Lexically-Based Terminology Structuring

Natalia Grabar, Pierre Zweigenbaum

STIM — DSI, Assistance Publique – Hôpitaux de Paris

& ERM 202, INSERM, Paris

& CRIM — INaLCO, Paris

{ngr,pz}@biomath.jussieu.fr

Abstract

Terminology structuring has been the subject of much work in the context of terms extracted from corpora: given a set of terms, obtained from an existing resource or extracted from a corpus, it consists in identifying hierarchical (or other types of) relations between these terms. The present work aims at assessing the feasibility of such structuring by studying it on an existing hierarchically structured terminology. Our overall goal is to test various structuring methods proposed in the literature and to check how they fare on this task. The specific goal at the present stage of our work, which we report here, is focussed on lexical methods that match terms on the basis on their content words, taking morphological variants and synonyms into account. We describe experiments performed on the French version of the US National Library of Medicine MeSH thesaurus. We compare the lexically-induced relations with the original MeSH relations and measure recall and precision metrics, taking two different views on the task: relation recovery and term placement. This method proposes correct term placement for up to 26% of the MeSH concepts, and its precision can reach 58%. After this quantitative evaluation, we perform a qualitative, human analysis of the 'new' relations not present in the MeSH. This analysis shows, on the one hand, the limits of the lexical structuring method. On the other hand, it reveals some specific structuring choices and naming conventions made by the MeSH designers, and emphasizes ontological commitments that cannot be left to automatic structuring.

Keywords Terminology structuring, morphological variants, synonyms, medicine, MeSH.

1 Introduction and Background

Terminology structuring, i.e., organizing a set of terms through semantic relations, is one of the difficult issues that have to be addressed when building terminological resources (Jacquemin and Bourigault 2002; Nazarenko and Hamon 2002). These relations include subsumption or hypernymy (the *is-a* relation), meronymy (*part-of* and its variants), as well as other, diverse relations, sometimes called 'transversal' (e.g., *cause*, or the general *see also*).

Various methods have been proposed to discover relations between terms (see (Jacquemin and Bourigault 2002) for a review). We divide them into *internal* and *external* methods, in the same way as (McDonald 1993) for proper names. Internal methods look at the constituency of terms, and compare terms based on the words they contain. These term matching methods can rely directly on raw word forms (Bodenreider *et al.* 2001), on morphological variants (Jacquemin and Tzoukermann 1999), on syntactic structure (Bourigault 1994; Jacquemin and Tzoukermann 1999) or on semantic variants (synonyms, hypernyms, etc.) (Hamon *et al.* 1998). External methods take advantage of the context in which terms occur: they examine the behavior of terms in corpora. Distributional methods group terms that occur in similar contexts (Grefenstette 1994). The detection of appropriate syntactic patterns of cooccurrence is another method to uncover relations between terms in corpora (Hearst 1992; Séguéla and Aussenac 1999).

The present work aims at assessing the feasibility of such structuring by studying it on an existing, hierarchically structured terminology. Ignoring this existing structure and starting from the set of terms it contains, we attempt to discover hierarchical term to term links and compare them with the preexisting relations.

We test various structuring methods proposed in the literature and check how they fare on this task, focussing on internal, lexical methods. We adopt the lexical inclusion hypothesis (Kleiber and Tamba 1990): if one term is a subpart of an other one, a hyponymic relation is likely to exist between them. In this experiment, we process raw terms, but we also take into account their morphological variants and synonyms.

We also analyze 'new' induced relations. 'New' means that these induced relations are not present in the original hierarchical structure of the MeSH thesaurus. Although they count as 'noise' in our first, quantitative evaluation, they might nevertheless reflect useful links. Performing this analysis allows us to propose a more precise evaluation of the methods and their results and to point out some inherent limits.

After the exposition of the data we used in our experiments (section 2), we present methods (section 3) for generating hierarchical links between terms through the study of lexical inclusion and for evaluating their quality with appropriate recall and precision metrics. Results are discussed in section 4. We then

present the analysis of some ‘new’ induced relations and attempt to propose a typology of term dependency in these relations (section 5). We finally discuss the limits of lexical methods for the structuring task (section 6).

2 Terminological and lexical material for this study

The series of experiments presented here use an existing hierarchically structured thesaurus, the MeSH (section 2.1), a ‘stop word’ list (section 2.2), morphological knowledge (section 2.3) and synonyms (section 2.4).

2.1 The MeSH biomedical thesaurus

The Medical Subject Headings (MeSH, (NLM 2001)) is one of the main international medical terminologies (see, e.g., (Cimino 1996) for a presentation of medical terminologies). The MeSH is a thesaurus specifically designed for information retrieval in the biomedical domain. It is used to index the international biomedical literature in the Medline bibliographic database. The French version of the MeSH (INSERM 2000) contains a translation of these terms (19,638 terms) and their synonyms (note that the MeSH is revised each year, so that current numbers are different). It happens to be written in unaccented, uppercase letters. As many other medical terminologies, the MeSH has a hierarchical structure (figure 1): ‘narrower’ concepts (children) are related to ‘broader’ concepts (parents). The MeSH specifically displays

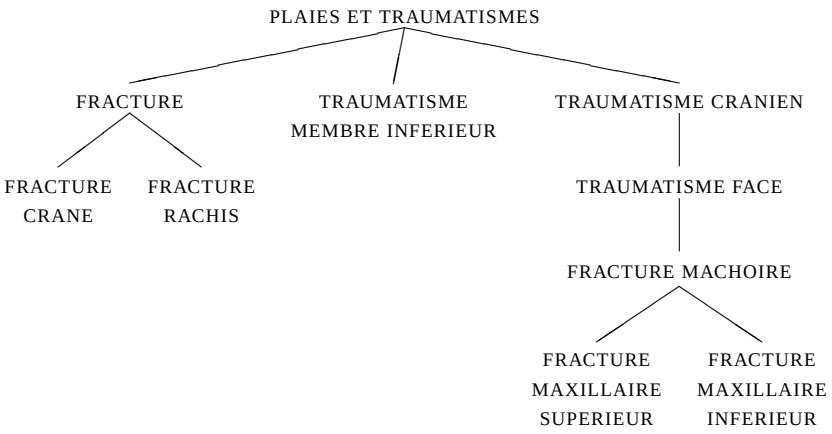

Figure 1: The MeSH thesaurus (excerpt).

a rich, polyhierarchical structure: each concept may have several parents. For instance, the MeSH editors positioned the concept “fracture of the jaw” (FRACTURE MACHOIRE, figure 1) as a child both of “fracture”, since it is a kind of fracture, and of “face trauma” (TRAUMATISME FACE), since a fracture is a kind

of trauma and a trauma located in a part of the face is a kind of trauma of the face. The MeSH contains 26,094 direct child-to-parent links and (under transitive closure) 95,815 direct or indirect child-to-ancestor links.

2.2 Stop word list

'Stop word lists' are used to remove from term comparison very frequent words which are considered not to be content-bearing, hence 'non-significant' for terminology structuring. The stop word list used in this experiment is a short one (15 word forms). It contains the few grammatical words (articles and prepositions) which occur frequently within MeSH terms:

au, aux, d', de, des, du, en, et, l', la, le, les, ses, un, une

2.3 Morphological knowledge

Previous work has acknowledged morphology as an important area of medical language processing and medical information indexing (Pacak *et al.* 1980; Wingert *et al.* 1989; Grabar *et al.* 2003) and of term variant extraction (Jacquemin and Tzoukermann 1999). In this work, we apply morphological knowledge to the terminology structuring task.

Typically, three types of morphological relations are considered:

- *Inflection* produces the various forms of a same word according to number, gender, person, tense, etc.: *'intervention'* – *'interventions'*, *'acid'* – *'acids'*. Reducing an inflected form to its canonical form, or *lemma*, is called lemmatization. A lemma and its inflected forms share the same part of speech.

- *Derivation* is used to obtain, e.g., the adjectival form of a noun (noun *'aorta'* ↔ adjective *'aortic'*, verb *'intervene'* ↔ noun *'intervention'*, adjective *'human'* ↔ adverb *'humanely'*). Derivation often deals with words of different parts of speech. Reducing a derived word to its base word is called stemming.

- *Compounding* combines several radicals, here often of Greek or Latin origin, to obtain complex words (e.g., *'aorta'* + *'coronary'* yields *'aortocoronary'*), the so-called *neo-classical compounds*.

