousse, E., Grabar, N., Ruch, P., Le Duff, F., Thirion, B., Darmoni, S.(2003). UMLF: a Unified Medical Lexicon for French. AMIA Annu. Symp. Proc. :1062.

Zweigenbaum, P. and Grabar, N.(2003). Corpus-based associations provide additional morphological variants to medical terminologies. AMIA Annu. Symp. Proc. :768-72.

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

Joost Buysschaert
Hogeschool Gent
Departement Vertaalkunde, Groot-Brittanniëlaan 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 Taalvlinder (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 co-operation 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 Schaüble 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 (http://debussy.hon.ch/cgi-bin/HONselect?search). A (partial) Dutch version commissioned by the *Nederlands Tijdschrift voor Geneeskunde*, codenamed MSHDUT2004, is obtainable for research purposes (though not for commercial purposes) .

Yet traditional terminologists have been quick to point out errors in the existing translations and have claimed that they are "rough and ready" conversions only. While this claim is awaiting substantiation (i.e. via a detailed review), it is true that the translation of extensive lists like the MeSH headings, spanning several specialisms, is a very time-consuming task (if it is to be done well) so that the fast creation of equivalent lists is at least suspicious. There are also other aspects that traditionalists are likely to frown upon; but also aspects that they tend to ignore and that the NLP supporters are much better at. Examples of either category are explored and illustrated below.

3.1. The issue of evidence

The creation of one-to-one lists relieves the makers of the arduous task of giving evidence. Traditional terminologists like to quote their sources in evidence; the term is given in one or more original fragments of text ("contexts"), with a detailed reference to the source. Sometimes the reference is to an informant. These details are often absent from a machine readable glossary. While this is understandable, it should be a matter of principle that even when machine readable lists do not give quoted examples or other evidence, the lists should somehow be backed up by a database that does give these data.

3.2. The issue of synonyms

Machine readable glossaries prefer to believe in the fiction that technical vocabularies have one term for one concept. While this is the ideal situation in a normative approach (and was also the situation envisaged by the founding father of terminology, Eugen Wüster), it definitely does not hold true of medical terminology. Monolingual medical dictionaries of English illustrate that the same concept is often referred to by a whole series of synonyms. The treatment of a patient with drugs, for example, can alternatively be termed *drug treatment*, *pharamacotherapy, pharmacologic therapy*, *pharmacological treatment* or *medication therapy*. The International dictionary of medicine and biology (Landau et al., 1986), in particular, has a habit of quoting many alternatives. While some of these may be related terms rather than true synonyms (and while it is wise also in other respects to make a distinction between "true synonyms" and "near synonyms" / "extra synonyms", cf. 3.4 below) , it remains undeniably true that the use of alternative names is common in medicine.

Where terminologies are used for indexing, there is a feeling that synonyms should be disregarded and that preference must be given to a favoured term (the normative approach). The human translator, however, knows that each of the alternative terms may crop up in a text so s/he is interested in having them all recorded in the termbase.

Yet even for NLP purposes, it is interesting not only to establish reference terms but also to link them up with synonyms (or even cognate words). This is already done in document retrieval. Here too, the detailed groundwork 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 (*anemia* versus *anaemia*); 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):

nor	This term was found in a normative source
leg	This term is the legally used term
pri	This term has been found used in only one primary source
2pri	This term has been found used in only two primary sources
3pri	This term has been found used in *at least* three primary sources
sec	This term has been found used in only one secondary source
2sec	This term has been found used in only two secondary sources
3sec	This term has been found used in *at least* three secondary sources
neo	This term is a neologism created by the terminologist (and preferably sanctioned by a domain specialist).

