Automatic Acquisition of Synonym Resources and Assessment of their Impact on the Enhanced Search in EHRs

N. Grabar[1, 2, 3]; P.-C. Varoutas[4]; P. Rizand[4]; A. Livartowski[4]; T. Hamon[5]
[1]Centre de Recherche des Cordeliers, Université Paris Descartes, Paris, France;
[2]INSERM, U872, Paris, France;
[3]HEGP AP-HP, Paris, France;
[4]Institut Curie, Département d'Information Médicale, Paris, France;
[5]LIPN-UMR 7030, Université Paris-Nord – CNRS, Villetaneuse, France

Keywords
Natural language processing, artificial intelligence, terminology, hospital information systems, France

Summary
Objective: Currently, the use of natural language processing (NLP) approaches in order to improve search and exploration of electronic health records (EHRs) within healthcare information systems is not a common practice. One reason for this is the lack of suitable lexical resources. Indeed, in order to support such tasks, various types of such resources need to be collected or acquired (i.e., morphological, orthographic, synonymous).
Methods: We propose a novel method for the acquisition of synonymy resources. This method is language-independent and relies on existence of structured terminologies. It enables to decipher hidden synonymy relations between simple words and terms on the basis of their syntactic analysis and exploitation of their compositionality.
Results: Applied to series of synonym terms from the French subset of the *UMLS*, the method shows 99% precision. The overlap between thus inferred terms and the existing sparse resources of synonyms is very low. In order to better integrate these resources in an EHR search system, we analyzed a sample of clinical queries submitted by healthcare professionals.
Conclusions: Observation of clinical queries shows that they make a very little use of the query expansion function, and, whenever they do, synonymy relations are rarely involved.

Correspondence to:
Natalia Grabar
Centre de Recherche des Cordeliers
Inserm U872 EQ 20 SPIM
15 rue de l'Ecole de Medecine
75006 Paris
France
E-mail: natalia.grabar@spim.jussieu.fr

Methods Inf Med 2009; 48: 149–154
doi: 10.3414/ME9213
prepublished: February 18, 2009

1. Introduction

The *Prométhée* framework [1], conceived and deployed at the Institut Curie[a], enables the aggregation, cross-interrogation, visualization and simple statistical analysis of heterogeneous biomedical data, including the hospital's electronic health records (EHRs). Since 2002, this framework is used on a daily basis by the institute's healthcare professionals, in various cognitive contexts [2]. *Prométhée*'s current full-text search engine is limited in that users must, within their queries, employ keywords which, at best, would match words and expressions directly used in health records. Guessing such 'best' keywords is a difficult task: language variation is rich and often unpredictable. Within the cancer field, area of expertise of the Institut Curie, the International Classification of Diseases for Oncology (ICD-O) [3] is used to query structured elements of the EHRs. Currently, the version 3 of the ICD-O is used at the Institut Curie, but terms from previous versions have been used before: they all co-exist in documents within the Prométhée framework. However, within queries, ICD-O terms do not allow to cover all the expressions occurring in EHRs: up to 48% of queries include textual elements.

In this work, we propose a novel approach for acquisition of resources for an improved search within *Prométhée*, when based on free-text queries or queries based on ICD-O categories. We shall first describe existing lexical resources suitable for enhancing *Prométhée*'s full-text search capacities. We then propose a novel high-quality method for deciphering elementary synonymy relations between simple terms or words. This method is language-independent and based on the identification of syntactic invariants. We evaluate the obtained resources. Finally, we perform an analysis of a sample of clinical queries and estimate the effect these acquired synonymy resources would have on the enhancement of the text search of clinical documents through *Prométhée*.

2. Background

Within the NLP field, we distinguish, in respect with term processing levels, among morphological, orthographic and synonymous lexica. Morphological lexica aim at registering links between lexical units that are morphologically and semantically related

[a] www.curie.net

through the inflectional ((*cell, cells*), (*eye, eyes*), (*examination, examinations*)), derivational ((*ovary, ovarian*), (*large, enlargement*), (*cancer, cancerous*)) or compositional ((*carcinoma, adenocarcinoma*), (*bronchiolar, bronchioloalveolar*), (*nucleus, macronucleus*)) processes. Orthographic lexica take into account spelling variations, like (*speciality, specialty*), (*behavior, behaviour*), (*paralyse, paralyze*), which appeared further to some historical or geographical reasons: both spellings are possible and convey the same semantics. Finally, lexica of synonyms record lexical units that are closely related semantically but do not present formal similarity, such as in the following examples: (*cancer, malignant tumor*), (*disease, pathology*), (*abscess, empyema*).