The morphological knowledge we used consists of {*lemma, derived or inflected form*} pairs of word forms where the first is the 'normalized' form and the second a 'variant' form. Both forms of such a pair should have similar meanings, so that they can be substituted for one another in term comparison

operations. Therefore we rely on inflectional knowledge and derivations that do not change word meaning. We have left compounding aside for the time being, since the words it relates may have more distant meanings.

2.3.1 Inflectional knowledge

For inflection, we have two lexicons of {*lemma, inflected form*} word pairs. The first one (*lem-gen*) is based on a general lexicon (ABU, `abu.cnam.fr/DICO`) which we have augmented with pairs obtained from medical corpora processed with a tagger/lemmatizer (in cardiology, hematology, intensive care, and drug monographs): it contains up to 308,812 pairs (where the inflected form is different from the lemma). The second lexicon (*lem-med*) is the result of applying rules acquired in previous work (Zweigenbaum *et al.* 2001) from two medical terminologies (ICD-10 and SNOMED) to the vocabulary of the MeSH, ICD-10 and SNOMED. This gives a total of 3,470 pairs.

Lemmatization can be ambiguous when an inflected form can be obtained from several lemmas (e.g., *"souris"* → *"souris/N"* (Engl. *'mouse'*) and *"sourire/V"* (Engl. *'to smile'*)). In that case, we have adopted a brute force approach which merges the two corresponding morphological families and chooses one lemma as unique representative for both.

2.3.2 Derivational knowledge

For derivation, we also used resources from (Zweigenbaum *et al.* 2001) which result in 2,418 pairs {*base, derived form*} (*stem-med*). The approach used to build these resources is based on the presence of an initial string of four common characters in both the base and derived words. Derivations with suppletive and allomorphic stems, such as {'stomach', 'gastric'}, cannot then be detected.

To improve these resources, we extracted from the SNOMED International Nomenclature (Côté 1996) simple synonymous terms with different syntactical categories and filtered them manually. This allowed us to enrich the derivations with suppletive and allomorphic derivational knowledge (462 word pairs, some of which were already present in *stem-med*), which increases the total number of derivational pairs to 2,716 (*allom*).

To obtain a better coverage of the domain, these morphological resources will still need to be improved and enriched.

2.4 Synonyms

As for derivation, we used here two sets of synonyms: one is general (*syno-gen*) and the other specific to the medical domain (*syno-med*).

The specialized set of synonyms (*syno-med*) has been extracted from the *Masson* medical dictionary (online at `www.AtMedica.com`); it contains 831 pairs of single-word synonyms.

The set of general synonyms (*syno-gen*) comes from the general-language dictionary *Le Robert*,[1] which lists 140,141 pairs of single-word synonyms. But, as we shall see below, this synonym knowledge is fairly noisy when applied to medical terms. Therefore, we decided to adapt it through the study of an 8.5 million word medical corpus (Grabar and Zweigenbaum forthcoming). This corpus contains hospital documents (patient discharge summaries) collected from haematology and cardiology departments and a large set of public documents in the domains of nephrology and neurology and more generally signs and symptoms, collected through the CISMeF catalog and index for French medical Web sites (Darmoni *et al.* 2000). We explored several approaches to filter general language synonyms:

- We used nine lexico-syntactic patterns for synonymy from (Séguéla and Aussenac 1999), such as:
  "X appelé Y" (Engl. *'X called Y'*)
  "X est défini comme 1-MOT Y" (Engl. *'X is defined as 1-WORD Y'*)
  "X n'est autre que 1-MOT Y" (Engl. *'X is no other than 1-WORD Y'*)
  They allow us to find, for instance, the following synonyms in the following contexts:
  - *"gonflement"* - *"oedème"* (Engl. *'swelling - edema'*): *"L'oedème est défini comme un gonflement palpable produit par l'expansion du volume interstitiel liquidien."*
    'Oedema is defined as a palpable swelling…'
  - *"rhinopharynx"* - *"cavum"* (Engl. *'nasopharynx - cavum'*): *"Le rhinopharynx appelé cavum est situé sous la base du crâne, en arrière des fosses nasales, au-dessus de l'oropharynx et en avant des 2 premières vertébres cervicales."*
    'The nasopharynx called cavum is located…'

  These patterns never match in hospital documents. Nevertheless, on the Internet documents, they match with 46 pairs of synonyms.

- (Lame 2002) noticed that in legal documents, the coordination markers *"et"* (Engl. *'and'*) and *"ou"* (Engl. *'or'*) coordinate terms with a close meaning and represent valid semantic relations. If these coordination markers link words of a given pair of synonyms from Le Robert, we assume that these

words are valid synonyms in the medical domain too. We have added an other coordination marker *"ni"* (Engl. *'nor'*), as well as the negation *"pas"* (Engl. *'not'*), since they are productive in the domain. It seems that coordination detects co-hyponyms rather than synonyms:

  - *"bruit" - "souffle"* (Engl. *'sound - wheezing'*): *"Examen cardiaque : bruits bien frappés aux 4 foyers sans souffle ni bruit surajoutés."*
  - *"orthopnée" - "dyspnée"* (Engl. *'orthopnea - dyspnea'*): *"Examen cardio-vasculaire : pas de dyspnée, pas d'orthopnée, présence d'oedèmes des membres inférieurs avec un godet positif."*

  With coordination, we match 1,736 pairs of synonyms in all our corpora.

- (Zweigenbaum *et al.* 2003) applied a statistical measure of association between two words, the 'log likelihood ratio' (Manning and Schütze 1999), for detecting morphologically related words. We have adapted this method and the corresponding program to the filtering of potential synonyms. The rationale is that if two 'general' synonyms co-occur more often than chance in a corpus at a distance smaller than a given window size N, then we can be more confident that they are actually used as synonyms in this corpus. This method has been run with a window of 2*150 full words (stop words being first removed). The general synonyms have been ranked in decreasing order of association, and the top 60% were kept. This selects 15,589 pairs of synonyms, among them:
  "abcès" - "phlegmon" (Engl. *'abscess - phlegmon'*), *"barrière" - "limite"* (Engl. *'barrier - limit'*), *"biopsie" - "ponction"* (Engl. *'biopsy - puncture'*), *"signal" - "appel"* (Engl. *'signal - call'*).

These three sets of filtered general-language synonyms were merged; the resulting set contains a total of 16,154 pairs of synonyms.

3 Lexical inclusion

The present work induces hierarchical relations between terms when the constituent words of one term lexically include those of the second term (section 3.1). We evaluate these relations by comparing them with the preexisting relations, computing precision and recall both for links and concepts (section 3.3).

3.1 Principles

The method we use here for inducing hierarchical relations between terms is basically a test of *lexical inclusion*: we check whether a term P (*parent*) is 'included' in another term C (*child*). We assume that this

type of inclusion is a clue of a hierarchical relation between terms, as in the following example: *"acides gras"* / *"acides gras indispensables"* (Engl. *'fatty acids'* / *'fatty acids, essential'*).

To detect this type of relation, we test whether all the content words of P occur in C. We test this on segmented terms with a gradually increasing normalization on word forms:

- basic normalization: conversion to lower case, removal of accents, punctuation marks, numbers and 'stop words' (introduced in section 2.2);
- normalization with morphological resources: lemmatization (with the two alternative inflectional lexicons presented in section 2.3.1) and stemming (with two derivational lexicons, see section 2.3.2);
- normalization with synonyms (see section 2.4): general-language and domain-specific sets of synonyms.

Individual words in terms are indexed separately to speed up the computation of term inclusion over all term pairs of the whole MeSH thesaurus. When these normalizations are applied, terms are indexed by their normalized words: we assume that P is lexically included in C if all normalized words of P occur in C.

3.2 Normalization of lexical variants

The gradually increasing normalizations we applied to our list of terms cover an increasingly large number of variations and induce an increasing number of hierarchical links between these terms. We list below the normalization sequences which were applied. *S-X* identifies each normlization step; and *X*, the entire normalization sequence ending at that step (we do not make this distinction for the first step, which is trivially identical to the corresponding sequence).