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 in 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 *artrose*) . 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>
- <descripGrp>
  <descrip type="PartOfSpeech">noun</descrip>
  </descripGrp>
- <descripGrp>
  <descrip type="Context">If the programmer establishes
distance telemetry with multiple devices, it lists each one
with a unique patient identifier.</descrip>
…
```

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
<Internat>
<Nl-term>kroeshaarsyndroom
<Equival>
<En-term>kinky hair syndrome
<Equival>
<Beeld>


<English>kinky hair syndrome [UPDATE JY]
<Trefwoord>kinky [UPDATE JY]
<Betrouwb>nor pri 3sec [UPDATE JY]
<Woordsrt>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 *palpatation* 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

Browne, A.C., McCray, A.T., Srinivasan, S. (2000). *The Specialist Lexicon*, Bethesda, Maryland: Lister Hill National Center for Biomedical Communications, National Library of Medicine. http://lexsrv3.nlm.nih.gov/SPECIALIST/Projects/lexicon/2006/release/LEXICON/DOCS/techrpt.pdf

Francopoulo, G., George, M., Calzolari, N., Monica Monachini, M, Bel, N., Pet, M., Soria, C. (2006) Lexical markup framework. Submission to LREC 2006. http://lirics.loria.fr/doc_pub/LMFPaperForLREC2006FinalSubmission6March06.pdf

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

Kerkhof, J.L. (2003). *Woordenboek Geneeskunde: E-N/ N-E* . Maarssen: Elsevier Gezondheidszorg.

Landau, S.I. et al. (eds.) (1986). *International dictionary of medicine and biology.* New York etc.: Churchill Livingstone, 1986.

MeSH Browser. http://www.nlm.nih.gov/mesh/MBrowser.html

Mostert, F.J. (2002). *Medisch Woordenboek: Engels-Nederlands, Nederlands-Engels.* Houten & Bohn Stafleu van Loghum

Multilingual Glossary of Technical and Popular Terms in Nine European languages. http://allserv.rug.ac.be/~rvdstich/eugloss/welcome.html

Reineke, D. (2005). Martif und TBX. Austauschformate für Terminologie. [Slides for] 2. Kölner Tagung "Softwarelokalisisierung". IIM der FH Köln.

Schaüble, P. et al. Cross-Language Information Retrieval. http://trec.nist.gov/pubs/trec6/papers/clir_track_US.ps.gz

Taalvlinder. http://www.ochrid.dds.nl/medici.htm

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

The specialist lexicon. http://lexsrv3.nlm.nih.gov/SPECIALIST/Projects/lexicon/2006/release/LEXICON/LEXICON.

Vander Stichele, R. (2004). *Impact of written drug information in patient package inserts. Acceptance and benefit/risk perception. Thesis submitted as partial fulfilment of the requirements for the Degree of Doctor in Medical Sciences*. Ghent University.

Woordenboek Ziekenhuistermen. Part of *Digitaal Ziekenhuis Nederland*. http://www.ziekenhuis.nl/

Designing an Ontology of Dialogue Elements Modeling Doctor-Patient Exchanges

Leslie Barrett
EDGAR Online, Inc.

Abstract

This paper describes a proposed addition to an existing ontology of medical terms used in medical point-of-service interactions, (i.e., doctor-patient communications). The ontology is contained as a module in an Arabic-English bi-directional machine-translation lexicon originally created to insure broad coverage in a commercial machine-translation application. Although the existing ontology covers far more than disease terms, including entries for words commonly used to describe symptoms, treatments, prescriptions and tests, it does not include words associated with the events in which they participate. In particular, connecting the instances of Disease, Symptom and Treatment words with events requires knowledge of related verb groups including verbs such as *drink, swallow, eat, hurt, throb, tingle* and the like. Poor translation of these lexical elements negatively impacts translation quality significantly. We therefore propose a strategy for enhancing the existing lexical resource with new verb links connecting a sub-network of these event elements to selected nodes in the ontology.

Introduction

The medical dialogue ontology (MDO) discussed here was originally created for use as a machine-translation lexicon model[1]. The point-of-service interactions for which the lexicon was intended (and ultimately the commercial MT system in which it was used) involve terminology pertinent to doctor-patient *dialogues* about medical conditions, not simply terminology specific to particular conditions themselves. For example, facilitating a successful communication between a doctor and a patient necessarily involves discussion not just of the patient's condition, but also of possible treatments for and symptoms of that condition. The MDO contains English words hierarchically organized into medical-dialogue categories each of which contains an Arabic translation. The intention of this resource is to have a language-independent lexical tool to use as a model for other translation pairs and other MT products in the same domain.

1.0 Improving Resources for Broader Coverage in Medical-Domain MT

Tuning an MT system for optimal performance in the domain of medical point-of-service interactions requires far more than improving the domain-specific lexical inventory. We have attempted through creation of the MDO to insure lexical coverage not simply for medical terms, but for terms used in *medical dialogues*. The remaining step, however, is to insure that dialogue-specific elements are covered as well. The MDO does not cover important linguistic aspects of medical discussions such as speaker intention, sentential type, speech act type or event type. Previous research has attempted to address this need in other domains. Levin et al (2003) developed a coding scheme for machine translation of spoken task-oriented dialogue. They argue that domain actions are the most relevant discourse unit for improving translation quality, and discuss the development of speech act and domain action classifiers.

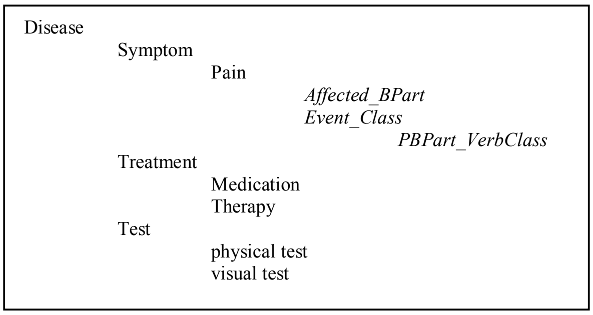
[1] The work was supported by DARPA grant #_________ completed in January 2005.

Levin et al (2003) point out that although speech acts are domain independent, task-oriented language tends to contain fixed expressions with domain specific functions. This applies to the medical dialogues in question here as well. For example, consumption verbs are found in dialogue segments involving prescriptions for treatment. The verbs *take, drink, eat, consume, ingest, swallow, inhale*[2] are generally followed by NPs representing food or medication, while *rest, apply, cover, bathe, wash, clean,* tend to be followed by NPs representing a body part affected by an injury. Dialogue turns involving patient feedback tend to have predictable verb patterns as well. They fall mostly into a class described in Framenet (see http://www.**framenet**.icsi.berkeley.edu) as the Perception_body class and include the following lexical set: *ache.v, ail.v, burn.v, goosebump.n, hurt.v, itch.v, pain.n, pain.v, prickle.v, smart.v, sting.v, tickle.v, tingle.v*. Other verb classes common in patient dialogue turns are verbs of experiencing like *feel, experience, have*. Members of the former class have a body part subject in the intransitive, or an experiencer subject and body part object in the transitive[3].

Like the domain-action model in Levin et al (2003) our ontological representation of medical dialogues could include elements representing domain actions. A system similar to the domain action classifications described in Levin et al. (2003) could be implemented as the interface with the MDO node (e.g. *treatment, symptom, etc.*) which would have links to topic-associated word classes[4]. The "Symptom" node, for example, in addition to containing words that represent instances of symptoms (and their translations), would also contain members of an associated verb class such as the Perception_Body class mentioned above.

An example of an MDO snippet with proposed additional links for enhancing event information is shown below in Figure 1.

Figure 1: Event Enhancement for Medical Dialogues



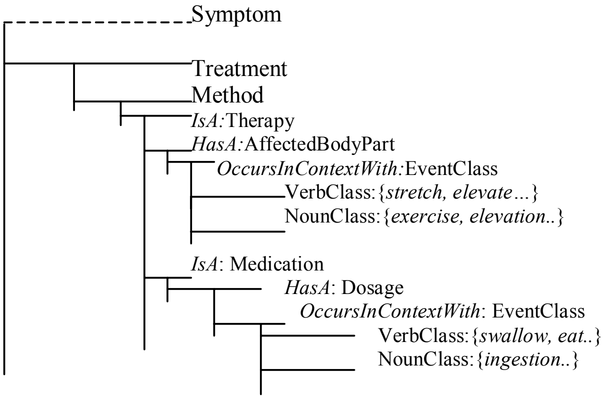
In the above example, words in the Perception Body Part Verb Class would be used in dialogues involving the verbal description of symptoms by patients. Elements of the verb class may also have commonly used nominalizations or synonymous phrases including verb+NP. For example, relations such as ache<v> → ache<n> or hurt<v> → have_pain<V,NP> can be represented in the lexicon by adding a feature on to the instance entry for verb class elements. Because the representations would not be syntactically static, we will refer to the proposed addition as "Event Class" type information. Figure 2 shows a more detailed representation of the proposed enhancements to the hierarchy.

[2] Covered partially by the "Ingestion" frame in Framenet and including the following lexical set there: *breakfast.v, consume.v, devour.v, dine.v, down.v, drink.v, eat.v, feast.v, feed.v, gobble.v, gulp.n, gulp.v, guzzle.v, have.v, imbibe.v, ingest.v, lap.v, lunch.v, munch.v, nibble.v, nosh.v, nurse.v, quaff.v, sip.n, sip.v, slurp.n, slurp.v, snack.v, sup.v, swig.n, swig.v, swill.v.* See http://framenet.icsi.berkeley.edu/index.php

[3] See Levin (1993) for details on the Causative/Inchoative alternation class that these verbs fall into.

[4] This small snippet of the MDO does not indicate the relations between nodes. The hierarchy shown is not an Is-A hierarchy but contains several types of relations that space does not allow a discussion of here. For example, a "Symptom" IS_AN_INDICATOR_OF a "Disease".

Figure 2: Snippet of Hierarchy of MDO Nodes with Event Class Nodes


Symptom
Treatment
Method
*IsA:*Therapy
*HasA:*AffectedBodyPart
*OccursInContextWith:*EventClass
VerbClass:{*stretch, elevate…*}
NounClass:{*exercise, elevation..*}
IsA: Medication
HasA: Dosage
OccursInContextWith: EventClass
VerbClass:{*swallow, eat..*}
NounClass:{*ingestion..*}

The bottom of the tree in this example represents an "event layer" denoting the events in which instances of the sibling node are participants. For example, (simplifying the "Dosage" contents somewhat) we might find participants in a dosage event to be an individual, a medication, and a frequency. Medications and individuals are typically related by events like "swallowing", "taking", "ingesting" and the like. Thus the "Dosage" node will have a sibling Event_Class node with appropriate verb and noun instances such as those found in the Framenet "Ingestion" class.

2.0 **Features and Instances**

The addition of verb-class information is meant to enhance the ontology with respect to its utility as a translation lexicon. It does, however, add some complexity by introducing the notion of parts of speech into an otherwise conceptual ontology. Furthermore, the existence of metonyms such as the previously cited *"I **have a pain** in my <body_part>"* for *"My <body_part> **hurts*** suggests that multiple parts of speech belong as instances of an "event class" and those instances may need to be stored with features representing part-of-speech, number and perhaps gender information. A typed feature-structure representation of a typical "EventClass" (here we use the "pain" class) instance might be as in figure 3, where the English and Arabic gloss, category and head-feature information are all represented and available for verbal and nominal features[5]:

Figure 3: Typed Feature Structure Model

