Semi-Automated Extension of a Specialized Medical Lexicon for French

Bruno Cartoni and Pierre Zweigenbaum

LIMSI-CNRS
BP 133, 91403 Orsay Cedex, France
bruno.cartoni@limsi.fr, pierre.zweigenbaum@limsi.fr

Abstract

This paper describes the development of a specialized lexical resource for a specialized domain, namely medicine. Based on the observation of a large collection of terms, we highlight the specificities that such a lexicon should take into account, and we show that general resources lack a large part of the words needed to process specialized language. We describe an experiment to feed semi-automatically a medical lexicon and populate it with inflectional information, which increased its coverage of the target vocabulary from 14.1% to 25.7%.

1. Introduction

Processing specialized languages requires specialized resources. Therefore, in domains such as medicine, specialized 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). In English, the UMLS Specialist Lexicon (McCray et al., 1994) is a large syntactic lexicon of biomedical and general English which gathers, in its last release[1], 432,822 base forms and 758,153 word forms. A "German Specialist Lexicon" (Weske-Heck et al., 2002) was also prepared to cover the words present in the German version of the International Classification of Diseases. 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.

The InterSTIS project[2] develops a terminology server whose goal is to provide access to the major French-language medical terminologies, together with controlled indexing methods. To let these methods take advantage of lexical information, a sub-goal of the project is to obtain a suitable coverage of the UMLF lexicon[3]. This raises issues of how to determine the desired coverage and which lexical information is useful in this context. These are key issues, since they will set the target for evaluating progress and results, and they may influence the kinds of methods which will be needed. Actual needs must then be assessed with respect to the coverage of existing lexicons. Methods must finally be found to increase coverage toward the target objective.

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

[2]This work was partially funded by project InterSTIS (ANR-07-TECSAN-010).

[3]More details on the general approach and on the desired targeted information are provided in (Cartoni and Zweigenbaum, 2010).

2. Extending a specialized lexicon

2.1. Coverage and contents

2.1.1. How to determine coverage

A specialized lexicon for medical sub-language should typically be able to recognize (i.e. to analyze) all the terms of the domain. In such a domain, terms are usually made of lexical units that are not always part of the general language. Two kinds of sources can be used to determine the desired coverage of the lexicon: a corpus or a terminology, both being representative of the sub-language of interest. The InterSTIS project is focused on medical terminologies and on their use in controlled indexing of textual documents. Therefore, a set of terms of all French-language medical terminologies (thesauri, classifications, nomenclatures) is the core of our target. We thus compiled the list of terms (henceforth, the *Term-Union*) contained in the major French-language terminologies of interest to InterSTIS: MeSH, SNOMED v3.5, ICD-10, MedDRA, etc.

The Term-Union contains 311,518 distinct terms linked to 203,300 unique concepts. Each term is linked to a conceptual representation: a Concept Unique Identifier (or CUI) ; a single concept can be expressed by more than one term. All the experiments described in this paper are based on this extended list of terms. Here below, we provide an excerpt of Term-Union. The first column is the CUI, the second is the source terminology and the third is a terminology-specific classifier.

```
C0000733 MSHF D00007 Traumatisme abdominal
C0000733 MSHF D00007 Traumatismes abdominaux
C0000733 MSHF D00007 Traumatismes de l'abdomen
...
C00001558 MSHF D00279 Voie cutanée
C00001558 MSHF D00279 Voie intradermique
C00001558 MSHF D00279 Voie percutanée
C00001558 MSHF D00279 Voie transcutanée
```

2.1.2. Target lexical information

When lexical resources are built for a specific purpose, it is important to have a clear idea of the kind of information (or lexical knowledge) that will be useful for the targeted task. A full lexical entry may include detailed information at each of the traditional levels of linguistics description: phonology, morphology, syntax, semantics, etc. But the needs of the target applications should be taken into account to determine which subset is really needed. A large proportion of medical NLP works target the recognition of these terms and their variants in text for indexing or information retrieval applications (Aronson, 2001). The target lexical information should consequently be able to support these tasks. Of course, many different kinds of variants can be addressed, and the choice has to be made according to the specificity of these variants.

A study of the Term-Union highlights interesting characteristics. First, terms of medical language are frequently made of more than one lexeme (e.g. *trouble congénital de la segmentation*). Second, it gathered an important number of term variations around each CUI. Out of the 154,594 CUI in Term-Union, a little bit less than the half (68,118 CUI) is associated with more than one terms. The obvervation of these variants brings to light three main types of variants that are primarily addressed in this project and are presented below. Others, such as semantic variation, will be considered later.