- *basic*: basic normalization (lower case conversion, removal of accents, punctuation marks, numbers and 'stop words').
  The *basic* normalization is performed in all normalization sequences;
- *S-lem-gen*: application of 308,812 {*lemma*, *inflected form*} pairs from general lexicon and lemmatized medical corpora.
  lem-gen = *basic* + *S-lem-gen*;
- *S-lem-med*: application of 3,470 {*lemma*, *inflected form*} pairs acquired on medical terms.
  lem-med = *basic* + *S-lem-med*;

- *S-stem-med*: application of 2,418 {*base*, *derived form*} pairs acquired on medical terms.
  stem-med = basic + S-lem-med + S-stem-med;

- *S-allom*: application of 462 allomorphic and suppletive pairs from Snomed, in addition to *stem-med*, total: 2,716 pairs.
  allom = basic + S-lem-med + S-stem-med + S-allom;

- *S-syno-med*: application of 831 pairs of synonyms from medical lexicon.
  syno-med = basic + S-lem-med + S-stem-med + S-allom + S-syno-med;

- *S-syno-gen*: application of 140,141 pairs of synonyms from general lexicon.
  syno-gen = basic + S-lem-med + S-stem-med + S-allom + S-syno-gen;

- *S-syno-gen-f*; application of 16,154 filtered general lexicon pairs of synonyms.
  syno-gen-f = basic + S-lem-med + S-stem-med + S-allom + S-syno-gen-f.

3.3 Evaluation

We evaluated the results obtained with this lexical inclusion approach by comparing them with the original structure in the MeSH. We were faced with two issues to perform this evaluation: the polyhierarchical nature of the MeSH and the transitivity of the *is-a* link. Two methods were considered to deal with the polyhierarchy issue (see figure 2).

1. The first method is interested in the number of links found, and compares these links with those originally present in the MeSH thesaurus: do we obtain all the links that pre-exist in the MeSH? In this measure, the gold standard is the full polyhierarchical structure of the MeSH, taking into account the multiple links which many concepts share with different parents.

2. The second method considers the positioning of individual MeSH concepts (terms) in the hierarchical structure of the thesaurus: can we place each concept in *at least one* suitable position in the emerging hierarchy? This is a relaxed gold standard, which requires that no concept be left unlinked: at least one link is expected to be found for each concept.

Indeed, in the case of a monohierarchical terminology, these two measures would be equivalent.

For both methods, we compute recall and precision metrics. The recall metric analyzes the completeness of the results, i.e., tries to find out whether all MeSH links are induced or whether all MeSH concepts are correctly positioned. The precision metric evaluates the proportion of correct links or positions among the induced results. To compute these metrics we need to agree on what a correct link is. Basically, it is

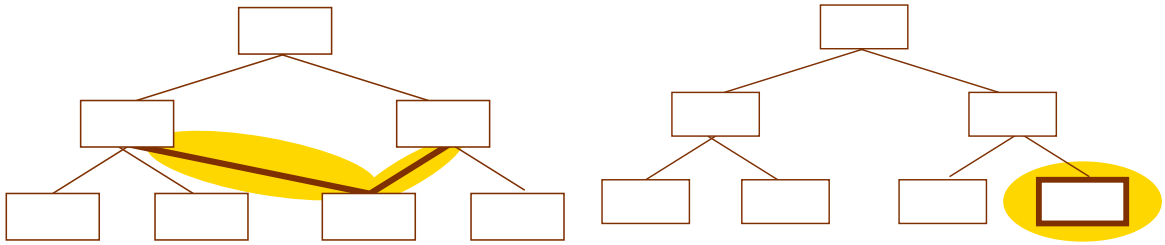

Figure 2: Evaluation of links (left) *vs.* evaluation of node placement (right). Links: all correct links for a given concept node must be found in the original structure of the MeSH. Nodes: a node must be correctly attached to the rest of the structure by at least one link.

a direct MeSH relation between a concept and one of its (immediate) parents (the examples in figure 2). However, because of the transitivity of the *is-a* relation, a link found between a concept and an ancestor higher in the hierarchy is also correct, although less specific. Therefore, we devised two versions of the recall and precision measures:

- strict (or direct): only the links to direct parents of a given concept are considered satisfactory, and
- tolerant (or indirect): a link to any ancestor is considered as correct.

Let us consider the number of links l returned by our link induction method. These links include d_c correct direct links, i_c correct indirect links and n non-MeSH links ($l = d_c + i_c + n$). The direct or strict recall R_s is measured by the number of correct direct links d_c found divided by the total number of direct links $D = 26,094$ in the MeSH. The tolerant recall R_t is measured by the number of correct direct or indirect links $d_c + i_c$ divided by the total number of links $D + I = 95,815$ in the MeSH:

$$R_s = \frac{d_c}{D}; \qquad R_t = \frac{d_c + i_c}{D + I}$$

The evaluation of the precision metric also takes into account both strict and tolerant appoaches; given d_c the number of correct direct links found, i_c the number of correct indirect links found, and n the number of non-MeSH links found, the strict precision P_s and the tolerant precision P_t are computed as:

$$P_s = \frac{d_c}{d_c + i_c + n} = \frac{d_c}{l}; \qquad P_t = \frac{d_c + i_c}{d_c + i_c + n} = \frac{d_c + i_c}{l}$$

We also tested a mixed scheme where the weight given to each link depends on the distance between the two concepts related by this link in the original hierarchical MeSH structure: the more distant these concepts, the lower the weight obtained by the induced link. However, this mixed scheme obtained results which were not very different from those of the tolerant scheme, so that for the sake of space we do not

present them here.

The evaluation of recall R_c and precision P_c for term placement uses similar formulas with counts of terms correctly placed rather than counts of links. The number c of different concepts which obtain at least one upgoing hierarchical link by our induction method includes d_{cc} correct direct links, i_{cc} correct indirect links and n_c non-MeSH links ($c = d_{cc} + i_{cc} + n_c$). The MeSH contains $C = 19,638 - 1$ different concepts with at least one parent (all concepts but the root). Therefore, for term placement, we have:

$$R_{cs} = \frac{d_{cc}}{C}; \qquad R_{ct} = \frac{d_{cc} + i_{cc}}{C}$$

$$P_{cs} = \frac{d_{cc}}{d_{cc} + i_{cc} + n_c} = \frac{d_{cc}}{c}; \qquad P_{ct} = \frac{d_{cc} + i_{cc}}{d_{cc} + i_{cc} + n_c} = \frac{d_{cc} + i_{cc}}{c}$$

The lexical inclusion methods and the evaluation procedure were implemented as Perl5 scripts.

4 Results

In this section, we first quantify the results obtained with lexical inclusion methods (section 4.1), and then compare them to the information contained in the MeSH (section 4.2).

4.1 Lexical inclusions obtained

The method described in section 3.1 has been applied to the 'flat' (unstructured) list of 19,638 terms ('main headings') of the MeSH thesaurus. Figure 3 shows quantitative results for the analysis of lexical inclusions and each type of normalization tested. The left panel shows the number of induced relations; let us note in comparison that the number of genuine hierarchical MeSH relations is 95,815. The right panel indicates the number of terms which have been linked with our methods; MeSH contains 19,638 linked terms in its hierarchy.

As expected, the number of links induced between terms increases when applying more resources for normalization. Inflectional knowledge compiled from the medical domain terminologies (*lem-med*) allows us to link more terms than inflectional knowledge from a general lexicon (*lem-gen*): 12,857 *vs.* 12,210 links. We observe the same situation for the positioning of terms, where we obtain a better coverage of terms when using specialized morphological knowledge (*lem-med*) than when using morphological knowledge from a general lexicon (*lem-gen*): 10,929 *vs.* 10,560 terms. Derivational word pairs again increase the number of links induced (14,695 with *stem-med vs.* 12,857 with *lem-med*) and of terms placed under a proposed parent (11,511 *vs.* 10,929). Using allomorphic derivations (*allom*) only slightly modifies the

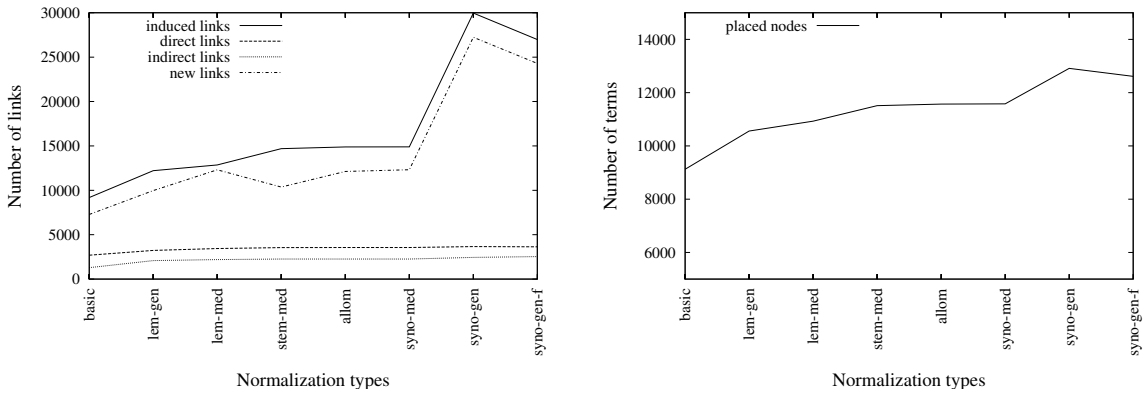

Figure 3: Quantification of induced relations between analyzed terms and of terms placed under a proposed parent by these links.

