Extension of a Specialised Lexicon Using Specific Terminological Data

Bruno Cartoni and Pierre Zweigenbaum
LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France

The paper describes methods for acquiring lexical information to implement a 'Unified Medical Lexicon for French' (UMLF) that aims at being a reference resource for NLP in the medical domain. We address four issues of lexical acquisition in a specialised domain. First, to assess the 'desired coverage' of lexical information, we use a large collection of French terms as a reference resource for the medical domain sublanguage. The collection contains close to 300,000 terms organised around conceptual identifiers. Second, by looking through this large amount of terminological data, we highlight the different kinds of information that might be useful to deal with typical terminological processing tasks, like variant recognition. The terminological variation phenomena that are very frequent in these terms are of three kinds: graphemic, inflectional and derivational variations. Third, we propose a model for organising the lexical information. Most of this model is inspired from existing specialist lexicons, but special emphasis is put on derivational morphological information. Finally, different kinds of acquisition methods are described, at the two levels of linguistic description that are addressed here: inflectional and derivational morphological knowledge. These methods allow acquiring an important amount of lexical data. For inflectional knowledge, the full paradigm is recorded, to provide information about all the possible inflected forms of lexical units within terms. Regarding derivational knowledge, specific derivation processes are targeted, in order to handle particular term variations. The relevance of the gathered derivational information is also assessed.

1. Introduction

Processing specialised languages requires specialised resources. In domains such as medicine, specialised lexicons are necessary to achieve typical Natural Language Processing (NLP) tasks, from POS-tagging to controlled indexing (Aronson 2001) and information extraction (Rindflesch et al. 2005). For the French language, the 'Unified Medical Lexicon for French' (UMLF) (Zweigenbaum et al. 2005) aims at being a reference resource for NLP in the medical domain.

In this project, which is a sub-project of the InterSTIS project[1], we start from the previous state of the UMLF lexicon and aim to provide it with a suitable coverage for the needs of the InterSTIS project. The notion of suitable coverage raises three main issues:

- How to determine the desired coverage?
- What kind of lexical information is useful?
- How to organise this information?

These three issues are related, and depend on the needs of the target applications of the lexicon. Once decisions are taken for these three questions, a fourth issue is raised, about the implementation methods that have to be put in place to acquire lexical information.

In this paper[2], we present the ongoing process of development of the UMLF lexicon. In Section 2, we provide a brief state of the art about existing specialised lexicons for medical

[1] This work was partly supported by project InterSTIS (ANR-07-TECSAN-010). The objective of the InterSTIS project is to develop a terminology server to support access to French-language medical terminology (http://www.interstis.org/).

[2] This paper focuses on the resources used to develop the lexicon and its organisation. For more technical details on the acquisition methodology, see (Cartoni and Zweigenbaum, forthcoming).

language, focusing on the way they are structured and recorded. In Section 3, we describe the way we determine the desired coverage, in order to assess the lexical knowledge that has to be gathered, and the specific information that has to be provided. This task is accomplished by gathering a large amount of terminological data from the medical domain. We then characterise the different linguistic phenomena that have to be described (Section 4). The next section presents the organisational strategy that has been taken under consideration to structure the lexicon (Section 5). Section 6 presents the data acquisition procedures that were used to acquire inflectional and derivational information, together with experiments on the validation of this information.

2. State of the Art: Existing Specialised Lexicons

Other specialised lexicons for the medical domain have already been conceived, namely for English and German. The English UMLS Specialist Lexicon (McCray et al. 1994) was the first of its kind. It consists in a large syntactic lexicon of biomedical and general English which gathers, in its last release[3], 432,822 base forms and 758,153 word forms. Each lexical entry gathers various syntactic, morphological and orthographic information such as spelling variants or inflectional variation.

The English Specialist Lexicon is distributed in two formats: a complete unit record and some specific relational tables. The complete unit record is a frame structure, where information is organised following a feature-attribute structure. All the information included in the frame structure is also expressed in relational tables (ten in total). The reason for creating these tables is that the lexicon can be used for different specific applications where only specific lexical information is needed. They are also convenient for loading into a relational database.