We notice that not all of these resources are equally well described for various languages and/or specialized areas. Currently, the morphological description of languages is the most complete and several languages are provided with at least inflectional lexica [4, 5]. In the biomedical domain, we can mention the widely used lexicon of the *UMLS* system [6] for English, and similar resources for other languages [7, 8]. However, few lexical resources can be found for the description of synonymy or orthographic relations. WordNet [9] proposes synonym relations for English, but the corresponding resources for other languages are not freely available. Besides, it has been shown that general lexica, for instance WordNet, are insufficient for specialized knowledge extraction [10]. Indeed, additional specialized information is crucial to improve their coverage. In order to provide a solution, we propose to exploit specialized terminologies, as several of them are created and continuously updated in the biomedical area, and decipher the "hidden" synonyms they contain. The question of availability of synonymy resources was addressed in two other works: within the WordNet resource [11], and within the biological area [12]. These works are related to our experience, but the method and results we propose are different.

3. Material

We use two types of material: 1) terminologies and resources for the synonymy acquisition and evaluation, and 2) logs of clinical queries for the analysis of impact the acquired resources may have. Our work has been performed using French resources. The examples in French are enclosed in braces "{ }", and their English translation in brackets "( )".

3.1 French UMLS

We use the French subset of the *UMLS* as the original resource from which elementary synonym relations are inferred. The goal of the *UMLS* is to merge various biomedical terminologies (currently over 100). It is a multipurpose resource, which includes concepts and terms from many different source vocabularies developed for very different purposes. The whole *UMLS* contains currently over 1.3 million concepts, 71,883 of them have labels in French. Concepts can group several labels, with same or very close meaning.

3.2 Existing Sets of Synonyms

In addition to the *UMLS*, two available sets of French synonyms were used for comparison purposes:

- set of general synonyms comes from the general-language dictionary *Le Petit Robert* [13], which lists 140,141 pairs of single-word synonyms;
- specialized set of synonyms *AtMedica* extracted from the Masson medical dictionary, formerly online[b]; it contains 831 pairs of single-word synonyms.

3.3 Clinical Queries

A sample of 2833 textual queries, submitted to the *Prométhée* framework by healthcare professionals of the Institut Curie during the 2003–2005 period, is used and analyzed for predicting a possible impact of the acquired resources on the search process.

4. Methods

The basic observation, which gives the foundation to our approach, is that within *UMLS*, terms can show the compositionality through the substitution of their components, for instance:

- `C0016627`: {*grippe, influenza, peste*} in {*Grippe aviaire, Influenza aviaire, Peste aviaire*} (*Avian influenza, Fowl plague*)
- `C0038352`: {*stomacal, gastrique*} (*gastric, stomach*) in {*Contenu gastrique, Contenu stomacal*} (*Gastric contents, Stomach contents*)

We propose a method for the generalization of this observation which allows to acquire a specialized lexicon of elementary synonym relations. We shall refer to the series of synonym terms as *original synonym relations*, and to the series of their substituted components as *induced* or *elementary synonym relations*. As in the given examples, this method exploits the compositional structure of terms and relies on existence of structured terminologies. The notion of compositionality, central within this method, assumes that the meaning of a complex expression is fully determined by 1) its syntactic structure, 2) the meaning of its parts, and (3) the composition function [14]. In order to exploit this principle, terms are first analyzed syntactically into head and expansion components, then specific inference rules are applied. Finally, the obtained results are evaluated and their possible impact on queries is analyzed.