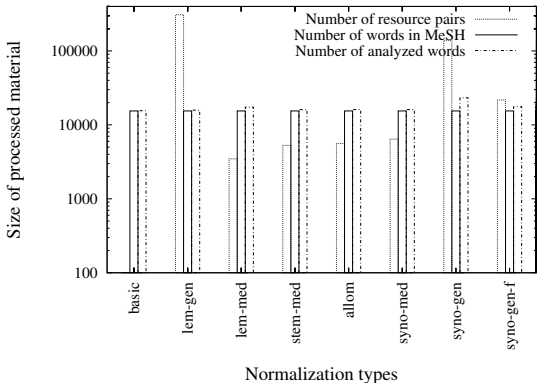

Figure 4: Number of words really processed in each normalization step.

results: 189 links and 61 placed terms more than in *stem-med*. Medical synonyms allow us to induce only 8 more links and to place 7 more terms. With the application of general language synonyms, the number of results increases drastically: 15,085 induced links and 1,314 terms placed more than with *allom*. When general language synonyms are filtered out, the difference is reduced, but still important: 12,102 induced links and 1,039 terms placed more than with *allom*.

Figure 4 shows (i) the total number of word pairs in each resource (*lem-gen* lexicon, *lem-med* lexicon, etc.), (ii) the number of words in MeSH (this is a constant provided as a reference: 15,446 after discarding sequences containing numbers), and (iii) the number of different words really used during each step performed. One can see that many word pairs in the resources applied are not used, particularly with general-language resources (*lem-gen* and *syno-gen*). The tuning of these resources to the processed domain can then be seen as useful and time-saving.

Table 1 shows examples of lexically included terms which we obtained with this method. For each type

Table 1: Examples of correct, lexically induced MeSH terms, and their English translations. 'Indirect' means that the MeSH includes a path of length > 1 from the parent to the child.

Type of normalization	Parent P	Child C
basic $_{direct}$	accouchement	accouchement provoque
	delivery	*labor, induced*
basic $_{indirect}$	acides gras	acides gras indispensables
	fatty acids	*fatty acids, essential*
lem-gen $_{direct}$	intervention chirurgicale	interventions chirurgicales obstetricales
	surgical procedures, operative	*obstetric surgical procedures*
lem-gen $_{indirect}$	intervention chirurgicale	interventions chirurgicales voies biliaires
	surgical procedures, operative	*biliary tract surgical procedures*
lem-med $_{direct}$	agents adrenergiques	inhibiteurs captage agent adrenergique
	adrenergic agents	*adrenergic uptake inhibitors*
lem-med $_{indirect}$	chromosomes humains	chromosome humain 21
	chromosomes, human	*chromosomes, human, pair 21*
stem-med $_{direct}$	aberration chromosomique, anomalies	aberrations chromosomes sexuels, anomalies
	chromosome abnormalities	*sex chromosome abnormalities*
stem-med $_{indirect}$	eosinophilie	poumon eosinophile
	eosinophilia	*pulmonary eosinophilia*
allom $_{direct}$	poumon	eau extravasculaire pulmonaire
	lung	*extravascular lung water*
allom $_{indirect}$	estomac	cellule parietale gastrique
	stomach	*parietal cells, gastric*
syno-med $_{direct}$	saccharose	sucrose alimentaire
	saccharose	*dietary sucrose*
syno-med $_{indirect}$	—	—
	—	—
syno-gen $_{direct}$	fracture machoire	fracture maxillaire inferieur
	jaw fractures	*mandibular fractures*
syno-gen $_{indirect}$	therapeutique	traitement par art
	therapeutics	*art therapy*

of normalization, it shows *parent* / *child* pairs corresponding to direct, then indirect relations in the original MeSH structure.

4.2 Evaluation of these lexical inclusions

In section 3.3 we presented the methods designed to evaluate the structuring results we obtain with a lexical inclusion analysis of terms. These methods allow us to evaluate recall and precision metrics both for relations between terms and for term positioning. In all the cases we take into account the nature of induced links (direct or indirect ones) by testing both strict and tolerant variants. Correctness is computed by comparing these links with the original MeSH structure. Remember that *strict recall* and *strict precision* only take into account direct links induced by our method or found in the MeSH; *tolerant recall* and *tolerant precision* take all the links into account.

Figure 5 shows the evaluation results, recall and precision, for the links induced, and figure 6 shows the same information for concept (term) placement.

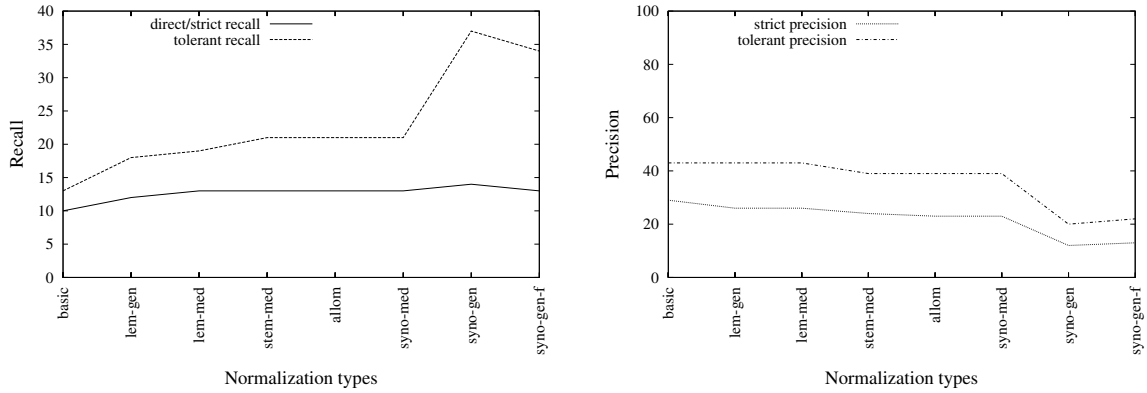

Figure 5: Evaluation of recall and precision for induced links.

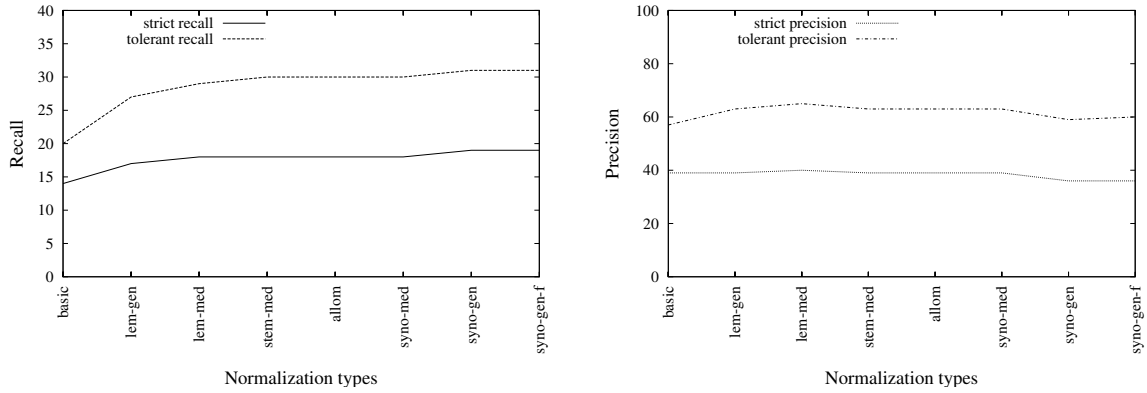

Figure 6: Evaluation of recall and precision for placed terms.