The 'German Specialist Lexicon' (Weske-Heck et al. 2002) was prepared to cover the words present in the German version of the International Classification of Diseases, and follows approximately the same organisation as its English counterpart. Due to German language specificities, more complex structures are provided to take into account inflectional information.

In the InterSTIS project, we took inspiration from these two Specialised Lexicons, and adapted some of their organisational structure to French language specificities. We particularly developed the derivational morphology side, as explained further in this article.

3. How to Determine Coverage?

A specialised lexicon for medical sub-language should typically be able to recognise (i.e. to analyse) all the terms of the domain. In such a domain, terms are made of lexical units that are not always part of the general language. Two kinds of sources can be used to determine the target list of words to be recorded in a specialised lexicon: textual corpora or sets of terms, both being large enough to be representative of the sub-language of interest. Since one of the main objectives of the InterSTIS project is the controlled indexing of textual documents with medical terms, the choice was made to use as a reference a *representative set of terms*.

[3] The English Specialist Lexicon is available at http://lexsrv3.nlm.nih.gov/SPECIALIST/Projects/lexicon/current/index.html [access date: 28/02/2010].

To obtain an extensive and representative set of terms from the French medical language, we compiled a list of terms (hereafter, the Term-Union) from various medical terminologies (thesauri, classifications, nomenclatures). Table 1 presents the terminologies that have been used, together with statistics on their numbers of terms.

Terminological database	**Number of terms**
ICD-10: *International Statistical Classification of Diseases and Related Health Problems, 10th revision,* French Translation (CIM10-FRE)	10,800
ICPC: *International Classification of Primary Care, French Translation* (ICPCFRE)	723
Medical Dictionary for Regulatory Activities Terminology (MedDRA), French Edition (MDRFRE)	67,784
Thésaurus Biomédical Français/Anglais, French translation of the *Medical Subject Headings (MeSH)* (MSHFRE)	76,295
French version of the *Minimal Standard Terminology of the European Society of Gastrointestinal Endoscopy* (MTHMSTFRE)	1,833
French version of SNOMED v3.5: *Systematized Nomenclature of Human and Veterinary Medicine* (SNMIGIPFRE)	150,410
World Health Organization Adverse Drug Reaction Terminology (WHOART), French Translation (WHOFRE)	3,673
Total	**311,518**

Table 1. Terminology database in Term-Union

More precisely, the terms present interesting characteristics, as shown in Table 2: each term is associated with a 'Concept Unique Identifier' (CUI), which comes from the UMLS Metathesaurus[4].

Number of Terms	311,518
Number of Unique Concepts (CUI)[5]	154,594
Number of concepts associated to more than one term	68,118

Table 2. Terms in Term-Union

Table 3 provides statistics about the number of lexical units in terms. It should be noticed that only a very small proportion of terms are made of one single term.

The particularly large amount of terms gathered in Term-Union is an important source of information and tends to provide an interesting picture of the different phenomena that the targeted lexicon should cover. By their very nature, terminologies are dense in specialised terms and therefore are highly relevant for the purpose of designing a specialised lexicon. On the other hand, free-text corpora may better represent current practice of medical language and may include terms which are absent from medical terminologies. However, the concept

[4] The UMLS Metathesaurus (Bodenreider 2004) is the union of more than one hundred biomedical terminologies, most of which are in English. It associates each term of these terminologies to a Metathesaurus concept; terms with the same meaning are mapped to the same Metathesaurus concept.

[5] 10,727 terms have no unique identifier.

orientation of a terminology such as Term-Union, with its organisation around CUIs, provides a direct handle on terminological variants, which can thus be analysed with confidence.

Mean term length (in lexical units)	3.81
Median	3
Terms with one single lexical unit	46,497
Number of lexical units	94,964

Table 3. Statistics on terms in Term-Union

Figure 1 contains an excerpt of Term-Union, presenting three CUIs and the different associated terms and information. The first column is the CUI, the second is the source terminology and the third is a terminology-specific identifier.