4.1 Preprocessing of Terminology: the Ogmios Platform

The aim of the terminology preprocessing step is to provide syntactic analysis of terms. Such analysis is crucial for our work: the proposed method exploits syntactic dependency relations and is based on syntactic invariants. ▶Figure 1 describes the approach we implement for obtaining such analysis. We use the Ogmios platform[c], which is suitable for the processing of large amounts of data and, moreover, can be tuned for a specialized linguistic domain. First, the TagEN [15] tool is used for the recognition of named entities, in order to help the forthcoming segmentation into words and sentences. Then, POS-tagging (assignment

[b] www.AtMedica.com

[c] www.lipn.univ-paris13.fr/~hamon/ALVIS/Debian/testing

[d] search.cpan.org/~thhamon/Lingua-YaTeA/

© Schattauer 2009

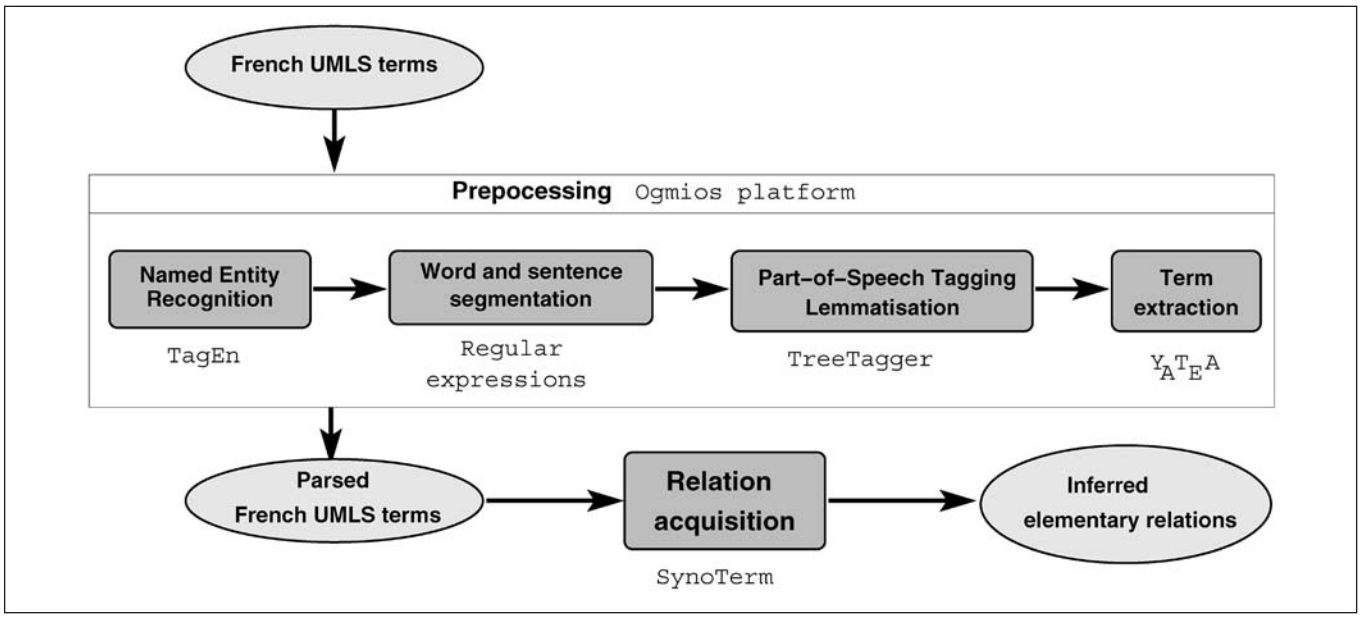

Fig. 1 General flowchart of the method

of syntactical categories like *Noun*, *Verb*, *Adjective*) and lemmatization (definition of the normalized form of words: *cancers* => *cancer*) is realized with TreeTagger [16]. The step of syntactic parsing of terms is carried with the rule-based term extractor YATEA[d]. The syntactic dependency relations between term components (head and expansion) are computed according to assigned POS tags and parsing rules defined within YATEA.

4.2 Acquisition of a Synonym Lexicon

In our previous work [10], we proposed to apply the semantic compositionality principle for inferring synonymy relations (*rel*) between complex terms. We then postulated that the composition process preserves synonymy and that the compositionality principle holds for complex terms. Roughly, this means that if the meanings $\mathcal{M}$ of two complex terms *A rel B* and *A′ rel B* are given by the following formulas:

$$\mathcal{M}(A\ rel\ B) = f(\mathcal{M}(A), \mathcal{M}(B), \mathcal{M}(rel))$$
$$\mathcal{M}(A'\ rel\ B) = f(\mathcal{M}(A'), \mathcal{M}(B), \mathcal{M}(rel))$$

for a given composition function *f*, and if *A* and *A′* are synonymous ($M(A) = M(A')$), then the synonymy of the complex terms can be inferred:

$$\mathcal{M}(A'\ rel\ B) = f(\mathcal{M}(A'), \mathcal{M}(B), \mathcal{M}(rel)) \quad (1)$$
$$= f(\mathcal{M}(A), \mathcal{M}(B), \mathcal{M}(rel)) \quad (2)$$
$$= \mathcal{M}(A\ rel\ B) \quad (3)$$