When we consider links, recall increases when applying more complete normalization knowledge. Morphological knowledge (inflection, derivation and then allomorphic variants) yields a slight improvement in recall: from 10.3% with *basic* normalization to 13.6 with *allom*, for strict recall, and from 13.7% to 21.6% for tolerant recall. And, not surprisingly, we notice that the recall (here tolerant, but the same is true of strict recall) of relations between terms obtained with morphological knowledge acquired from medical terminologies (*lem-med*, *stem-med* or *allom*) is higher (19.3%, 21.4% and 21.6%) than the recall corresponding to the use of the morphological knowledge compiled from the general lexicon (*lem-gen*, 18.3%). The application of medical domain pairs of synonyms produces an extremely small increase in both strict and tolerant recalls. With general language synonyms strict recall is augmented to 14% and tolerant recall to 37.4%. Filtered general synonyms increase strict recall 13.9% and tolerant recall to 34.6%.

The evolution of precision is opposite: injection of more extensive morphological (derivation *vs.* inflection) then synonym knowledge leads to taking more 'risks' for generating links between terms: *basic* strict precision decreases from 29.3% to 23.9% for *allom* and *syno-med*, down to 12.2% for *syno-gen* and 13.5% for *syno-gen-f*.

When accepting both direct and indirect links (tolerant approach), the precision obtained is higher than when only direct links are considered (strict approach). This corresponds to the fact that lexical inclusion directly identifies a number of links present in the MeSH as indirect hierarchical links: these are now counted as correct.

For instance, with basic normalization, the tolerant approach gives a precision of 43.3% and the strict approach 29.3%. With *lem-med* normalization, the precision is 43.8% (tolerant) or 26.8% (strict). For the *allom* and *syno-med* normalizations, the tolerant precision is 39% and the strict precision is 23.9%. And finally, for *syno-gen* and *syno-med*, the tolerant precision is respectively 20.4% and 22.8% and the strict preision 12.2% and 13.5%.

Depending on the normalization and on the evaluation scheme, up to 37.6% of the links found are correct (tolerant approach with *lem-med*), and up to 14% of the direct MeSH links are found by lexical inclusion (strict approach with *syno-gen*).

When we consider term placement into a common hierarchy, the shape of recall and precision curves varies less as more normalization knowledge is injected. As for the links, recall increases, but more slowly. Precision starts with an increase (*lem-med* step) and then decreases, especially at the *syno-gen* step.

Up to 31% of the MeSH concepts are correctly placed under their ancestors (tolerant approach with *syno-gen* or *syno-gen-f*); and the suggestions for term positioning are correct in up to 65% of the cases (tolerant approach with *lem-med*).

The induction of new links seems to be more sensitive to the injection of more linguistic knowledge, especially as for synonym resources. The precision of links decreases drastically with general-language synonyms, and is slightly improved with filtered synonyms. The evolution of recall is inverse: it increases in a large proportion. Terms placed, in contrast, have a much flatter evolution.

5 Human analysis of 'new' relations

The evaluations presented in the previous section quantify the match between the induced relations and existing MeSH relations. However, they give no explanation for the fact that 60% to 70% of the induced relations are not considered relevant by the MeSH. In the remainder of this paper, we examine why these terms are not hierarchically related in the MeSH, and what kinds of relations exist between them.

Adding modifiers or complements to a term generally produces a more specific term, e.g., *"infarctus"*, *"infarctus du myocarde"* (Engl. *'infarction'*, *'myocardial infarction'*). The *head* of this specific term is then the initial term (here, *"infarctus"*); we shall call *expansion* the rest of the specific term. Given two terms linked with an 'extra-MeSH relation', we therefore examine the position of the words of the 'parent' term in the 'child' term. We divide these extra-MeSH relations accordingly into three sets: (1) the parent concept is in the *head* position in the child concept: *absorption*/*absorption intestinale*; (2) the parent concept is in the tail (*expansion*) position in the child concept: *abdomen*/*tumeur abdomen*; (3) *other* types of positions. Each set of relations is sampled by selecting a 20% subset randomly, both from the basic normalization file (*basic*) and from the inflectional and derivational normalizations file (*stem-med*). Table 2 presents the number of analyzed relations (total = 194). The analyzed examples discussed in the rest of the text are displayed on tables 3 and 4.

Table 2: Relations to analyze: sample sizes.

Normalizations	Head	Expan.	Other
basic	22	31	14
stem-med	37	57	33

We encountered a few normalization errors due to overzealous derivational knowledge (table 3:5). For instance, *"contracture"* (a muscle disease) and *"contraction"* (normal muscle function) have both been stemmed to the same base word; the adjective *"biologique"* is derived from the noun *"biologie"*, but its sense is generally more specific than *"biologie"*.

In the remainder of this section, we first examine issues encountered when trying to identify the head of each term (section 5.1), then review in turn each analyzed subset: head (section 5.2), expansion (section 5.3) and other relations (section 5.4).

5.1 Finding the head

Terms are generally noun phrases, which can take a more or less degenerate form depending on the terminology designers. In French, the semantic head of a noun phrase is usually located at the beginning of this phrase (this contrasts with English, where the semantic head is generally at the end of NPs). As is often the case with terms, MeSH terms do not include determiners, so the semantic head is usually the first word. We therefore rely on a heuristic for determining 'head' and 'expansion' subsets: the head is the first word of the term, and the expansion is the last word. This is correct most of the time, but in some cases, listed in table 3:*5.1*, the semantic head is positioned at the end of the term, generally separated with a comma, a tradition sometimes followed in thesauri.

Table 3: Analysis of 'new' relations: *head* class. The number within parentheses references the section where a case is discussed. The English terms are those in the original MeSH thesaurus, with their original capitalization.

Type of issue	Induced parent P	Induced child C
(*5*) Inappropriate derivation	contracture	contraction musculaire
	Contracture	*Muscle Contraction*
	biologie	testament biologique
	Biology	*Living Wills*
(*5.1*) Inverted head	filoviridae	filoviridae, infections
	Filoviridae	*Filoviridae Infections*
	leishmania	leishmania tropica, infection
	Leishmania	*Leishmaniasis*
	quinones	quinone reductases
	Quinones	*Quinone Reductases*
	neurone	neurone moteur, maladie
	Neurons	*Motor Neuron Disease*
	syndrome	bouche main pied, syndrome
	Syndrome	*Hand, Foot and Mouth Disease*
(*5.2*) Translation problem	acide linoleique	acide linoleique alpha
	Linoleic Acids	*alpha-Linolenic Acid*
(*5.2.1*) Head is not genus	acides amines	acides amines, peptides et proteines
	Amino Acids	*Amino Acids, Peptides, and Proteins*
	personnalite	personnalite compulsive
	Personality	*Compulsive Personality Disorder*
	voix	voix oesophagienne
	Voice	*Speech, Esophageal*
(*5.2.2*) Ambiguous head	investissement	investissement (psychanalyse)
	Investments	*Cathexis*
	absorption	absorption cutanee
	Absorption	*Skin Absorption*
	goitre	goitre ovarien
	Goiter	*Struma Ovarii*
	acides	acides pentanoiques
	Acids	*Pentanoic Acids*
	acne	acne rosacee
	Acne Vulgaris	*Acne Rosacea*
(*5.2.3*) Ontological commitments	amyotrophies	amyotrophies spinales enfance
	Muscular Atrophy	*Spinal Muscular Atrophies of Childhood*
	hyperplasie	hyperplasie epitheliale focale
	Hyperplasia	*Focal Epithelial Hyperplasia*
	centre public sante	centre public sante mentale
	Community Health Centers	*Community Mental Health Centers*
	rectocolite	rectocolite hemorragique
	Proctocolitis	*Colitis, Ulcerative*
	penicillines	penicilline g
	Penicillins	*Penicillin G*

These cases must be hand-corrected and distributed into the non-head classes.