C0001399	MSHFRE	D000221	Adaptation de l'oeil
C0001399	MSHFRE	D000221	Adaptation des yeux
C0001399	MSHFRE	D000221	Adaptation oculaire
C0001399	SNMIGIPFRE	F-F0010	vision diurne
C0001399	SNMIGIPFRE	F-F0010	vision photopique
…			
C0001552	MSHFRE	D000277	Adjuvant médicamenteux
C0001552	MSHFRE	D000277	Adjuvants pharmaceutiques
…			
C0002019	CIM10FRE	R48.0	Dyslexie et alexie
C0002019	MDRFRE	10001664	Alexie et dyslexie

Figure 1. Excerpt from Term-Union

All the experiments described in this paper are based on this extended list of terms, used as the reference of the specificities of the French medical language. Further investigations are also planned with free-text corpora.

4. What Kind of Lexical Information is Useful?

When lexical resources are built for specific purposes and specialised vocabularies, it is important to have a clear idea of the kind of information (or lexical knowledge) that will be useful for the targeted task. The target lexicon will be used in future NLP projects on medical terminology. A large proportion of medical NLP works targets the recognition of these terms and their variants in indexing or information retrieval applications. To be able to process all these variants, the specialised lexicon should contain relevant information. Besides, a full lexical entry may include detailed information at each of the traditional levels of linguistic description: phonology, morphology, syntax, semantics, etc. Again, the needs of the target applications should be taken into account to determine which subset is really needed.

Browsing through the Term-Union allowed us to highlight interesting characteristics and to identify different variations. In this project, three types of variations have been identified and three corresponding aspects of lexical knowledge are targeted.

4.1. Graphical Variation

Spelling of highly specialised terms is sometimes flexible, as can be observed in Term-Union. For instance, Example (1) shows two graphemic variants found in the French MeSH Thesaurus (INSERM 2009).

Example (1)

équilibre acido-basique
équilibre acidobasique
[EN: acid-base balance]

In the Term-Union, 1,593 word-forms are recorded with and without a hyphen, and many other graphemic variations are observed, such as capitalisation. Term capitalisation can sometimes be meaningful, as in the names of animal species, but sometimes it is only a graphical convention of a particular terminological resource. The lexicon has to be able to address these variants, i.e. to recognise any graphemic variant of the same lexeme, whenever it is meaningful.

4.2. Inflectional Variation

Inflectional knowledge is important to assign each lexical item categorical and morphosyntactic information, together with its lemma. As shown in Example (2), both plural and singular forms of the same term can be found in Term-Union.

Example (2)

adaptation de l'oeil
adaptation des yeux
[EN: eye adaptation]

This variation is very frequent in corpus, and the lexicon has to be able to provide relevant information to recognise the plural form of a term recorded in singular form. Apart from usual grammatical words (preposition and determiner) the most common parts-of-speech in the terms of Term-Union are adjective and noun, and therefore the acquisition process (see Section 6) is focused on these two categories.

4.3. Derivational Variation

Derivational knowledge is particularly useful in medical terminology, because one term can have many 'morphosemantic' variants[6], as in Example (3), where both terms are recorded with the same CUI in Term-Union.

Example (3)

intoxication à l'alcool
intoxication alcoolique
[EN: alcohol intoxication]

Automatically linking *alcoolique* and *alcool* [EN: *alcoholic* and *alcohol*] through morphological analysis is an important asset that can also be implemented in the lexicon. Many different kinds of morphological variations can be implemented, and the main issue is to determine which derivational information is most relevant to deal with terminological variation.

[6] Morphology terminology is not very fixed. Here, we use 'morphosemantic' to describe all the morphological knowledge that is not 'morphosyntactic', i.e. derivation, construction, … even though in this specific project, only derivational processes are addressed.