In the current work, we assume that the inverse function exists and, given synonymous complex terms, can be applied for deducing elementary synonym relations. Our approach takes into account the internal structure of the complex terms. We assume that the syntactic dependency relation between components is preserved through the compositionality principle. Thus, we can infer elementary synonym relations between components of two terms if:

1. parsed terms are synonymous;
2. these components are located at the same syntactic position (head or expansion) and have the same POS tag;
3. the other components within terms are either synonymous or identical.

Parsed terms are represented as a terminological network, within which deduction of the elementary synonym relations is based on the three rules:

- Rule 1: If both terms are synonymous and their head components are identical, then an elementary synonym relation is inferred between expansion components. For instance, we infer the synonym relation {*gastrique, stomacal*} (*gastric, stomach*) from the synonym relation between terms *Contenu gastrique* (*Gastric contents*) and *Contenu stomacal* (*Stomach contents*) where the head component *contenu* (*contents*) is identical (▶Fig. 2).
- Rule 2: If both terms are synonymous and their expansion components are identical, then an elementary synonym relation is inferred between head components. For instance, we can infer the synonym relation {*grippe, influenza*} (*influenza*) from the original synonym relation between terms *Grippe aviaire* (*Avian influenza*) and

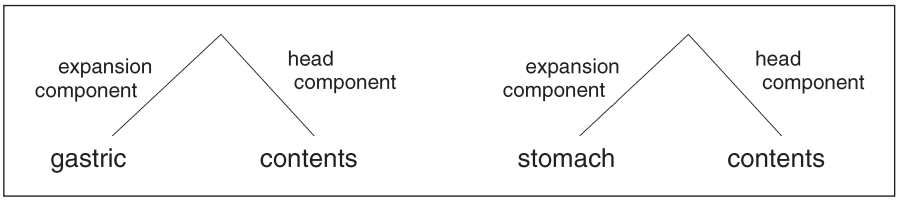

Fig. 2 Parsing syntactic trees of the terms *gastric contents* and *stomach contents* for the acquisition of synonymy relations {*gastric, stomach*}.

© Schattauer 2009

Influenza aviaire (*Avian influenza*) where the expansion component *aviaire* (*avian*) is identical.
- Rule 3: If both terms are synonymous and either their head or expansion components are synonymous, then an elementary synonym relation is inferred. For instance, we infer the synonym relation {*pustuleux, vésiculeux*} (*aphthous, vesicular*) from the synonym relation between terms *Angine pustuleuse* (*Aphthous pharyngitis*) and *Pharyngite vésiculeuse* (*Vesicular pharyngitis*) where the head components {*angine, pharyngite*} (*pharyngitis*) are already known synonyms.

4.3 Evaluation of the Acquired Synonyms

We perform manual validation of the inferred relations between words and simple terms. Each pair is examined, as well as its source series of synonyms. The accuracy of the inferred pairs is thus computed. Moreover, we perform comparison of the acquired synonyms with the existing similar resources: general synonyms from *Le Petit Robert* and medical synonyms from *AtMedica* and *UMLS* and compute the overlap between them.

4.4 Analysis of the Possible Impact on Results

Evaluation of the possible impact of the acquired resources on querying clinical documents at the Institut Curie is performed through the analysis of already submitted queries. Thus, we studied the strategies employed by users for enriching and expanding their queries. It should be noticed that currently *Prométhée* implements no automatic query expansion functionality, which means that users have to perform the expansion by themselves. Within the expanded queries, we analyze the kind of expansion involved implicitly by the users (morphology, synonymy, ...); finally, we analyze whether the acquired resources can help the expansion of these queries.

5. Results and Discussion

5.1 Preprocessing of Terminology: the Ogmios Platform

156,404 terms, corresponding to 71,883 French *UMLS* concepts have been fully parsed through the Ogmios platform. The 54,058 original synonym pairs were used to infer elementary relations.