```
[ lex:        hurt
  cat:        V
  class:        pain
  glossE:       #hurt
  glossA:       ملؤي
  headE:     [ agr:    []
              number:
              proper:
              verbal:   + ] ]
```

To capture the intuition that "have a pain" and "hurt" belong to the same semantic class, the category inventory could be increased to allow phrases. Thus, with the inclusion of a Verb Phrase <vp> category, "have_a_pain" could be part of the semantic set including simple verbs like "hurt" and "ache". The feature set shown in the example above is not a part of the disease-name instances collected so far in the database mainly because the ontology was designed to be an application-neutral lexicon. Software that utilizes the lexicon may have independent mechanisms for recognizing this information, and that must be programmed by hand when the lexicon is loaded. Also, because Arabic and English will have different valency patterns there may be feature slots with entries in only one language. This is the case in similar projects such as the Japanese-English Valency Dictionary described in Bond and Shirai (1997).

2.1 **Relations**

Because adding new layers to an ontology often involves adding new relations as well, we will briefly address how the event layer will fit in with its parent nodes. We propose a generalized new relation to handle the connections between these verb sets and the entity types (most often Body Parts) with which they are associated.

Although body part relations are well known to meet the criteria of antisymmetry, reflexivity and transitivity that defines mereological relations, the associations between the

[5] The feature structure shown is simplified somewhat as additional nominal and verbal features would be necessary to accommodate both English and Arabic. We omit here the Arabic Head feature set.

proposed verb classes and body parts is less clear[6]. Thus where "P" represents "Part", the following hold of the relations between body parts such as *digit* → *hand* → *arm*, etc (see Sowa 2000, Guarino and Welty 2001 for in depth discussion of partial ordering relationships, Gerstl and Pribbenow 1995 for a discussion of body part relations):

The relation between verbs and their arguments, however, becomes more complex as it involves semantic roles. There is no proto-role appropriate to link a concept "VerbClassX" with a concept "BodyPartX", but rather individual verbs pick out individual sets of semantic roles. For example, Framenet lists verbs in the Perception_Body_Part class as being associated with an Experiencer role[7] as well as a Body_Part. In the case of a typical instance this leads immediately to problems in creating a single relation. For example "hurt" can be associated with any body part in more than one way, and the type of association has an important medical implication:

My leg hurts. → suggests cause could be unknown and injury uncertain.

I hurt my leg. → suggests cause is known and an injury is suspected.

Thus a simple relation such as "HasSubject"[8] is inappropriate for connecting the Perception_Body_Part verb class with Body_Part and equally inappropriate for connecting individual verb instances with individual body part instances.

The issue of semantic role mapping and its challenges to ontological representation has been explored before. Davis and Barrett (2002) discuss issues of interfacing a hierarchy of semantic roles with the situations in which they are used. They point out challenges to inheritance in situation-types and the impact that has on the inheritance properties of the roles within those situations. For example, they point out the problem of Lehmann's (1977) situation-type hierarchy where complex situation types inherit from multiple parents. They give the example of *taking a trip in a car*, including the sub-events *unlocking the car* and *driving the car*. The former sub-event has a key as an instrument. If however the "taking a trip in a car" event inherits the roles of its sub-events unconstrained, then the key will be an instrument in *taking a trip in a car* as well, which is undesirable. Similarly, a discussion about Treatments might involve discussions about swallowing pills but not opening pill jars. Even more complex are instances of events which imply body part functions without stating them, such as "*swallow*", "*eat*", "*ingest*" and similar activities normally associated with taking medications (ultimately parented by the Treatment node).

If we avoid these complexities by linking at the class level, with a contextually-based relation like "OccursInContextWith" we will lose the similarity with other common ontological relations. This relation will not be Antisymmetric (i.e. because if A occurs in context with B, then B occurs in context with A) or Transitive (i.e. because if A occurs in context with B and B occurs in context with C, it is NOT true that A occurs in context with C) – although it will be Reflexive (i.e. because A occurs in context with A).

(P.1.1)	$OICWxx$	*Reflexivity*
(P.2.1)	$[\neg \forall (OICWxy \; \& \; OICWyx) \rightarrow x=y]$	*nonAntisymmetry*
(P.3.1)	$[\neg \forall (OICWxy \; \& \; OICWyz) \rightarrow OICWxz]$	*nonTransitivity*

[6] Although we take the position here that body parts are subject to transitivity see Gerstl and Pribbenow (1995) for opposing view.

[7] We ignore for now the issue of whether this role maps to the grammatical subject or object.

[8] Along the same lines, "HasExperiencer", while appropriate for the first example, would be inappropriate for the second, and all other cases where the Experiencer is the Agent of the experience. The agency in the example is what is pertinent in the medical context since it strongly implies the presence of injury.

However, given the main function of the ontology as a translation lexicon, this kind of class-based relation which refers only to textual dependence ignoring details of semantic relations among class instances, is the simplest solution. Therefore although the body parts themselves may be connected in a sub-hierarchy of mereological relations, the relation of the Body_Part class to the Perception_Verb class will be simple and practically semantically vacuous. Individual verb and body part instances will not be connected[9].

An example below[10] shows dialogue information added to the topical lexical entries in a typed-feature-structure containing the elements show in Figure 4[11].

Figure 4: Lexical Entry Containing Event Information

```
[ lex:       bruise
  cat:       N
  glossE:    #bruise
  glossA:     مدكي
  head:    [ agr:   [ 3s:  + ]
             proper:  -
             verbal:  - ]
  treatment:       [ [ type: Medication [1]
                       EventClass: Ingest [2]