In the InterSTIS project, we mainly focus on the relational adjectives and their links with nouns. Relational adjectives are said to be very frequent in specialised domains (L'Homme, 2004) and are very often used to catch morphological variants in term extraction tasks (Daille, 1999). Relational adjectives are derived from nouns; they designate a relation between *(i)* the entity denoted by the noun they are derived from and *(ii)* the entity denoted by the noun they modify (Fradin 2007). In a noun phrase such as *muscle abdominal* [EN: abdominal muscle], the adjective *abdominal* designates the relation between the head noun (*muscle)* and the base-noun of the adjective: *abdomen.* From this relation comes also the fact that the same concept can be expressed by a prepositional phrase (*muscle de l'abdomen* [EN: muscle of abdomen]), which is often the case in terminological variation.

Another related phenomenon is the prefixation of relational adjectives. When prefixed, relational adjectives are the formal base of the adjectives, but on the semantic side, the prefixation rule applies to the nominal base. For example, in a prefixed adjective such as *anticancéreux* [EN: anticancerous], prefixation in *anti* actually applies to the base noun *cancer*, and can be paraphrased as *against cancer*. Consequently, it is interesting to gather information linking prefixed relational adjectives to base-nouns.

In the same line, we also take into account a specific link that seems to be very relevant in medical variation: the link between deverbal adjective and deverbal noun, as in the two term variants in Example (4).

Example (4)

comportement agressif
comportement d'agression
[EN: aggressive behaviour / behaviour of aggression]

The two lexemes (*agressif* and *agression*) are coined on the same base-verb (*agresser* [EN: to assault]); their link is consequently recorded in a specific way.

All these links are very often said to be specific to terminological variants in specialised domains. But other morphological relations would be considered, like deverbal nouns (with their relation to the verbs) or qualifier adjectives (with their relations to the quality noun).

Beyond graphemic, morphosyntactic and morphosemantic knowledge, many other kinds of lexical knowledge can be implemented (e.g. semantic information such as synonymy or hypernymy), but the priority was set on the above-described variations, since they were noted as very frequent in Term-Union.

5. How to Organise Lexical Information?

To address all the above-listed features, the desired lexicon should contain descriptions at three different levels. Following what was done for similar lexicons in other languages (cf. Section 2), all this information is represented in specific relational tables that can be easily gathered in a database or compiled into a structured data file following appropriate guidelines.

5.1. General Organisation

In terms of general modelling and formatting for information exchange, a standard framework such as the Lexical Markup Framework (LMF[7]) is fully appropriate to describe all the kinds of information that need to be recorded.

At the time of writing, only prospective work was carried out to build a full lexical unit record. Figure 2 provides only an example of the model we intend to implement (entry indexes are not numbered yet), inspired by the UMLS Specialist Lexicon.

```
{base=abdominal
 entry=XXX
        cat=adj
        gvariant=reg
        inflection_pattern=3
        rel_adj_of=abdomen|noun|YYY
}
```

Figure 2. Example of a full lexical entry record

In this example entry for the adjective *abdominal*, a selected number of fields are presented. First, the POS is given (cat=adj) followed by the gvariant field (*graphemic variant*, here, nothing to be recorded, hence regular). Then, instructions about the inflectional pattern are recorded, and a morphosemantic instruction is given, providing the link to the base noun of this relational adjective (*adjective rel_adj_of=abdomen|noun|YYY*).

These unit records are built by compiling information contained in the specific relation tables described below.

5.2. Organisation of Graphical and Inflectional Knowledge

The different spellings of the lexeme are listed, and linked with the variant that is considered to be the 'reference'. For example, if a hyphenated word is also found without a hyphen, the two forms are listed in an appropriate table, as shown in Figure 3.

…	
laryngo-pharyngé	laryngopharyngé
recto-colite	rectocolite
micro-dosées	microdosées
intra-osseuse	intraosseuse
…	

Figure 3. Excerpt of relational table for graphemic variants

For inflection, the full inflectional paradigm is provided for each lexeme, with necessary information on the lemma and on the POS-tag, following the Grace/Multext format [8], as shown in Figure 4.

sérofibrineux	sérofibrineux	Afpms
sérofibrineuse	sérofibrineux	Afpfs
sérofibrineux	sérofibrineux	Afpmp
sérofibrineuses	sérofibrineux	Afpfp

Figure 4. Excerpt of relational table for inflectional paradigms

[7] http://www.lexicalmarkupframework.org [access date 23/02/2010].