5.2 'Head' subset

Let us first discard a case in which there seems to be a translation error (table 3:*5.2*). An examination of the structure of the English MeSH and a search on Web pages show that in the French MeSH, *acide linoleique alpha* should read *acide linolenique alpha*, which is a kind of *acide linolenique* (and not a kind of *acide linoleique*). The induced relation is therefore incorrect; with the correct spelling, the lexical inclusion *"acide linolenique"*/*"acide linolenique alpha"* would reveal a correct hierarchical relation.

5.2.1 The head is not the 'genus' of the term

We encountered cases where the whole term did not have an *is-a* relation with the head as defined above. This happens in two types of situations shown on table 3:*5.2.1*.

The first situation is due to syntactic reasons. In the induced relation in table 3:*5.2.1*, *"acides amines, peptides et proteines"* is an enumeration, with the sense of a logical OR. It is therefore the genus term, of which each of the components (e.g., *"acides amines"*) is a sub-type.

The second situation is due to semantic reasons. Lexical induction of hierarchical relations assumes inheritance of the defining features of the genus term (e.g., a *'fatty acid, essential'* is a kind of *'fatty acid'*). However, it is well known that this is not always true: a *'plaster cat'* is not a *'cat'* (i.e., a mammal, etc.). This is sometimes modeled as a type coercion phenomenon. We found quite a few 'plaster cats' in our terms, two of which are shown next in table 3:*5.2.1*.

For instance, *"personnalite"* here describes 'behavior-response patterns that characterize the individual', whereas *"personnalite compulsive"* describes a mental disorder. Disorders (or diseases) are objects different from behaviors in the MeSH.

5.2.2 Term naming conventions and ambiguous heads

Head ambiguity depends on the choice of term names in the terminology (here, the MeSH thesaurus). Terms like *"absorption"*, *"investissement"*, etc., have specific senses that make them polysemous. To determine a precise sense, these terms have to be specialized by their contexts, as in the examples listed in table 3:*5.2.2*. Here, *"investissement"* alone (Engl. *'investment'*) has the financial sense, whereas in *"investissement (psychanalyse)"*, it has its more generic sense. In a similar way, *"absorption"* has a specific meaning in chemistry, and *"goitre"* alone is a disorder of the thyroid gland. These cases are often non-ambiguous in the original English version of the same terms: for instance, *"investissement (psychanalyse)"*

is a translation of Engl. *'cathexis'*.

A related case occurs when the name of a parent term is underspecified. In the MeSH, the term *"acides"* refers to *'inorganic acids'*;[2] and in medical French, *"acne"* alone means *"acne vulgaris"*: the convention adopted is to use these single words to name the corresponding concepts. Therefore, terms built around these heads do not refer to children of the concept referred to by the head term.

5.2.3 Ontological commitments

Finally, some induced links (see table 3:*5.2.3*), although absent from the MeSH, are potentially correct *is-a* links, but the designers of the MeSH have made a different modeling choice.

A general representational choice in the MeSH, as in other medical terminologies (e.g., SNOMED), is to differentiate on the one hand "signs or symptoms" and on the other hand "diseases" (a more fully characterized pathological state). This is the case for *"amyotrophies"* and *"hyperplasie"* ("signs or symptoms") *vs.* *"amyotrophies spinales enfance"* and *"hyperplasie epitheliale focale"* ("disease" of the nervous system, of the mouth). For some reason, a *"centre public sante mentale"* is considered not to share all the attributes of a general *"centre public sante"*, which prevents them from being in a parent-child relationship: they are only siblings in the MeSH thesaurus. *"Penicillines"*, in the MeSH, have been chosen to refer to a therapeutic class of drugs (under *'antibiotics'*, under *'chemical actions'*), whereas *"penicilline g"* is considered as a chemical substance.

The structuring involved in these instances reflects the ontological commitments of the terminology designers, and cannot be recovered by lexical inclusion.[3]

5.3 'Expansion' subset

When a 'parent' term is in the 'expansion' position (end position) in a 'child' term, we assume that the semantic head of the child term is different from that of the suggested parent; the induced relation is indeed expected not to be *is-a*.

Some of the main error cases found are close to those for the 'head' subset. Among others, we find again enumerations such as *"antineoplasiques et immunodepresseurs"* (table 4:*5.3.1*; see also subsection 5.2.1) and syntactic ambiguity (see also subsection 5.2.2): in table 4:*5.3.2*, the word *"oncogene"* is a noun in the first term and an adjective in the second one.

Many of the relations found in the 'expansion' subset are partitive. In table 4:*5.3.3*, we can find relations between human body parts, a continent and its population groups, and chemical substances.

In some instances (table 4:*5.3.4*), a general type of link between terms can be detected, but in most

Table 4: Analysis of 'new' relations: *expansion* class. The number within parentheses references the section where a case is discussed.

Case	Induced parent P	Induced child C
(*5.3.1*) Head is not genus	immunodepresseurs	antineoplasiques et immunodepresseurs
	Immunosuppressive Agents	*Antineoplastic and Immunosuppressive Agents*
(*5.3.2*) Ambiguous head	oncogene	antigene viral oncogene
	Oncogenes	*Antigens, Viral, Tumor*
(*5.3.3*) Partitive relations	abdomen	muscle droit abdomen
	Abdomen	*Rectus Abdominis*
	amerique centrale	indien amerique centrale
	Central America	*Indians, Central American*
	argent	nitrate argent
	Silver	*Silver Nitrate*
(*5.3.4*) Causal relation	myxome	virus myxome
	Myxoma	*Myxoma virus*
(*5.3.5*) Thematic relation	comportement alimentaire	troubles comportement alimentaire
	Feeding Behavior	*Eating Disorders*
	hopital	capacite lits hopital
	Hospitals	*Hospital Bed Capacity*
	services sante	fermeture service sante
	Health Services	*Health Facility Closure*
	macrophage	activation macrophage
	Macrophages	*Macrophage Activation*
	bovin	pneumonie interstitielle atypique bovin
	Cattle	*Pneumonia, Atypical Interstitial, of Cattle*
(*5.3.6*) Derived adjectives	cubitus	nerf cubital
	Ulna	*Ulna Nerve*
	genes	epreuve complementation genetique
	Genes	*Genetic Complementation Test*

Table 5: Analysis of 'new' relations: *other* class. The number within parentheses references the section where a case is discussed.

Case	Induced parent P	Induced child C
(*5.4.1*) Insertion in parent	bacterie aerobie	bacterie gram-negatif aerobie
	Bacteria, Aerobic	*Gram-Negative Aerobic Bacteria*
(*5.4.2*) Variety of relations	arteres	anevrysme artere iliaque
	Arteries	*Iliac Aneurysm*
	hepatite b	virus hepatite b canard
	Hepatitis B	*Hepatitis B Virus, Duck*
	encephalite	virus encephalite equine ouest
	Encephalitis	*Encephalitis Virus, Western Equine*
	sommeil	troubles sommeil extrinseques
	Sleep	*Dyssomnias*
	irrigation	liquide irrigation endocanalaire
	Irrigation	*Root Canal Irrigants*
	maladie	assurance maladie personne agee
	Disease	*Medicare*
(*5.4.3*) With derived adjectives	cellules	molecule-1 adhesion cellulaire vasculaire
	Cells	*Vascular Cell Adhesion Molecule-1*
	chimie	produits chimiques inorganiques
	Chemistry	*Inorganic Chemicals*
	dent	implantation dentaire sous-periostee
	Tooth	*Dental Implantation, Subperiosteal*
(*5.4.4*) Chemical compounds	cytochrome c	ubiquinol-cytochrome c reductase
	Cytochrome c	*Ubiquinol-Cytochrome-c Reductase*
	diphosphate	uridine diphosphate acide glucuronique
	Diphosphates	*Uridine Diphosphate Glucuronic Acid*
	lysine	histone-lysine n-methyltransferase
	Lysine	*Histone-Lysine N-Methyltransferase*
(*5.4.5*) Syntactic ambiguity	cilie	cellule ciliee externe
	Ciliophora	*Hair Cells, Inner*

other cases (table 4:*5.3.5*), we have what looks like a specific thematic relation between a predicate and its argument. Note that some of the correct expansion relations involve adjectival derivations of nouns: in table 4:*5.3.6* "*cubital*" and "*genetique*" are correctly derived from "*cubitus*" and "*gene*".

5.4 'Other' subset

In this last subset, the 'parent' term can be at any position in the 'child' term other than head or expansion. It can also be non-contiguous, accepting modifiers or some other intervening elements. All these cases are actually similar to those of the 'expansion' subset except those of the form in table 5:*5.4.1* where "*bacterie*" remains the head of the term.