                 [list [1]: [ nounsyn: [ lex:  analgesic
                  cat:       N
                 glossE: #analgesic
                 glossA: # نِّكَسُم
                head: [agr: 3s +]
                    proper: -
                    verbal: - ]
                instance: [ lex:  ibuprofin
                 cat:       N
                 glossE: #ibuprofin
                glossA: # نيفوربوبيأ
                 head: [agr: 3s +]
                  proper: -
                  verbal - ] ]
              [list [2]: [verbsyn: [lex: swallow
                  Cat: V
                 glossE: #swallow
                glossA: # علبي
                head: [agr: ] ]]]
                                ]
```

3 Conclusion

We have proposed an additional link to the Medical Dialogue Ontology (MDO) designed to increase coverage and thereby improve translation outputs. We recognize the fact that medical dialogues, like dialogues in any other domain, converge around events. Knowledge of those events is a crucial part of domain knowledge. This is encoded in relations between entities and domain objects such as treatments, diagnoses, tests, symptoms, effects and the like. We have proposed a location within the existing MDO taxonomy for these "event-related" elements and a semantically "light" relation connecting them to other nodes.

We propose that the addition of these nodes and features will improve translation quality by improving coverage on the one hand and improving the selection of correct word-senses at the same time.

References

Francis Bond and Satoshi Shirai.1997. Practical and Efficient Organization of a Large Valency Dictionary. Workshop on Multilingual Information Processing, held in conjunction with NLPRS'97.

Anthony R. Davis and Leslie Barrett. 2002. Relations among Roles. In Proceedings of Ontolex 2002, 9-16 Las Palmas

Peter Gerstl and Simone Pribbenow. 1995. Midwinters, end games, and body parts: a classification of part-whole relations. Int. J. Hum.-Comput. Stud. 43(5-6): 865-889 (1995)

Fritz Lehmann. 1997. Big Posets of Participatings and Semantic Roles. In P. Eklund, G. Ellis and G. Ellis (eds.). *Conceptual Structures: Knowledge Representation as Interlingua*. Heidelberg: Springer

Beth Levin. 1993. English Verb Classes and Alternations: A Preliminary Investigation. University of Chicago Press, Chicago, IL

[9] Nor will the verbs be connected in any sub-hierarchy similar to the WordNet *troponym* relation.

[10] To save space the Head feature information is shown for English only

[11] For simplicity we will show only the "treatment" child node of Disease in place of the full range, and show one instance each of wordlists.

Lori Levin, Chad Langley, Alon Lavie, Donna Gates, Dorcas Wallace, and Kay Peterson. 2003. Domain Specific Speech Acts for Spoken Language Translation. IN Proceedings of the 4th SIGDIAL Workshop on Discourse and Dialogue 2003.

John F. Sowa. 2000. Knowledge Representation -- Logical, *Philosophical and Computational Foundations*. Brooks/Cole, 2000

Chris Welty and Nicola Guarino. 2001. Support for Ontological Analysis of Taxonomic Relationships. *J. Data and Knowledge Engineering.* **39**(1):51-74. October, 2001.

Populating ontologies in biomedicine and presenting their content using multilingual generation

V. Karkaletsis, A. Valarakos, C.D. Spyropoulos

Software & Knowledge Engineering Laboratory,
Institute of Informatics and Telecommunications,
National Centre for Scientific Research “Demokritos”,
P. Grigoriou & Neapoleos str., 15310 Aghia Paraskevi Attikis, Greece,
{vangelis, alexv, costass}@iit.demokritos.gr

Abstract

This paper discusses the use of ontologies for the representation and management of domain and language-specific knowledge in the field of biomedicine. It outlines a methodology for the semi-automatic population of domain knowledge from relevant corpora exploiting natural language processing and machine learning techniques and proposes the combination of ontology population with natural language generation techniques for rendering the content of the populated ontologies in different natural languages. The paper presents the results from populating a formally defined ontology on allergens acquiring instances from PubMed abstracts, and rendering its content in English.

1. Introduction

Ontologies are widely used for formalizing and organizing the knowledge of a particular domain of interest. This facilitates knowledge sharing and re-use by both people and systems. Ontologies are becoming increasingly important in the biomedical field since they enable knowledge sharing in a formal, homogeneous and unambiguous way. Knowledge in a rapidly growing field such as biomedicine is usually evolving and therefore an ontology maintenance process is required to keep ontological knowledge up-to-date. This paper outlines our methodology for building a formally defined ontology (using OWL[1], the emerging standard for specifying ontologies) and populating it exploiting natural language processing and machine learning techniques, domain specific corpora, and an ontology editor. The application of this methodology in the allergens domain is presented.

In this paper, we propose the combination of ontology population with natural language generation techniques for rendering the content of the populated ontologies in different natural languages. For this purpose, we exploit the M-PIRO authoring tool which is used for porting NLG technology to new domains. This tool provides an ontology editor which enables the creation and maintenance of ontologies as well as the import of existing OWL ontologies. The authoring tool enables also the creation and maintenance of language-specific resources for an ontology (lexicons, grammars), in order to be used by a natural language generation (NLG) engine for describing the ontology’s content in different languages. In addition to the use of the authoring tool in the population methodology, for creating the initial ontology and validating the acquired instances, we also propose its exploitation for presenting the contents of the evolving ontology using natural language descriptions in different languages.

We argue for the benefit of using OWL ontologies along with the domain-dependent linguistic resources that are necessary for NLG systems to produce textual descriptions about the ontologies. This would allow content (e.g., information about allergens) to be published on the Semantic Web in the form of OWL ontologies, with different NLG engines playing the role of browsers that would be responsible for rendering the content in different natural languages.

Section 2 of the paper discusses the use of ontologies in the field of biomedicine providing examples from the domain of allergens. Furthermore, it briefly presents related work on ontology population and natural language generation. Section 3 presents our methodology for populating ontologies and rendering the content of the populated ontology in different natural languages using multilingual generation. Section 4 presents the application of our methodology in the allergens domain. Finally, section 5 concludes summarising the current status of our work and presenting our future plans.