[8] http://aune.lpl.univ-aix.fr/projects/multext/ [access date 23/02/2010].

5.3. Organisation of Derivational Knowledge

The derivational lexicon also contains relational tables that provide morphological information for complex words. Each table represents a specific morphological link between a derived lexeme of a particular category and its base lexeme.

For example, a relational table (shown in Figure 5) provides information for relational adjectives and their base nouns, like *abdominal* and *abdomen*, as explained in Section 4.3.

…	
abdominal	abdomen
aplasique	aplasie
appendiculaire	appendicule
arachnéphobique	arachnéphobie
arachnoïdien	arachnoïde
argentique	argent
…	

Figure 5. Excerpt of relational table for relational adjectives

Other relational tables provide information for prefixed relational adjectives. In this case, each table is specific to one particular morphosemantic relation. For instance, Table 'anti_adj_noun' (see Figure 6) provides information for the link between the adjective *anticancéreux* [EN: anticancerous] and the noun *cancer*, as previously explained in Section 4.3.

…	
anticancéreux	cancer
antilymphocytaire	lymphocyte
antimalarique	malaria
antimicrobien	microbe
antimigraineux	migraine
anti-mitochondrie	mitochondrie
antiasthmatique	asthme
…	

Figure 6. Excerpt of relational table for prefixation in anti- of relational adjectives

A third example of a relational table is the one presented in Figure 7 for the relation between deverbal adjectives and deverbal nouns sharing the same verb base.

…
implosif|implosion
incisif|incision
inclusif|inclusion
intégratif|intégration
intensif|intension
invasif|invasion
…

Figure 7. Excerpt of relational table for deverbal adjectives and deverbal nouns

Other relations are currently under consideration, for instance deverbal nouns and their relation to their base-verbs.

6. Acquisition of Lexical Information

6.1. Initial State of the UMLF

A first version of the UMLF lexicon was produced in the UMLF project[9] by gathering lexical entries from lexicons of the project partners, with a focus on the lexical database compiled at HUG (Hôpitaux universitaires de Genève) (Baud et al., 1998). This lexicon contained 17,192 lexical units (5,353 adjectives and 11,799 nouns), together with their complete inflectional paradigms (36,965 word forms).

In order to evaluate the coverage of the UMLF, i.e. its lexical completeness, we confronted it with the Term-Union. The confrontation was performed on single words after case folding. Out of the 94,964 word-forms from the Term-Union, 81,595 forms were unknown from the UMLF in its initial state. Consequently, the acquisition of inflectional knowledge focuses on these remaining word-forms.

6.2. Acquisition of General Lexicon Knowledge

In any specialised language, some of the terms may be composed of lexical units that are common to the general lexicon. Although these lexical units might have a special linguistic behaviour, their morphosyntactic characteristics are generally identical in both specialised and general languages. Consequently, the first obvious step is to obtain inflectional knowledge from a general lexicon.

To perform this task, we used the general, large-coverage French lexicon Morphalou, a free lexicon[10] which contains 67,376 lemmas and 524,725 word-forms. Out of the 81,595 word-forms unknown from the UMLF, only 6,617 word-forms were found in Morphalou. These forms were consequently added to the UMLF, together with the rest of their inflectional paradigm.

These figures show that the 78,978 forms that remain unknown from Morphalou are specific to the medical domain. They represent around 80% of the number of lexical units within the Term-Union, which emphasizes the specificity of the vocabulary in the terms included in Term-Union.