5.2 Acquisition of a Synonym Lexicon

The three rules have been applied to the terminological network formed with 76,240 original synonym terms (54,058 original synonym pairs) and generated 1196 pairs of elementary synonym relations. The general observation is that only three inferred pairs ({*affection, maladie*} (*affection, disease*), {*maladie, syndrome*} (*disease, syndrome*) and {*cancer, tumeur maligne*} (*cancer, malignant tumor*)) are inferred on 10 to 14 series of original synonyms, while the majority of them are supported by singular series of terms. The acquired synonym pairs can be classified according to their linguistic features, for instance:
- Orthographic variants: {*acathisie, akathisie*} (*akathisia*), {*embolie, embole*} (*embolus*)
- Abbreviations: {*ARNt, ARN transfert*} (*tRNA, transfer RNA*), {*biop, biopsie*} (*biopsy*), {*EEG, electro-encéphalogramme*} (*EEG, electroencephalogram*)
- Named entities: {*bartholin, duverney*}, {*côlon, valsalva*}, {*saint jean, rhumatismale*}
- Ellipse: {*insuffisance artérielle, insuffisance*} (*artery insufficiency, insufficiency*), {*adrénergiques, récepteurs adrénergiques*} (*adrenergic, adrenergic receptor*)
- Scientific vs. popular words: {*maladie, pathologie*} (*disease, pathology*), {*abcès, empyème*} (*abscess, empyema*)
- Morphologically related words: {*spasmodique, spastique*} (*spastic*), {*vermiculaire, vermiforme*} (*vermiform*)
- Most induced synonym pairs link entities for which no common formal features can be observed: {*augmentation volume, hypertrophie*} (*enlargement*), {*grave, sévère*} (*severe*), {*cancer, tumeur maligne*} (*cancer, malignant tumor*)

5.3 Evaluation

5.3.1 Manual Evaluation

Manual evaluation of the totality of generated pairs has shown that 99.3% (n = 1188) are correct, 0.08% (n = 1) rejected and 0.6% (n = 7) not known. The erroneous pair has been generated from the *UMLS* concept C0038814: {*Coup de soleil, Sensibilité au soleil*} (*Solar sensitiveness*) and is due to the fact that one of the original terms (*Coup de soleil*) is not compositional: its correct decomposition and alignment with its synonym is impossible for this reason.

We can see that the efficiency of the proposed method is very high, which is certainly due to the use of controlled terminological data. Moreover, the inferring rules strongly exploit the syntactic scheme within the syntactically analyzed terms. These factors contribute to the acquisition of high-quality synonym pairs. In similar works [17, 18], the inferring of synonym terms does not involve the syntactic analysis of terms. It is based on string matching, and it is interesting to observe that in this case the precision of the induced synonyms drops to 32% [17]. As already mentioned, a very large number of elementary synonymous pairs has been deduced on the basis of only one original pair of synonyms. This contrasts with results we could obtain from the biological terminology *Gene Ontology* [19], where the productivity is so much higher [20], for instance 274 original synonym pairs for inducing {*breakdown, catabolism*} or 240 original pairs for inducing {*catabolism, degradation*}. Indeed, *Gene Ontology* is known to be a highly compositional terminology [21]. One reason for the small productivity observed with French *UMLS* could be that the *UMLS* results from the merging of multiple terminologies, each one implementing different strategies for coining their term labels. On the contrary, *Gene Ontology* has been using the same guidelines for several years[e]. As for the validity of elementary synonyms inferred from singular original term pairs, it should be verified using other terminologies or corpora. A possible weak point of the method is the induction of synonym pairs composed of named entities (n = 26),

[e] www.geneontology.org/GO.usage.shtml

© Schattauer 2009

whose use is reduced to the medical area and, possibly, to certain terms only. For instance:

- {*bartholin, duverney*} inferred from C0004768 *Glandes de Bartholin* and *Glandes de Duverney* (*Bartholin glands, Duverney's gland*),
- {*saint jean, rhumatismale*} inferred from C0152113 *Chorée rhumatismale* and *Chorée de Saint Jean* (*Rheumatic chorea*)

Another possible weak point are ellipses (n = 102): use of *insuffisance* (*insufficiency*) instead of *insuffisance artérielle* (*artery insufficiency*), etc. While widely used in medical, and especially in clinical literature, the use and substituability of elliptic expressions should be studied in more detail: even if they are given as equivalent terms in structured terminologies, the situation can be different in free text documents where, according to contexts, they may correspond to synonymy or hyperonymy relations.