2. Related work

Ontologies represent the solution to the semantic and structural heterogeneity that appears in database schemata since they are able to provide a shareable, consistent and formal description of the semantics of the information source (Noy & Klein, 2004). Various biomedical communities have created several ontologies in order to address the interoperability problem between the various database applications (Baker et al. 1999) or to provide a common vocabulary and semantics (Gene Ontology; Schulze-Kremer, 1998; Giudicelli & Lefranc, 1999). In the context of the “Open Biological Ontologies” project[2], several biomedical ontologies or controlled vocabularies can be found. However, the knowledge in a domain ontology is usually evolving, especially in dynamic domains such as most of the domains in biomedicine. For example, new allergen names or variants of existing ones appear frequently in the literature, following or not the nomenclature.

Considering that the regular update of a domain ontology is crucial for its reliability and quality, the process of ontology maintenance is a necessity in the area of biomedicine. For instance, in the allergens domain, various databases and lists exist most of which are

[1] http://www.w3.org/TR/owl-ref/

[2] http://obo.sourceforge.net

available for free on the web. Their schemata are more or less similar, concentrating mostly to allergen's name, the species it occurs in and the protein associated with, along with its links to GenBank and SwissProt through their accession number. The main problem of these schemata is related to the differences occurring in the meaning of their categories (semantic heterogeneity) as well as their structure (structural heterogeneity). For example, some schemata use the term '*trivial name*' to refer to the allergen's common name and the term '*description*' to refer to the protein associated with, in contrast to others that use the term '*common name*' and '*biochemical id*', respectively. Also, some of the databases provide unstructured information (see ALLALLERGY[3]). Moreover, these schemata are ambiguous because they do not provide rigor definitions of the vocabulary uses, e.g. what is the meaning of source, whether it should be filled with allergen sources or proteins, etc. And finally, many of them are out-of-date since they are not updated regularly. The above problems motivated us to design and build a formally defined ontology for the allergens domain which would be machine exploitable and could be populated acquiring new instances from domain specific corpora (Valarakos et al. 2005, 2006).

Ontology building is not a trivial task and requires special attention in order to build a useful and machine exploitable ontology (Noy et al. 2004; Pinto & Martins 2004). This involves mainly the selection of concepts to be included in the ontology, the specification of concepts' properties and relations, addition of concepts' instances. Concerning ontology maintenance this mainly involves adding new instances (ontology population), as well as new concepts, properties and relations (ontology enrichment). The most recent approaches for ontology maintenance involve the use of machine learning techniques to identify regularities, which could lead to the identification of interesting concepts and relations. On the task of ontology population, most of the work that has been done is related to information extraction from unstructured natural language text or semi-structured HTML pages. The extent to which the text is structured determines the depth of the required linguistic analysis, in order to identify concept instances. A representative example of ontology population work is presented in (Craven et al. 2000). They state the need for constructing and maintaining knowledge bases with information coming form the Web and stress the need of formally storing information in knowledge bases which results in a more effective and intelligent information retrieval exploiting knowledge-based inference capabilities. A recent work on ontology population is the autonomous KnowItAll system (Etzioni et al. 2004) that incrementally extracts information from the web in an unsupervised way given only an initial ontology of a particular knowledge representation formalism. The works presented in (Vargas-Vera et al. 2002; Harith et al. 2003) are focusing mainly in the extraction of instances from textual corpora using information extraction systems. In these efforts, the training examples for the extraction systems are provided by manually annotating a corpus, whereas our approach relies on the automatic creation of the training examples exploiting an initial version of the domain ontology. In (Ciravegna et al. 2003), learning is enforced by integrating information from various structured sources, e.g. databases and digital libraries. A rule-based approach is adapted in (Kiryakov et al. 2003) aiming to tackle the problem as a named entity recognition task combining linguistic analysis and manually crafted rules to populate an ontology that contains many generally used entity types such as persons, companies etc.

Concerning Natural Language Generation (NLG), a strand of work has been devoted to the generation of textual descriptions of objects from symbolic information in ontologies and databases. An example of such work is ILEX (O'Donnell et al. 2001), which was demonstrated mostly in the museums domain, producing personalised English descriptions of exhibits. More recently, the M-PIRO project (Isard et al. 2003) developed a multilingual extension of ILEX, which has been tested in a variety of domains, including museum exhibits and items for sale. A major problem in this and many other NLG sub-areas is the difficulty of obtaining source symbolic information in forms compatible with the requirements of the language generators. This issue has mainly been addressed so far by extracting source information from structured and semi-structured data (Dale et al., 1998), and by developing authoring tools that help in the creation of source information and domain-dependent linguistic resources. Such tools were developed, for example, in DRAFTER (Hartley & Paris, 1997), ITRI's WYSIWYM systems (Van Deemter & Power, 2003), and M-PIRO (Androutsopoulos et al. 2002, 2006).

In recent years, considerable effort has been invested in the Semantic Web, which can be seen as an attempt to develop mechanisms that will allow computer applications to reason more easily about the semantics of the resources (documents, services, etc.) of the Web. A major target is the development of standard representation formalisms, that will allow ontologies to be published on the Web and be shared by different computer applications. The emerging standard for specifying ontologies is OWL, an extension of RDF. In NLG systems that describe objects, pre-existing OWL ontologies can provide much of the required source information, reducing the authoring effort and providing a common standard representation to generate from (Androutsopoulos et al. 2005).