1. graphemic variations: Spelling of highly specialised terms is sometimes flexible. For example, *équilibre acido-basique* and its graphemic variant *acidobasique* [EN: acid-base balance] are found under the same identifier (CUI) in the French MeSH Thesaurus (INS, 2009). 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 name 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.
2. inflectional variations: Inflectional knowledge is important to assign each lexical item categorical and morphosyntactic information, together with its lemma. Both plural and singular forms of the same term can be found in Term-Union, like in *adaptation de l'oeil* and its variants *adaptation des yeux* [EN: eye adaptation]. This variation is also 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.
3. derivational variations: Derivational knowledge is particularly useful in medical terminology, because one term can have many "morphosemantic" variants, as in *intoxication à l'alcool* which is recorded with the same CUI in Term-Union as *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.

To be able to process all these variants, the specialized lexicon should contain relevant information. The section below presents the organisation of the various lexical resources that are currently under construction.

2.1.3. Organization of the specialized lexicon

Following what was done for similar lexicons in other languages (cf. section 1), 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. We present here the three types of relational tables that are targeted to cover the necessary descriptions at the three different levels described above.

1. graphemic variation: since the spelling of highly specialized terms is sometimes flexible, the different spellings of a lexeme should be listed, and linked with the variant that is considered to be the 'reference' (i.e. for any hyphenated word also found without a hyphen, a specific resource should contain the two forms.
   ```
   inter−maxillaire|intermaxillaire
   insulino−sécrétantes|insulinosécrétantes
   scléro−cornéenne|sclérocornéenne
   ```
2. inflection: inflectional knowledge is important to assign to each lexical item categorical and morphosyntactic information together with the lemma, in order to recognized inflected forms of a term. Consequently, a full inflectional lexicon should provide necessary information for the full paradigm[4], as in:
   ```
   sérofibrineux|sérofibrineux|Afpms
   sérofibrineuse|sérofibrineux|Afpfs
   sérofibrineux|sérofibrineux|Afpmp
   sérofibrineuses|sérofibrineux|Afpfp
   ```
3. derivation: relational tables for derivation provide morphological information for constructed words. Each table represents a specific morphological link between a derived lexeme of a particular category and its base lexeme. For example, the relational table Xsfx-X (shown below) provides information for relational adjectives and their base nouns.
   ```
   ...
   abdominal|abdomen
   aplasique|aplasie
   appendiculaire|appendicule
   arachnéphobique|arachnéphobie
   arachnoïdien|arachnoïde
   argentique|argent
   ...
   ```

All this information can be then summed up and represented in a standard framework such as the Lexical Markup Framework (Francopoulo et al., 2009).

2.2. Coverage of the initial state of the UMLF lexicon

A first version of the UMLF was produced at the first stage of the UMLF project by gathering lexical entries from lexicons of the project partners, with a focus on the lexical

[4] Inflectional information is encoded following the Grace/Multext formathttp://aune.lpl.univ-aix.fr/projects/multext/.

database compiled at the Geneva University Hospital (Baud et al., 1998). This lexicon contained 17,192 lexical units (5,353 adjectives and 11,799 nouns), together with their complete inflectional paradigms (36,211 word forms). To evaluate its coverage, i.e. its lexical completeness, we confronted it with the Term-union. The confrontation was performed on single words after case folding.

2.3. Obtaining entries from general lexicons

In any specialized 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 behavior, their morphosyntactic characteristics are generally identical in both specialized 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 2.0[5] which contains 67,376 lemmas and 524,725 word forms.

2.4. Learning morphosyntactic information from existing lexicons

To minimize human work to acquire inflectional knowledge for the remaining word-forms, we tested automatic methods. The task we want to perform is three-fold: for any unknown word-form, the objective is *(i)* to get its morphosyntactic information (i.e. the POS and the gender and number information) *(ii)* to obtain its lemma and *(iii)* to complete its full inflectional paradigm (e.g. an adjective has to be recorded with its 4 forms: masculine-singular, feminine-singular, masculine-plural, feminine-plural).