5.3.2 Comparison with Existing Synonym Resources

Comparison between the induced elementary synonymous pairs and existing synonyms shows that overlap is very low. Thus, we found only 36 common pairs with the directly available synonyms within *UMLS*, such as {*tumeur maligne, cancer*} (*malignant tumor, cancer*) or {*saignement, hemorragie*} (*bleeding, hemorrhage*). Thus, the proposed method is useful because it deciphers " hidden " synonyms which are otherwise not accessible. Similarly, we found two common pairs with the *AtMedica* resource and 105 with the *Le Petit Robert* resource. In the first case, results point out the complementarity between different resources of synonyms from the medical area. As for *Le Petit Robert* overlap, the overlap is bigger; still it covers only small part of both resources. It proves that general language resources contain specialized medical vocabulary, although it is not very rich. The difference between them is not surprising as their purpose, as well as addressees and aimed applications, are different. For instance, their use for terminology structuring and knowledge extraction has shown that such general lexica are insufficient for specialized domains [10] and should be completed with specialized resources. Indeed, specialized domains make use of concepts too specific to occur within a general language lexicon.

5.4 Analysis of the Possible Impact on *Prométhée* Queries

2833 textual queries submitted to *Prométhée* have been analyzed. Within these queries, the possibility to enrich them with synonyms or equivalent terms becomes possible through two options:

1. use of disjunction function with the " | " sign
   m(é|e)ningiome: disjunction on the accented character (*é|e*)
   (*ut(é|e)r|endocol|exocol*): additional disjunction on related query terms (*endocol|exocol*)
2. use of boolean queries, and especially of the "OR" operator
   mesotheliome OR *m(é|e)soth(é|e)liome*
   mesonephrique OR *mesonephroide*
   sarcome d'ewing OR *ost(é|e)osarcome*

Among the analyzed queries, less than 200 queries use the expansion option. Within *Prométhée*, these possibilities are processed differently. Normalizing accented characters, which is the most frequent context of use of disjunction function, is performed automatically by the system: when a user's query contains an accented character, *Prométhée* systematically allows it to be replaced by the corresponding unaccented character (i.e., *é* ⇒ (*é|e*)). Thus, even when accents are missing in clinical documents, the system can remedy this. Such normalization is a simple and efficient way to improve the sensitivity of queries. As for boolean queries, they are composed by users using multiple visual elements (widgets) of the *Prométhée* user interface. Finally, use of the disjunction character " | " is done manually by users within the same user interface element. These two last possibilities have the same impact on results.

We first analyzed *Prométhée* queries and tried to categorize them. Not surprisingly, these categories are close enough to those we obtain within the acquired resources:

- Orthographic variants: {*mesotheliome, m(é|e)soth(é|e)liome*} (*mesothelioma*)
- Morphologically related words: {*droit, droite*} (*right*), {*mesonephrique, mesonephroide*} (*mesonephric*)
- Abbreviations: {*CNBPC, carcinome bronchique (à|a) petites cellules*} (*small cell carcinoma of the lung*)
- Named entities: mainly names of medical professionals and of patients
- Synonyms (except the previously cited types): {*forage, biopsie*} (*biopsy*), {*gros noyaux, haut grade nucl(é|e)aire*} (*macronucleus*), {*sarcome d'Ewing, ost(é|e)osarcome*} (*Ewing's sarcoma*)

But we found other types of semantic relations between query words, which can be explained by the cognitive paradigm of search process: users look for notions linked to some therapeutic processes, to drugs used or to possible lexicalization of concepts:

- Drugs: {*adriamycine, vp16*}, {*herceptin, trastuzumab*}
- part-of, is-a: {*ut(é|e)r, endocol, exocol*} (*uterus, endocervix, exocervix*), {*oeil, choroide, r(é|e)tine, globe oculaire*} (*eye, choroids, retina, eyeball*), {*scanner, examen*} (*scan, examination*)
- Associated: {*embolie, phl(é|e)bite, thrombose*} (*embolism, phlebitis, thrombosis*), {*ovaire, ut(é|e)rus*} (*ovary, uterus*), {*CR, consultation, passage, entr(é|e)e*} (*discharge summary, visit, appearance, entrance*)

Let's mention that some queries are very heavy and use over 1100 characters and 50 disjunctions (both | and boolean). We assume it can be useful to apply linguistic resources and dictionaries in order to ease the creation of textual queries, both complex and simple. Moreover we noticed that, as users must perform the query expansion by themselves, they do it according to their training, knowledge, and recently activated concepts in their brain. An automatic option for the query expansion would normalize this process.