3. Populating an ontology and rendering its content in different natural languages

This paper proposes the combined use of ontology population with natural language generation aiming at a common infrastructure that will enable, on one hand, the population of existing ontologies with knowledge acquired from domain-specific corpora, and on the other hand, the rendering of the content of the populated ontologies in different natural languages. This infrastructure exploits an existing methodology for ontology population which uses machine learning and natural language processing techniques (Valarakos et al. 2005, 2006). In addition, it exploits an existing methodology for authoring NLG applications for new domains (Androutsopoulos et al. 2002, 2006).

The idea for this combination followed after some initial experiments of how domain-dependent language resources could be best embedded in OWL ontologies. As noted in (Androutsopoulos et al. 2005), this embedding would lead to 'language-enabled' ontologies opening up

[3] http://www.allallergy.net/

another possibility for publishing content on the Semantic Web. However, ontologies are evolving and this must be the case also for 'language-enabled' ontologies. We believe it is worth trying to move towards this direction, as there are large potential gains for both the NLG community and the users of the emerging Semantic Web.

In the following sub-sections we present the main features of our technologies for ontology population and authoring NLG applications. The application of both these technologies in the field of biomedicine (allergens) is presented in section 4.

3.1. Ontology population

Our population methodology populates the ontology with new instances, as well as with their properties and relations, located in domain specific corpora. The initial ontology can be created manually using an ontology editor. Alternatively, in the case of existing ontologies or other resources (e.g. lexicons, thesauri) for a specific domain, these can be imported and updated in order to form the initial ontology to be populated.

The key idea behind our approach is that we can keep the instances of the domain ontology up-to-date in a semi-automatic way, by periodically re-training an information extraction system using a domain specific corpus. The methodology does not rely on a manually annotated corpus but uses the already known instances of the ontology to annotate the corpus. More specifically, our methodology involves the following processing stages (details can be found in (Valarakos et al. 2006)):

- Ontology-Based Annotation. This stage exploits the instances found in the initial ontology to automatically annotate the domain specific corpus. The ontology instances are fed to a lookup engine that finds all their occurrences in the corpus using regular expression patterns.
- Recognition and Classification of Instances. A named entity recognition and classification module is trained using machine learning techniques on the annotated corpus derived from the previous stage. The trained module is capable of recognizing new instances.
- Knowledge Refinement. A compression-based clustering algorithm is employed at this stage for identifying typographic variants of each instance.
- Extracting Properties and Relations. A shallow parser is used to extract the instances' properties and the relations that hold between instances. For this purpose, it uses a set of patterns that employ lexical, syntactic and semantic features. There are different patterns for each property and relation we try to extract.
- Validation and Insertion. At this stage the domain expert validates the extracted instances, properties and relations derived from the previous stages. He/she inserts then the validated information into the ontology. The outcome of this stage is a new version of the ontology containing knowledge acquired from the domain specific corpus.

A new iteration begins with the new version of the ontology (see Fig. 1). The iterative process will stop when no more changes in the ontology are possible.

3.2. Ontology content presentation

The NLG authoring tool was developed in the M-PIRO project (Calder et al. 2005; Isard et al. 2003), using the ILEX system (O'Donnell et al. 2001) as a starting point. In contrast to work on ILEX which had focused mostly on the generation of English descriptions, M-PIRO targeted multilingual generation, which required a clear separation of language-specific processes and resources from language independent ones; the system currently supports English, Italian, and Greek.

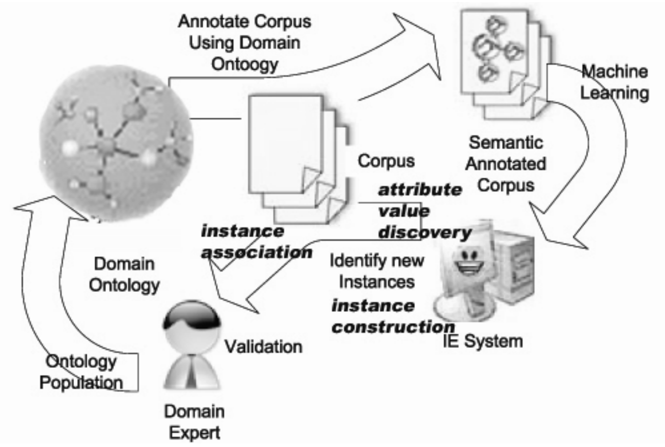


Figure 1: The stages of the ontology population methodology

M-PIRO's authoring tool allows the authors, i.e., the persons responsible for porting M-PIRO's technology to a new domain, to modify all the domain-dependent resources: the ontology, language resources, and the end-user stereotypes (these are used for tailoring the generated descriptions to the users' preferences and knowledge). M-PIRO generates texts from the ontology that encodes domain knowledge in the form of concepts, concepts' instances (entity types and entities correspondingly in M-PIRO's terminology), concepts' properties and relations between concepts. Properties and relations are expressed using fields. At any entity type, it is possible to introduce new fields, which then become available at all the entities that belong to that type and its subtypes.

M-PIRO relies on large-scale grammars, one for each supported language, to convert sentence specifications to surface text. These grammars can be treated as domain independent for M-PIRO's purposes. However, a part of the lexicon that the grammars employ has to be filled in by the authors when the system is ported to a new domain. The domain-dependent lexicon contains entries for nouns and verbs, and when moving to a new domain, the authors enter the base forms of the nouns and verbs they wish the system to use, and there are facilities to generate the other forms automatically.

For each field of the ontology and each language, the authors have to specify at least one template or clause (micro-plans in MPIRO's terminology) that specifies how the field can be expressed in that language. The author specifies the template or clause to be generated in abstract terms, by specifying, for example, the verb to be used, the voice and tense of the resulting clause, etc. The verb of the clause is specified by selecting a verb entry from the domain-dependent lexicon.

Much of the authoring effort when porting M-PIRO's technology to a new domain, has to be devoted to the definition of the available entity types (concepts) and the fields that express properties and relations. If a well-thought OWL ontology already exists for the specific domain, the authoring process can be accelerated by importing the ontology into the authoring tool. Thereafter, the authors can focus on populating the ontology with entities (instances) and adding the necessary domain-dependent linguistic resources (lexicon entries, micro-plans, etc.). For ontology population the authors can employ the methodology and tools presented in the previous sub-section. In addition, they can employ a functionality of the authoring tool to import instances automatically from data obtained from relational databases via (Androutsopoulos et al. 2005).