The next examples in table 5:*5.4.2* reproduce the general cases of the 'expansion' subset with additional modifiers. In some of them (table 5:*5.4.3*), adjectival derivation is involved. Some relations, shown on

table 5:*5.4.4*, are characteristic of the language of chemical compounds.

The 'other' subset also hosted a morphosyntactic ambiguity (table 5:*5.4.5*) where the words *"cilie"* (noun, an invertebrate organism) and *"ciliee"* (inflected form of adjective *"cilie"*, which characterizes a type of cell) are conflated by lemmatization. This error is mainly due to the fact that the MeSH is written with unaccented uppercase letters: the adjective is actually spelled *"cilié/ciliée"*, which would be unambiguous here.

6 Discussion

6.1 Lexical inclusion with linguistic normalization

We presented in this paper an experiment in terminology structuring. We tested some 'internal' methods for this task, relying on the detection of lexical inclusion among terms. We consider that a parent term P is lexically included in a child term C if all words of P occur in C, and assume that this is a clue of its being a parent (ancestor) of C. To help this analysis we applied several kinds of normalizations, first basic then making use of morphological knowledge and finally of synonyms.

Whereas basic lexical inclusion detects easily identifiable relations between terms by matching identical words in these terms, linguistic knowledge allows us to obtain hierarchical dependencies between terms that are more based on the 'meanings' of these terms. These semantic similarities were detected through the morphological analysis and synonym resources we applied. Lemmatization adds flexibility with inflectional variants. Morphological stemming allows us to link terms which contain words that, though different, are formally similar and have closely related meanings. In addition, pairs of synonyms help to induce relations between terms by matching words that share strong semantic features, at least in some contexts.

6.2 Evaluating relations between terms

To assess the induced relations we compared them with the original structure of the MeSH. We evaluated both the induced links and the placement of terms. With linguistic resources and depending on the evaluation scheme, up to 37.6% of the links found were correct, and up to 14% of the direct MeSH links were recovered by lexical inclusion. Up to 31% of the terms were correctly placed under their ancestors; and the placement advices were correct in up to 65% of the cases.

Morphological normalization was found to be useful to identify not only already existing relations (section 4.2), but also 'new' relations (section 5). This confirms previous work by Jacquemin & Tzoukermann

(1999). Synonym resources do as well, but a human analysis of induced extra MeSH links remains to be done.

6.3 General *vs.* domain-specific resources

Our observation on the relative contributions of general and domain-specific resources (morphological and/or synonyms) is similar to that made by (Hamon *et al.* 1998). General-language resources allow us to increase recall, while domain-specific resources are better for improving precision. Adapting general-language resources to the domain processed should help to strike a balance between the two. In addition, general-language resources seem to be too 'general', so that only a very small part is really involved in the processing. This is an additional reason why their filtering through domain corpora can be considered useful.

We tested three approaches for filtering synonyms: lexico-syntactic patterns, coordination marks and association strength in a corpus. All these approaches are based on the co-detection of synonyms in a common syntagmatic context. The first two use a lexical and syntactic context, while the last one relies on a 'graphic' window size and statistical measures. This latter approach erroneously confirms, for instance, a synonymy relation between:

- *"dernier"* - *"culot"* (Engl. *'last - bottom (of bottle) / small bottle'*):
  "Le dernier culot reçu date du 13/07/97. (la dernière transfusion sanguine faite ...)"
  (Engl. *'The last bottle received is dated 13/07/97. (the last blood transfusion done...)'*)

- *"signal"* - *"appel"* (Engl. *'signal - call'*):
  "Plus qu'un symptôme parfois très désagréable, le prurit constitue un véritable signal d'appel pour des maladies aussi diverses que la gale, les affections cutanées, les hémopathies malignes ou l'insuffisance rénale chronique."
  (Engl. *'More than a symptom which is often very unpleasant, pruritus constitutes an actual call signal for diseases...'*)

The first pair contains polysemic words which can co-occur in a term. The words in the second pair are semantically close, but constitute here a multiword expression. It might be more relevant to use a distributional approach (Nazarenko *et al.* 2001) which would check instead the *paradigmatic* substitutability of candidate synonyms.

6.4 Beyond the gold standard

The only expected and evaluated relations in this experiment were the hierarchical relations that exist in the original structure of the MeSH thesaurus. Nevertheless, we assume that the methods applied here should also allow us to induce other potentially correct hierarchical relations, as well as other types of relations beyond the original MeSH hierarchy. Therefore, we also presented a human analysis of automatically, lexically-induced term relations that were not found in the terminology from which the terms were obtained (the MeSH thesaurus). However, a more detailed analysis remains to be done before considering the automatic induction of a typology of these 'new' relations.

In our analysis, spurious relations came from several sources. A few cases are due to abusive morphological and synonym normalization; errors in term names (translation errors) were also uncovered. We made a distinction between *head* and *expansion* positions of the *parent* term in its *child*. One would expect that relations where the parent is in head position would be correct; however, this is not always true. The putative head of a term is sometimes not correctly identified because of specific thesaural constructs (the 'comma' form) and chemical constructs (*"quinone reductases"* are kinds of *"reductases"*) which display head inversion, and because of enumerations. An additional situation is that of a term that does not share an *is-a* relation with its syntactic head (the *plaster cat*). Furthermore, the head word may not have a stable meaning: it may be syntactically ambiguous (*"cilie"*), polysemous (*"investissement"*) or underspecified (*"acne"*). The remaining *head* cases reveal specific modeling options, or 'ontological commitments', of the terminology designers: the relations induced might be considered semantically valid, but were discarded in the MeSH because of overall structuring choices. These choices cannot be predicted with the lexical methods used here, and seem to be the most resistant to attempts at automatic derivation. They also show that what is correct is not necessarily useful for a given terminology.

The *expansion* cases may be useful to propose other relations than *is-a*: we displayed partitive relations, but left to further work a classification of the remaining ones. The UMLS semantic network relations (NLM 2003) might be a relevant direction to look into to represent such links.

6.5 Towards more linguistic preprocessing

The occurrences of syntactic ambiguity suggest that morphosyntactic tagging could be useful. The methods specifically designed for detection of syntactic and morpho-syntactic term variants (Bourigault 1994; Jacquemin and Tzoukermann 1999) might then be more efficient and less error-prone. We must stress however that this may not be an easy task, since most of the MeSH terms are not syntactically well-formed (few determiners and prepositions, inverted heads) and contain rare, technical words that are likely to be

absent from most electronic lexicons. Working on an accented version of MeSH terms would give more precise results too. Suggesting accented forms for unknown words has been the subject of previous work (Zweigenbaum and Grabar 2002), and an accented French MeSH has now been prepared by the CISMeF team and by the official MeSH translators at INSERM.

7 Conclusion

In summary, lexical inclusion accounts for a non-negligible part of the hierarchical concept organization in the MeSH thesaurus; and the use of morphological and synonym knowledge significantly increases this proportion. As could have been hypothesized, trying to place a concept at one position in the hierarchy is more successful than finding all the links from this concept to its parents in a polyhierarchical terminology.

A simple analysis of lexical inclusions shows that in many cases a hierarchical dependency between (medical) terms can be detected. This allows us to obtain an important number of hierarchical relations between these terms. This information should be useful when performing a terminology structuring task.

It should be possible to adapt our method to the induction of other types of relations between terms: synonymy (equivalence of the terms detected through morphological and synonym resources) and antonymy (insertion of negation) (Hamon *et al.* 1998; Daille 2003). To detect and evaluate more relations between terms, other methods for terminology structuring may be applied, such as those presented in section 1. Previous work has shown that results from these different methods seem to be complementary (Kavanagh 1995; Grabar and Jeannin 2002). We plan to test them in the same context as the lexical inclusion experiments presented here.

Aknowledgements

We thank Roger Côté for the French version of Snomed, the US National Library of Medicine for making available the ICD and MeSH terminologies (among many others) through the UMLS, INaLF for lending us the series of synonyms of the general-language lexicon *Le Robert*, and Béatrice Daille and Marie-Claude L'Homme for their numerous comments which helped us to clarify the exposition of many points in this paper.

Notes

1 According to the colleagues from whom we obtained this set of synonyms, the edition of the dictionary is probably that of 1979.