6.3. Acquisition of Inflectional Information

As previously mentioned, recording inflectional information is a key issue to detect different inflectional variants of the same term. For this specific task, we used a learning algorithm that is based on the frequent endings of existing recorded lexical entries. This method implements the algorithm of Tanguy and Hathout (2007: 295) to learn from a reference lexicon the association between the full tag of a word-form (POS plus gender and number information) and the form of this word. The algorithm then guesses the possible tag(s) of each unknown word. The learning phase of the program is based on the endings (from the longest to the smallest) of the different entries of the reference lexicon. The program computes the most frequent tag of each final character string. Then, for each unknown word from the Term-Union, the program proposes one or more tag(s), according to what it has learned from the reference lexicon. To increase the quality of the guessing (and to avoid a manual verification), two different reference lexicons are used to perform the task, and the two results are

[9] http://www-test.biomath.jussieu.fr/umlf/ [access date 01/03/2010].

[10] http://www.cnrtl.fr/lexiques/morphalou/ [access date 23/02/2010].

compared: precision should be better when both results agree. The first one is the general lexicon Morphalou (cf. Section 6.2 above) and the second one is the UMLF itself (in its initial version). In total, 30,137 word-forms have been analysed the same way based on the two reference lexicons.

This cross-validation ensures a good quality for the guessing. Indeed, an evaluation based on a sample of 1,000 entries shows that only 82 (8.2%) were wrongly labelled. An error analysis shows that only 12 were actual labelling errors (e.g. 'accidentellement', an adverb, was labelled as a noun—since there is no adverb in the two reference lexicons—and 'kascher' was labelled as a noun instead of an adjective). Proper names are the main source of mistakes since their endings are not predictable. They represent 59.7% of the errors, and could be excluded easily in a preprocessing step (e.g. by using a specific resource such as that described in (Bodenreider and Zweigenbaum, 2000)). Other errors are Latin words, which should also be addressed in a preprocessing step by using dedicated resources. We can assume that with appropriate preprocessing to exclude lexical units that are resistant to 'ending guessing', the process is efficient enough.

6.4. Acquisition of Morphosemantic Knowledge

To acquire morphosemantic knowledge, i.e. morphological links between complex lexeme and base lexeme, we mainly made use of a morphological analyser of French: DériF (Namer 2009). DériF is a rule-based morphological analyser that describes word-construction processes. Given as input a word token (provided with its POS tag), DériF provides a complete analysis of that word, i.e. constructional information (the rule that coins the word, the involved affix and the base-lexeme) and a gloss which provides information of meaning.

For this experiment, we provided DériF with a list of adjectives (acquired by the morphosyntactic acquisition procedure, cf. Section 6.3 above) and filtered its ouput to obtain a simple list of 'complex lexemes −base -lexeme', as explained in Section 5.3. Manual checking was also performed to ensure good quality of the data.

Extra information was gathered from existing resources, such as the resources provided by Memodata[11] which are lexical resources that contain interesting lexical relations such as synonymy and derivation link.

In total, as a first step of resources construction, we compiled 3 relational tables for the adjectives, as shown in Table 4.

Relational table	**Example**	**Number of entries**
AdjRel – N (X – Xsfx)	abdominal|abdomen	2091
PrefAdjRel – N (PrefXsfx – X)	antiasthmatique|asthme	864
AdjDeverb – NDeverb (Xif – Xion)	auditif|audition	124

Table 4. Relational tables for adjectives

[11] Memodata is one of the partners of the InterSTIS project: http://www.memodata.com/ [access date 02/25/2010].

As previously mentioned, the PrefAdjRel list is in fact split into sub-tables according to the meaning of the prefixation process (e.g. one relational table is set up for prefixation in *anti,* another one for prefixation of uniqueness – *mono, uni*, ...).

6.5 Validating the Relevance of the Morphosemantic Relation Table

Building the relation tables as explained above relies on assumptions drawn from the observation of term variation in Term-Union. A second experiment was performed to confirm the importance of such relational tables in looking for terminological variants. We looked in Term-Union for term variants that actually contained both members of a pair listed in these tables. Table 5 summarises the results of this experiment. For each list, the number of entries recorded is provided, together with the number of terms variants (types and tokens) found in Term-Union. Table 6 provides examples of term variation detected by each list.