2.4.1. Guessing the tag

To achieve the first objective, we used the algorithm of (Tanguy and Hathout, 2007, p. 295) to acquire the full tag of a word form (POS + gender and number info) in a reference lexicon, and then to guess 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. For each final character string, the program calculates the most frequent tag. For the longest final character string (8 or more), all the possible tags are recorded. Otherwise, only the most frequent tag is kept, except for the adjectives, since in French, adjective are fully inflected in French, and one single word-form can be both masculin and feminine (like *alcoolique*).

To enhance the quality of the output, two different reference lexicons were used. The first one is the general lexicon Morphalou (c.f. section 2.3 above) and the second one is the UMLF itself (in its initial version). The learning program is run on the two lexicons, and only the lexical units that have been guessed the same way are kept.

2.4.2. Acquisition of the full paradigm

Once full tags have been guessed for each word form, the next step is to acquire the complete paradigm (i.e. the four forms and the lemma). Based on a pattern model that is

Table 1: Example adjective inflectional paradigms

m.s.	f.s.	m.pl.	f.pl.	Example
(.*)	(.*)	(.*)s	(.*)s	pulmonaire|Afpms pulmonaire|Afpfs pulmonaires|Afpmp pulmonaires|Afpfp
(.*)el	(.*)elle	(.*)els	(.*)elles	artériel|Afpms artérielle|Afpfs artériels|Afpmp artérielles|Afpfp
(.*)x	(.*)se	(.*)x	(.*)ses	veineux|Afpms veineuse|Afpfs veineux|Afpmp veineuses|Afpfp

made of 9 "productive" inflectional paradigms for adjectives, and 3 for nouns (Table 1 provides three examples of adjective inflectional paradigms) the algorithm tries to cluster together word-forms of the same pattern. The algorithme uses a lexical trie based approach to cluster all the guessed forms that belong to the same paradigm. The more member of one paradigm are found, the more confident we can be on the guessing part. If one of the members of the paradigm is missing, it tries to generate it, based on the pattern model.

When one of the members of the pair is the canonical form (masculine singular for adjectives and singular for nouns), the lemma can be automatically generated. Otherwise, it can be hypothesized by means of the pattern, but this latter case requires human checking.

3. Results

3.1. Initial coverage and acquisition from a general lexicon

81,595 out of the 94,964 distinct word forms in the Term-Union were not found in the initial version of the UMLF. These 81,595 word-forms were further processed as described above to add entries to the UMLF. As shown in Table 2, 6,617 out of the 81,595 remaining word forms were known from Morphalou. These words are common medical words such as *alité, auscultatoire* or *cardiographique*. They were consequently added to the UMLF, together with the rest of their inflectional paradigms.

Table 2: Coverage: initial version and first extension of the lexicon

	Known words entries	Remaining words to describe
Term-Union		94,964
Initial UMLF	19,599	81,595
Morphalou	6,617	74,978

Interestingly, and expectedly, the 74,978 forms that remain unknown from Morphalou are specific to the medical domain (like *adrénocorticotrophine*, *cérébrosclérose*, *circon-*

[5] http://www.cnrtl.fr/lexiques/morphalou/

volutionnelle or *macracanthorhynchose*). They represent 79% of the number of lexical units within the Term-Union, which shows the specificity of the vocabulary in the terminologies included in Term-Union.

3.2. Acquisition and consensus guessing

Among the 74,978 unknown forms, 34,612 received one or more tags from the guessing based on Morphalou, and 30,579 from the guessing based on UMLF. But since the guessing program allows more than one tag, there are actually 44,515 analyses provided by the Morphalou-based program, and 35,438 analyses provided by the UMLF-based program. Amongst all these possible tagged lexical units, 30,137 were analyzed the same way with the two reference lexicons. This consensus guessing yields an interesting validation of the output.

We evaluated a sample of 1,000 entries, and found that only 82 were wrongly labelled (8.2%; see Table 3). An

Table 3: Evaluation of a sample of 1000 guessed entries and classification of errors

Type	Number
Correct	918
Errors	82

Error Type	Number
Wrong label	12
Proper names	49
Latin words	5
English words	1
Spelling/segmentation	10
Other	5

error analysis shows that only 12 were real POS labelling errors (e.g. "accidentellement", an adverb, was labeled as a noun—since there is no adverb in the two reference lexicons—or "kascher" labeled 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 special resource (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.