2 Note, though, that if *'inorganic acids'* were named this way, it would be impossible to link it by lexical induction to other, more specific types of inorganic acids.

3 They might be amenable to distributional methods if their contexts of occurrence are different enough.

References

Bodenreider, O., Burgun, A. and Rindflesch, T. C. (2001). "Lexically-suggested hyponymic relations among medical terms and their representation in the UMLS". In *TIA 2001 – Terminologie et Intelligence artificielle*, 11–21, Nancy: INIST-CNRS.

Bourigault, D. (1994). "Extraction et structuration automatiques de terminologie pour l'aide à l'acquisition de connaissances à partir de textes". In *Proceedings of the 9 th Conference RFIA-AFCET*, 1123–1132, Paris, France: AFCET.

Cimino, J. J. (1996). "Coding systems in health care". In van Bemmel, J. H. and McCray, A. T. (eds.), *Yearbook of Medical Informatics '95 — The Computer-based Patient Record*, 71–85. Stuttgart: Schattauer.

Côté, R. A. (1996). *Répertoire d'anatomopathologie de la SNOMED internationale, v3.4*. Université de Sherbrooke, Sherbrooke, Québec.

Daille, B. (2003). "Conceptual structuring through term variations". In Bond, F., Korhonen, A., McCarthy, D. and Villavicencio, A. (eds.), *Proceedings ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, 9–16, Sapporo, Japan.

Darmoni, S. J., Leroy, J.-P., Thirion, B., Baudic, F., Douyere, M. and Piot, J. (2000). "CISMeF: a structured health resource guide". *Methods of Information in Medicine* **39**(1), 30–35.

Grabar, N. and Jeannin, B. (2002). "Contribution de différents outils à la construction d'une terminologie pour la recherche d'information". In Bachimont, B. (ed.), *Proceedings of the 13 th French Knowledge Engineering Workshop*, Rouen, France.

Grabar, N. and Zweigenbaum, P. (forthcoming). "Productivité à travers domaines et genres : dérivés adjectivaux et langue médicale". *Langue Française* (140), 102–125.

Grabar, N., Zweigenbaum, P., Soualmia, L. and Darmoni, S. J. (2003). "Matching controlled vocabulary words". In Baud, R., Fieschi, M., Le Beux, P. and Ruch, P. (eds.), *The New Navigators: from Professionals to Patients,Proceedings Medical Informatics Europe*, volume 95 of *Studies in Health Technology and Informatics*, 445–450, Amsterdam: IOS Press.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Natural Language Processing and Machine Translation. London: Kluwer Academic Publishers.

Hamon, T., Nazarenko, A. and Gros, C. (1998). "A step towards the detection of semantic variants of terms in technical documents". In Boitet, C. (ed.), *Proceedings of the 17 th COLING*, 498–504, Montréal, Canada.

Hearst, M. A. (1992). "Automatic acquisition of hyponyms from large text corpora". In Zampolli, A. (ed.), *Proceedings of the 14 th COLING*, 539–545, Nantes, France.

INSERM (2000). *Thésaurus Biomédical Français/Anglais*. Institut National de la Santé et de la Recherche Médicale, Paris.

Jacquemin, C. and Bourigault, D. (2002). "Term extraction and automatic indexing". In Mitkov, R. (ed.), *Handbook of Computational Linguistics*. Oxford: Oxford University Press. *To appear*.

Jacquemin, C. and Tzoukermann, E. (1999). "NLP for term variant extraction: A synergy of morphology, lexicon, and syntax". In Strzalkowski, T. (ed.), *Natural language information retrieval*, volume 7 of *Text, speech and language technology*, chapter 2, 25–74. Dordrecht & Boston: Kluwer Academic Publishers.

Kavanagh, J. (1995). "The text analyzer: A tool for extracting knowledge from text". Master's thesis, University of Ottawa. Available at `http://www.site.uottawa.ca/~kavanagh/Thesis/`.

Kleiber, G. and Tamba, I. (1990). "L'hyperonymie revisitée : inclusion et hiérarchie". *Langages* **98**, 7–32. L'hyponymie et l'hyperonymie (dir. Marie-Françoise Mortureux).

Lame, G. (2002). *Construction d'ontologie à partir de textes – Une ontologie du droit dédiée à la recherche d'information sur le Web*. PhD dissertation, École des Mines, Paris.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

McDonald, D. D. (1993). “Internal and external evidence in the identification and semantic categorization of proper names”. In Boguraev, B. and Pustejovsky, J. (eds.), *Corpus Processing for Lexical Acquisition*, 61–76. Cambridge, MA: MIT Press.

Nazarenko, A. and Hamon, T. (2002). “Structuration de terminologie : quels outils pour quelles pratiques ?”. *Traitement automatique des langues* **43**(1), 7–18.

Nazarenko, A., Zweigenbaum, P., Habert, B. and Bouaud, J. (2001). “Corpus-based extension of a terminological semantic lexicon”. In Bourigault, D., Jacquemin, C. and L’Homme, M.-C. (eds.), *Recent Advances in Computational Terminology*, 327–351. Amsterdam: John Benjamins.

NLM (2001). *Medical Subject Headings*. National Library of Medicine, Bethesda, Maryland. `http://www.nlm.nih.gov/mesh/meshhome.html`.

NLM (2003). *Semantic Network*. National Library of Medicine, Bethesda, Maryland. Available at `www.nlm.nih.gov-research-umls-META3.HTML`.

Pacak, M. G., Norton, L. M. and Dunham, G. S. (1980). “Morphosemantic analysis of -ITIS forms in medical language”. *Methods of Information in Medicine* **19**, 99–105.

Séguéla, P. and Aussenac, N. (1999). “Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine”. In Teulier, R. (ed.), *Proceedings of the 3 rd French Knowledge Engineering Workshop*, 79–88, École Polytechnique, Palaiseau, France.

Wingert, F., Rothwell, D. and Côté, R. A. (1989). “Automated indexing into SNOMED and ICD”. In Scherrer, J. R., Côté, R. A. and Mandil, S. H. (eds.), *Computerised Natural Medical Language Processing for Knowledge Engineering*, 201–239. Amsterdam: North-Holland.

Zweigenbaum, P., Darmoni, S. J. and Grabar, N. (2001). “The contribution of morphological knowledge to French MeSH mapping for information retrieval”. *Journal of the American Medical Informatics Association* **8**(suppl), 796–800.

Zweigenbaum, P. and Grabar, N. (2002). “Restoring accents in unknown biomedical words: application to the French MeSH thesaurus”. *International Journal of Medical Informatics* **67**(1–3), 113–126.

Zweigenbaum, P., Hadouche, F. and Grabar, N. (2003). “Apprentissage de relations morphologiques en corpus”. In Daille, B. (ed.), *Proceedings of TALN 2003 (Traitement automatique des langues naturelles)*, 285–294, Batz-sur-mer: ATALA IRIN.

Authors' addresses

Natalia Grabar, Pierre Zweigenbaum

STIM, 91, boulevard de l'Hôpital, 75634 Paris Cedex 13, France

{ngr,pz}@biomath.jussieu.fr

http://www.biomath.jussieu.fr/~{ngr,pz}/

About the authors

Natalia Grabar is finishing a Ph.D. which relates medical terminology and the morphology of medical words. She has been contrasting the prevalence of morphological operations in diverse medical corpora and genres and in newspaper articles. She devised methods to acquire morphological knowledge, including inflection, derivation and neoclassical compounds, from structured medical terminologies. Conversely, she applied this morphological knowledge, on the one hand to study how it can help structure a terminology, and on the other hand to help indexing into controlled vocabularies. Aside from her Ph.D. work, she also took part in several French and European projects dealing with Web-based terminology acquisition (SAFIR) and the detection of racist Web sites (PRINCIP).

Pierre Zweigenbaum's work focusses on Natural Language Processing in Medicine, from morphology to syntax to knowledge representation. He has coordinated or taken part in European projects (MENELAS, NLPAD, DOME) on medical language processing, and currently coordinates the French project UMLF on the development of a unified medical lexicon for French. He teaches NLP in Artificial Intelligence, Language Engineering and Medical Informatics curricula, is Chief Editor of the French journal Traitement Automatique des Langues, has been Vice-President of ATALA, the French NLP Society for five years, and is President of the standing committee of TALN, the French conference on Natural Language Processing.