4. Application to the allergens domain

The ontology population methodology has been applied so far in the biomedicine field (allergens) and in the e-commerce field (laptops offers). In both cases, OWL ontologies were created using existing resources (databases and lists) and following well established design criteria aiming at the development of formally defined ontologies. In the allergens case, the ontology was built by two biologists and a knowledge engineer exploiting the IUIS allergen list[4] and documents that describe the allergen nomenclature.

Figure 2 illustrates the allergen domain ontology, ellipses stand for concepts whereas arrows denote concept relations.

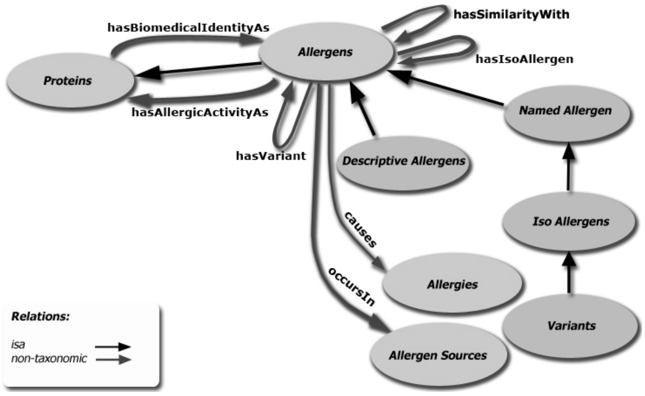


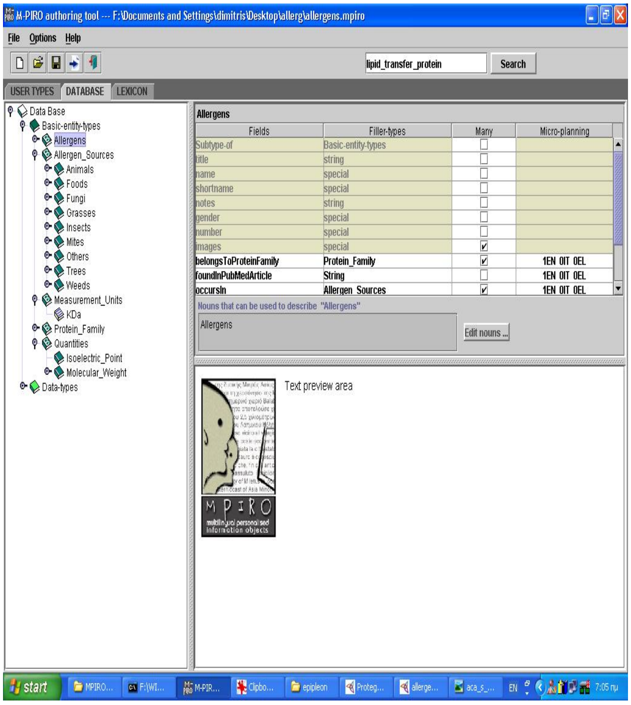
Figure 2: The allergens ontology

The properties for the four main types of concepts (allergens, proteins, allergen sources and allergies) along with the number of instances filling each property are depicted in the following table:

		Number of Instances
Allergens	Scientific name	*311*
	common name	*56*
	isoelectric point	*22*
	molecular weight	*171*
Allergen Sources	Scientific name	*194*
	common name	*175*
Proteins	Name	*185*
Allergies	Name	*54*
Total		*1168*

As it is noted in (Valarakos et al. 2006), starting from two different initial ontologies containing 15.59% and 34.94% of the "gold" ontology and a corpus of 279 PubMed abstracts on allergens, the coverage was increased to 68.3% and 81.9% respectively in two iterations. As a general remark, we can say that our approach presents a very good performance on locating separate instances filling concepts' properties. But, at the end, what is important, is whether it manages to fill the whole "template" correctly, that is whether it locates all the properties and relations found in the abstracts for each target allergen instance. For instance, the system may manage to locate correctly in a PubMed abstract the following instances: *Pen c 1* (scientific name of an allergen), *Penicillium citrinum* (scientific name of an allergen source), *33 KDa* (molecular weight of an allergen), *7.1* (isoelectric point of an allergen). The next step is to recognize that these instances can be grouped together to fill the whole "template" that represents the complete allergen instance.

We measured this in a second experiment (Valarakos et al. 2005), where we used a subset of the testing corpus containing 182 allergen instances. Our system managed to find correctly, either all or some of the properties and relations, in 168 out of the total 182 cases. In the 168 correct cases, the main problem was that the system didn't manage to locate all the information existing in the abstracts. This was due to the fact that our system, in its current state, extracts properties and relations only from single sentences. In case the information occurs in other sentences (for example, a sentence may provide input on an allergen's molecular weight without mentioning the allergen's name but just referring to it with a pronoun or other expression), this information is missed from the final "template", although it may be found in the previous processing stages.



lipid_transfer_protein

Allergens

Fields	Filler-types	Many	Micro-planning
Subtype-of	Basic-entity-types	☐	
title	string	☐	
name	special	☐	
shortname	special	☐	
notes	string	☐	
gender	special	☐	
number	special	☐	
images	special	☑	
belongsToProteinFamily	Protein_Family	☑	1EN 0IT 0EL
foundInPubMedArticle	String	☐	1EN 0IT 0EL
occursIn	Allergen_Sources	☑	1EN 0IT 0EL

Nouns that can be used to describe "Allergens"

Allergens

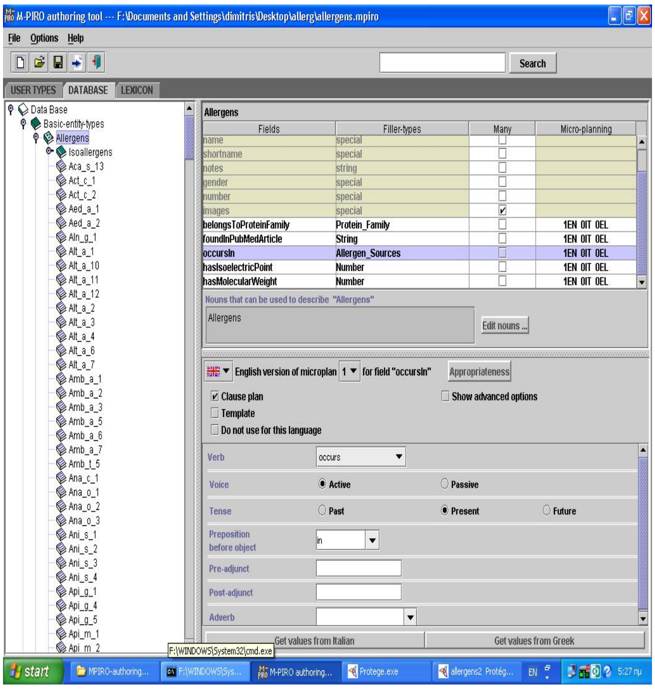
Figure 3: The allergens ontology imported in the authoring tool

[4] http://www.allergen.org/List.htm

Therefore, the major problem is related to the lack of a co-reference resolution module. This would enable us to include in the stage of properties and relations' extraction, also those sentences containing co-references to the instances of allergen names. We are currently working on improving the natural language processing stage by employing a co-reference analysis module in order to take into account missing information found in sentences containing co-reference to allergens names.

Furthermore, we are trying to improve the interaction with the domain expert through the use of an authoring tool which apart from the ontology editor will also provide additional functionalities. The M-PIRO authoring tool is used for this purpose, since it provides an ontology editor with import and export functionalities in OWL. Figure 3 depicts the allergens ontology using the authoring tool.

In order to create a 'language-enabled' ontology, we used the tool editors for adding language-specific resources. Nouns and verbs were added in the domain-specific lexicon to express the names of concepts, properties and relations in a natural language (only English currently). In addition, templates and clauses were created in order to specify how the properties and relations can be expressed in that language (see Figure 4).



Allergens

Fields	Filler-types	Many	Micro-planning
name	special	□	
shortname	special	□	
notes	string	□	
gender	special	□	
number	special	□	
images	special	☑	
belongsToProteinFamily	Protein_Family	□	1EN 0IT 0EL
foundInPubMedArticle	String	□	1EN 0IT 0EL
occursIn	Allergen_Sources	□	1EN 0IT 0EL
hasIsoelectricPoint	Number	□	1EN 0IT 0EL
hasMolecularWeight	Number	□	1EN 0IT 0EL

Nouns that can be used to describe "Allergens"

Allergens

Edit nouns ...

English version of microplan 1 for field "occursIn" Appropriateness

☑ Clause plan
□ Template
□ Do not use for this language
□ Show advanced options

Verb: occurs

Voice: ● Active ○ Passive

Tense: ○ Past ● Present ○ Future

Preposition before object: in

Pre-adjunct:

Post-adjunct:

Adverb:

Get values from Italian | Get values from Greek

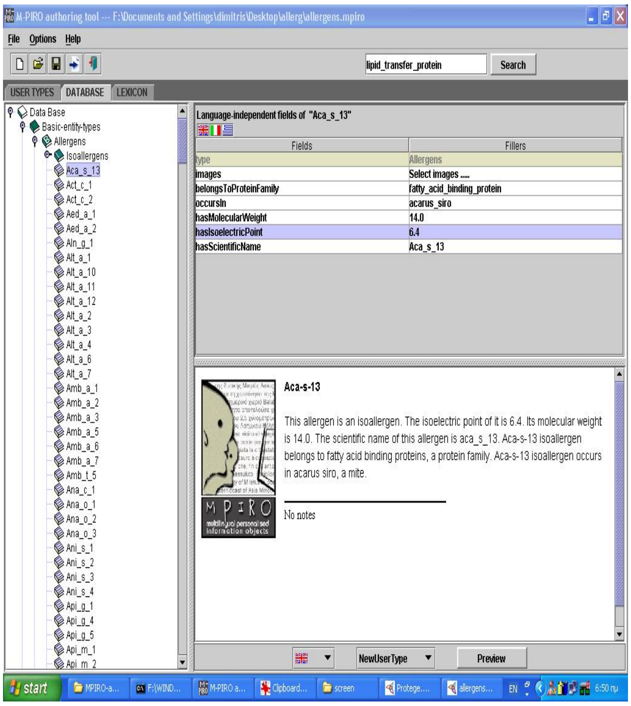
Figure 4: The 'language-enabled' allergens ontology: adding clauses

The resulting 'language-enabled' ontology can then be used by the natural language generation (NLG) engine for generating descriptions of the ontology's content in the supported language. Figure 5 depicts such a dynamically generated description.

More than one language can be supported by enriching the 'language-enabled' ontology with domain-specific lexicons, templates and clauses for each new language. Therefore, the authoring tool can support the multilingual presentation of the contents of the biomedical evolving ontologies, enforcing their publishing on the Semantic Web.

5. Concluding remarks