Relational table	Terms variants (tok.)	Terms variants (type)	Number of entries
X – Xsfx	1409	192	2091
PrefXsfx – X	2	1	864
Xif – Xion	66	9	124

Table 5. Term variants extracted thanks to the relational tables

Surprisingly, the PrefAdjRel – N relation, even though it is often described, is not frequent in Term-Union. Of course, Term-Union does not contain all the possible variants, and maybe not this specific kind of variation, which might be more frequent in free-text.

Relational table	Examples
X – Xsfx	Aberrations **autosomiques** Anomalies des **autosomes** Anémie **aplasique** anémie par **aplasie** Tumeurs de la **glande** péri-anale Tumeurs **glandulaires** périanales
PrefXsfx – X	Calcul dans la **glande** salivaire Calculs salivaires **intraglandulaires**
Xif – Xion	Facteur d'**inhibition** de la prolactine Facteur **inhibiteur** de la libération de prolactine Abaissement du seuil de **convulsions** Seuil **convulsif** abaissé Syndrome pulmonaire **obstructif** Syndrome d'**obstruction** pulmonaire chronique

Table 6. Examples of term variations extracted with the relational tables

The PrefXsfx variation look-up was performed with a script based on character strings. It highlights another interesting terminological variation, between prefixed relational adjective and prefixed noun. In these cases, prefixed nouns are used as adjectives, as shown in Examples (5) and (6).

Example (5)

*Immunoglobuline anti**lymphocytaire***
*Immunoglobulines anti-**lymphocytes***
[EN : antilymphocyte immunoglobulin]

Example (6)

*Anticorps anti**nucléaire***
*anticorps anti-**noyaux***
[EN :antinuclear antibody**]**

This phenomenon is interesting to consider, and probably deserves to be recorded in another relational table.

This last experiment shows that empirical study is very interesting to validate the relevance of lexical morphological data. Of course, this kind of experiment depends greatly on the data on which it is performed. Corpus-based terminological variant extraction might give different results. Nonetheless, these results remain an interesting motivation for pruning lexical morphology data in a lexical knowledge representation context.

7. Concluding Remarks and Further Work

In this article, we presented the state of development of a specialised French lexicon for the medical domain. We described the specific information that is needed, and the different procedures for acquiring specialised lexical knowledge that have been investigated.

As a reference resource, we use a large terminological database that is concept-oriented. Its orientation and its importance (more than 300,000 terms) provide an interesting material to characterise and test the specific phenomena that need to be described.

Three kinds of knowledge are targeted for the French UMLF lexicon: graphemic, inflectional and derivational. While graphemic variation knowledge is simply gathered from investigation amongst Term-Union words, morphological knowledge acquisition requires specific methods. We presented the method for the acquisition of inflectional information that is based on frequent ending. Another experiment is under process, using machine learning techniques, and more specifically the Conditional Random Fields (CRF) model (Lafferty et al. 2001). This model is appropriate to learn the full morphosyntactic tag of the lexical unit, because it can take advantage of the regular and limited morphosyntactic patterns followed by terms. First results appear to be very encouraging.

Derivational knowledge requires yet different methods and tools. We used a morphological analyser and some manual verification to acquire these data, and to organise them in a specific format. These data are shaped as static knowledge (compared to a dynamic morphological analysis) and therefore their relevance should be tested. The simple experiment we presented, consisting in confronting recorded lexical relations with term variants, allows assessing the relevance of the recorded resource, and highlighted some other interesting variations.

Nowadays, acquisition of lexical knowledge is becoming less resource- and time-consuming, so the question is less about how to gather information, than which information to record. The question of the relevance of the data is more than ever crucial.

References

Aronson, A.R. (2001). 'Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program'. In *Journal of the American Medical Informatics Association* 8 (suppl.). 17–21.