3.3. Acquisition of the full paradigm

Out of the 30,137 word forms, the algorithm captured 4,453 paradigms (incomplete or not), grouping 9,352 wordforms. 3,308 paradigms were found for adjectives; Table 4

Table 4: Adjective paradigms

members	number	forms
2	2892	5784
3	399	1197
4	17	68

provides detailed information about the captured adjective paradigms—with 2, 3 or 4 members—and the number of forms they contain. 514 complete paradigms were found for nouns. Moreover, 621 adjectival paradigms were found with 2 or 3 members, but without the canonical forms (i.e. masculine singular). For now, only the adjective paradigms which contained a canonical form were automatically extended. In total, we automatically completed 3,212 adjectival paradigms (12,848 word forms).

3.4. Improvement of the coverage of the lexicon

After this extension, 17,828 forms from the Morphalou lexicon were added to UMLF, and 8,088 from the semi-automated acquisition explained above. In total, UMLF now contains 62,127 forms, together with their full inflectional paradigms. But this figure does not reflect the improvement of the coverage of the lexicon for the targetted domain. To compute this improvement, a comparison was performed at each step with the reference word-forms from the Term-Union. As shown in Table 5, coverage improvement with the simple method of acquisition is very encouraging.

Table 5: Extensions to the UMLF lexicon. Coverage is measured as a percentage of the 94,964 forms in Term-union

Source	Forms added	Still unknown in Term-union	Coverage
UMLF-v1	36,211	81,595	14,1%
Morphalou	17,828	74,978	21,0%
Acquisition	8,088	70,602	25,7%

4. Discussion and Conclusion

In this article, we presented the state of development of a specialized French lexicon for the medical domain, and we described the needed specific information. Based on the fact that acquisition of specialized lexical knowledge requires appropriate data, we showed how a large terminological database coming from various sources can be a very useful resource to characterize the phenomena that need to be described and to focus the acquisition of inflectional information. The important amount of data helped to get enough examples of inflected items, which allowed us to acquire quickly some of the needed information to feed the lexicon. We are currently investigating other machine-learning techniques (such as Conditional Random Fields), to learn from the data found in Term-Union and to improve the coverage of the inflectional lexicon. For derivational knowledge, other rule-based techniques are under consideration.

5. References

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

Robert H. Baud, Christian Lovis, Anne-Marie Rassinoux, Pierre-André Michel, and Jean-Raoul Scherrer. 1998.

Automatic extraction of linguistic knowledge from an international classification. In Branko Cesnik, Charles Safran, and Patrice Degoulet, editors, *Proceedings of the 9th World Congress on Medical Informatics*, pages 581–585, Seoul.

Olivier Bodenreider and Pierre Zweigenbaum. 2000. Stratégies d'identification de noms propres à partir de nomenclatures médicales parallèles. *Traitement automatique des langues*, 41(3):727–757.

Bruno Cartoni and Pierre Zweigenbaum. 2010. Extension of a specialised lexicon using specific terminological data: the unified medical lexicon for french (umlf). In *Proceedings of EURALEX*.

Gil Francopoulo, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet, and Claudia Soria. 2009. Multilingual resources for NLP in the lexical markup framework (LMF). *Language Resources and Evaluation*, 43(1):57–70, March.

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

Alexa T. McCray, S. Srinivasan, and A. C. Browne. 1994. Lexical methods for managing variation in biomedical terminologies. In *Proceedings of the 18^{th} Annual SCAMC*, pages 235–239, Washington. Mc Graw Hill.

Thomas C. Rindflesch, Marcelo Fiszman, and B. Libbus. 2005. Semantic interpretation for the biomedical literature. In H Chen, S Fuller, WR Hersh, and C Friedman, editors, *Medical informatics: Advances in knowledge management and data mining in biomedicine*, pages 399–422, Berlin / Heidelberg. Springer.

Ludovic Tanguy and Nabil Hathout. 2007. *Perl pour les linguistes*. Hermes-Sciences Lavoisier, Paris.

G. Weske-Heck, Albrecht Zaiß, M. Zabel, Stefan Schulz, Wolfgang Giere, M. Schopen, and Rudiger Klar. 2002. The German Specialist Lexicon. *Journal of the American Medical Informatics Association*, 8(suppl).

Pierre Zweigenbaum, Robert H. Baud, Anita Burgun, Fiammetta Namer, Éric Jarrousse, Natalia Grabar, Patrick Ruch, Franck Le Duff, Jean-François Forget, Magaly Douyère, and Stéfan Darmoni. 2005. A unified medical lexicon for French. *International Journal of Medical Informatics*, 74(2–4):119–124, March.