The paper discusses the use of ontologies in biomedicine and outlines the methods and tools we have used for developing and populating an ontology about allergens. Using existing resources about allergens we have created an initial formally defined ontology written in OWL, and populated it from PubMed corpora on allergens using an iterative ontology population methodology. The process of validating the acquired instances in each iteration can be performed using an ontology editor similar to the one provided by the M-PIRO authoring tool which enables the import and export of OWL ontologies.



Language-independent fields of "Aca_s_13"

Fields	Fillers
type	Allergens
images	Select images .....
belongsToProteinFamily	fatty_acid_binding_protein
occursIn	acarus_siro
hasMolecularWeight	14.0
hasIsoelectricPoint	6.4
hasScientificName	Aca_s_13

Aca-s-13

This allergen is an isoallergen. The isoelectric point of it is 6.4. Its molecular weight is 14.0. The scientific name of this allergen is aca_s_13. Aca-s-13 isoallergen belongs to fatty acid binding proteins, a protein family. Aca-s-13 isoallergen occurs in acarus siro, a mite.

No notes

Figure 5: The 'language-enabled' allergens ontology: generating descriptions

The added-value of the authoring tool is that it also enables us to add language specific resources (lexicons, grammars in the form of templates and clauses) in order to present the ontology's content in different languages. This would allow biomedical content to be published on the Semantic Web in the form of OWL ontologies, with different NLG engines playing the role of browsers that would be responsible for rendering the content in different natural languages.

We are currently working on improving the OWL-import functionalities of the authoring tool and testing it using the allergen ontology and resources for additional languages. Our next step is to implement the whole process of ontology creation, population and presentation in another domain of biomedicine.

6. Acknowledgements

This work is supported by the R&D project OntoSum which is partially funded by the Greek General Secretariat of Research & Technology (GSRT) under the PENED funding programme. OntoSum partners involve NCSR

"Demokritos" (Coordinator), University of the Aegean, EXODUS S.A., and Dartmouth College.

The authoring tool was initially developed by NCSR "Demokritos" in the context of the FP5-IST M-PIRO project. After the end of the project (Feb. 2003), the tool is being constantly updated by a team of researchers from NCSR "Demokritos" and the Athens University of Economics & Business.

7. References

Androutsopoulos I., D. Spiliotopoulos, K. Stamatakis, A. Dimitromanolaki, V. Karkaletsis, and C.D. Spyropoulos. (2002). Symbolic Authoring for Multilingual Natural Language Generation, In *Methods and Applications of Artificial Intelligence, LNAI 2308*, I.P. Vlahavas and C.D. Spyropoulos (eds), pp. 131-142.

Androutsopoulos I., S. Kallonis, V. Karkaletsis. (2005). Exploiting OWL Ontologies in the Multilingual Generation of Object Descriptions, *Proceedings of ENLG-2005*, Aberdeen, 08-10/08/2005.

Androutsopoulos I., J. Oberlander, V. Karkaletsis. (2006). Source Authoring for Multilingual Generation of Personalised Object Descriptions, *Natural Language Engineering*, Cambridge University Press, (to appear)

Baker P.G., C.A. Goble, S. Bechhofer, N.W. Paton, R. Stevens, and A Brass. (1999). An Ontology for Bioinformatics Applications, *Bioinformatics*, 15(6), pp. 510-520.

Calder J., A. Melengoglou, C. Callaway, E. Not, F. Pianesi, I. Androutsopoulos, C.D. Spyropoulos, G. Xydas, G. Kouroupetroglou and M. Roussou. (2005). "Multilingual personalized information objects". In Stock, O. and Zancanaro, M. (Eds.), *Multimodal Intelligent Information Presentation*, pp. 177-201, Springer.

Ciravegna F., Dingli A, Guthrie D, Wilks Y. (2003). Integrating Information to Bootstrap Information Extraction from Web Sites, *In Proceedings of the IJCAI Workshop on Information Integration on the Web*, pp. 9–14, 2003

Craven M., DiPasquo D, Freitag D et al. (2000). Learning to Construct Knowledge Bases from the World Wide Web. *Journal of Artificial Intelligence*, 118(1/2), pp. 69-113.

Dale R., S.J. Green, M. Milosavljevic, C. Paris, C. Verspoor, and S. Williams. (1998). Dynamic document delivery: generating natural language texts on demand. In Proc. of the *9th International Conference and Workshop on Database and Expert Systems Applications*, pp. 131–136, Vienna, Austria.

Etzioni O., Kok S, Soderland et al. (2004). Web-Scale Information Extraction in KnowItAll (Prelimenary Results), In: the *13th International World Wide Web conference (www2004)*, p. 100-110, New York.

Gene Ontology WWW resources: http://www.geneontology.org

Giudicelli V., M.-P. Lefranc. (1999). Ontology for immunogenetics: the IMGT-ONTOLOGY, *Bioinformatics*, 15(12), pp. 1047-1054.

Harith A., Sanghee K, Millard DE, Weal MJ, Hall W, Lewis PH, Shadbolt N. (2003). Web based knowledge extraction and consolidation for automatic ontology instantiation, *Workshop on Knowledge Markup and Semantic Annotation (KCap'03)*, Sanibel Island, Florida, USA.

Hartley A., and C. Paris. (1997). Multilingual document production – from support for translating to support for authoring. *Machine Translation*, 12(1–2), pp. 109–129.

Isard A., J. Oberlander, I. Androutsopoulos and C. Matheson. (2003). Speaking the Users' Languages. *IEEE Intelligent Systems*, 18(1), pp. 40-45.

Kiryakov A., Popov B, Ognyanoff D, Manov D, Kirilov A, Goranov M. (2003). Semantic Annotation, Indexing, and Retrieval, In *Proceedings of the 2nd International Semantic Web Conference (ISWC2003)*, Florida, USA, LNAI vol. 2870, pp. 484-499, Springer-Verlag.

Noy N. F., D.L. Rubin, M.A. Musen. (2005). Making Biomedical Ontologies and Ontology Repositories Work, *IEEE Intelligent Systems*, 19(6), pp. 78-81

Noy N.F. and M. Klein. (2004). Ontology Evolution: Not the Same as Schema Evolution, *Knowledge and Information Systems*, 6, pp. 428-440.

O'Donnell M., C. Mellish, J. Oberlander, and A. Knott. (2001). ILEX: an architecture for a dynamic hypertext generation system. *Natural Language Engineering*, 7(3), pp. 225–250.

Pinto H.S., J.P. Martins. (2004). Ontologies: How can they be built?, *Knowledge and Information Systems*, Springer-Verlag, 6, pp. 441-464.

Schulze-Kremer S. (1998). Ontologies for Molecular Biology. *Proceedings of the 3rd Pacific Symposium on Biocomputing*, Hawaii, Singapore, pp. 693-704.

Valarakos A., V. Karkaletsis, D. Alexopoulou, E. Papadimitriou, C.D. Spyropoulos. (2005). Populating an Allergens Ontology Using Natural Language Processing and Machine Learning Techniques, *Proceedings of AIME-2005*, LNAI, Issue 3581, pp. 264-273.

Valarakos A., V. Karkaletsis, D. Alexopoulou, E. Papadimitriou, C.D. Spyropoulos, and G. Vouros. (2006). Building an Allergens Ontology and Maintaining it using Machine Learning Techniques. *Computers in Biology and Medicine Journal* (to appear)

Van Deemter K., and R. Power (2003). High-level authoring of illustrated documents. *Natural Language Engineering*, 9(2), pp.101–126.

Vargas-Vera M., Domingue J, Lanzoni M, Stutt A, Ciravegna F. (2002). MnM: Ontology driven semi-automatic support for semantic markup. In: Asuncion Gomez-Perez, V. Richard Benjamins (eds.), *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02)*, 1-4 October 2002 - Siguenza (Spain), LNAI 2473, Springer Verlag.