Baud, R.H.; Lovis, C.; Rassinoux, A.-M.; Michel, P.-A.; Scherrer, J.-R. (1998). 'Automatic extraction of linguistic knowledge from an international classification'. In Cesnik, B.; Safran, C.; Degoulet, P. (eds.). *Proceedings of the 9th World Congress on Medical Informatics*. Seoul. 581–585.

Bodenreider, O. (2004). 'The Unified Medical Language System (UMLS): integrating biomedical terminology'. In *Nucleic Acids Research* 32 (Database issue). D267-270.

Bodenreider, O.; Zweigenbaum, P. (2000). 'Stratégies d'identification de noms propres à partir de nomenclatures médicales parallèles'. In *Traitement automatique des langues* 41 (3). 727–757.

Cartoni, B. ; Zweigenbaum, P. (forthcoming). 'Semi-Automated Extension of a Specialized Medial Lexicon for French'. In *Proceedings of LREC 2010*. Malta.

Daille, B. (1999). 'Identification des adjectifs relationnels en corpus'. In *Proceedings of TALN'99*, Cargèse. 105-114.

Fradin, B. (2007). 'On the semantics of denominal adjectives'. In Ralli, A.; Booij, G.; Scalise, S. (eds.). *Proceedings of the 6th Mediterranean Morphology Meeting*. University of Patras.

INSERM. (2009). *Thésaurus Biomédical Français/Anglais. Institut National de la Santé et de la Recherche Médicale*. Paris.

Jacquemin, C.; Tzoukermann, E. (1999). 'NLP for term variant extraction: A synergy of morphology, lexicon, and syntax'. in T. Strzalkowski (ed.). *Natural Language Information Retrieval*. Boston: Kluwer. MA.Ch. 2, 25–74.

Lafferty, J.; McCallum; A.; Pereira, F. (2001). 'Conditional random fields: Probabilistic models for segmenting and labeling sequence data'. In *Proceedings of ICML-01*. 282–289.

L'Homme, M.C. (2004). 'Adjectifs dérivés sémantiques (ADS) dans la structuration des terminologies'. In *Actes. Terminologie, ontologie et représentation des connaissances, Université Jean-Moulin Lyon-3, 22-23 janvier 2004.*

McCray, A.T.; Srinivasan, S.; Browne, A.C. (1994). 'Lexical methods for managing variation in biomedical terminologies'. In *Proceedings of the 18th Annual SCAMC*, Washington: Mc Graw Hill. 235–239.

Namer, F. (2009). *Morphologie, lexique et traitement automatique des langues: l'analyseur DériF*. Paris: Hermès-Lavoisier.

Namer, F.; Zweigenbaum, P. (2004). 'Acquiring meaning for French medical terminology: contribution of morphosemantics'. In Fieschi, M.; Coiera, E.; Li, Y.C.L. (eds.). *10th World Congress on Medical Informatics, volume 107 of Studies in Health Technology and Informatics*. Amsterdam: IOS Press. 535–539,

Rindflesch, T.C.; Fiszman, M.; Libbus, B. (2005). 'Semantic interpretation for the biomedical literature'. In: Chen, H.; Fuller, S.; Hersh, W.; Friedman, C; (eds.). *Medical informatics: Advances in knowledge management and data mining in biomedicine*. Berlin/Heidelberg: Springer. 399–422.

Tanguy, L.; Hathout, N. (2007). *Perl pour les linguistes*. Paris: Hermes-Sciences Lavoisier.

Weske-Heck, G.; Zaiß, A.; Zabel, M.; Schulz, S.; Giere, W.; Schopen, M.; Klar, R. (2002). 'The German Specialist Lexicon'. In *Journal of the American Medical Informatics Association* 8 (suppl.).

Zweigenbaum, P.; Baud, R.H.; Burgun, A.; Namer, F.; Jarrousse, E.; Grabar, N.; Ruch, P.; Le Duff, F.; Forget, J.-F.; Douyère, M.; Darmoni, S. (2005). 'A unified medical lexicon for French'. In *International Journal of Medical Informatics* 74 (2–4). 119–124.