In the final stage of the current work, we looked for a possible overlap between the acquired resources and the *Prométhée* queries. Although the semantic typology of the acquired synonymy relations and of the *Prométhée* queries is similar, the recovery between them is small. As a matter of fact, only orthographic and morphological variations could be processed with the currently acquired resources, or with the previously [8] acquired ones. The synonymy variation of queries, as actually done by users, has no relation with the acquired resources, nor with existing resources (*Le Petit Robert*, *AtMedica*): notions involved seem to be specific to the area con-

© Schattauer 2009

cerned (oncology) and more specialized terminologies should be used to possibly detect them. Notice that the analyzed queries could also be used as a possible resource for query expansion. Additionally, we can speculate that users are not completely aware of the possibilities provided by *Prométhée* and the extent to which queries can be expanded. Indeed, sometimes the boolean operator OR is used for composing a syntagmatic (and not paradigmatic) query, like in: *lymphome OR orbitaire* for searching the term *lymphome orbitaire* (*orbital lymphoma*) or using the query *drain OR redon* for searching the term *drain de Redon* (*Redon's drain*). This means that online suggestion of query expansion should be implemented in order to make such possibilities and their correct use known. As for non-synonymous queries (part-of, is-a, associated relations), if taken into account by the *Prométhée* expansion process, they are hardly predictable and should be based on the previously submitted queries and resources similar to co-occurrencies of the *UMLS*. Finally, still for the objective of analysis of the possible impact of the acquired resources, we will analyze textual documents of the Curie databases and shall have a more precise opinion about the idiolect of the Institut Curie .

6. Conclusion and Perspectives

Within healthcare information systems, exploration of EHR content is a current and challenging field. Although the NLP approaches could be suitable to address this need, there is a huge need in various types of linguistic resources. For instance, semantic resources such as lexica of synonyms are missing especially in specialized domains. In this paper, we propose a novel method for filling in this gap and inferring synonymy relations between words and simple terms. This method exploits the compositionality principle and relies on the existence of structured terminologies. It applies a set of rules based on syntactic dependency analysis within terms. In this article, the proposed method has been applied to the *UMLS* subset of French terms. It provides high-quality results: the manual evaluation showed that over 99% of the inferred relations are correct. The comparison with the available resources of synonyms, such as those directly available in *UMLS* and sparse resources like *AtMedica* and *Le Petit Robert*, shows that their overlap is very low. The observed differences seem to indicate that these resources should be combined and used at the same time within NLP tools. First analysis of a sample of queries submitted to *Prométhée* by healthcare professionals shows that these queries are rarely enriched with synonyms, and that these synonyms have but little overlapping with the already existing or acquired resources.

In the near future, we plan to apply the inferred resource to the Institut Curie's EHR corpora (accessible via *Prométhée*), thus further evaluating it during the detection of new synonymy relations between terms used by the institute's healthcare professionals. Additionally, a more thorough analysis of queries will be performed. We shall then better predict a possible impact of such resources in user queries. Finally, the detected (and validated) synonyms will be implemented within the *Prométhée* full-text search engine, where they will play a dual role: 1) during the query parsing phase (the phase during which the question submitted by the user is compiled into a query, understandable by the search engine) they will permit textual query expansion and/or normalization, 2) during the query results visualization phase, they will support on-demand exploratory analysis and classification of textual results. All these efforts contribute to the semantic interoperability in the biomedical area. We have successfully applied the same method to a biological terminological resource Gene Ontology in English [20], which reinforces that it is language and domain-independent. In the future, we would like to test this method on other languages, knowledge domains or terminologies, which is possible as long as 1) the required linguistic processing can be realized and 2) synonym relations between complex terms are available.

References

1. Varoutas PC, Rizand P, Livartowski A. Using category theory as a basis for a heterogeneous data source search meta-engine: The Promethee framework. Lecture Notes in Computer Science (MSFP 2006) 2006; 4019: 381–388.
2. Stroetmann V, Jones T, Ambroise D, et al. Institut Curie, Paris, France: Elios and prométhée. Technical report, eHealth Impact, Information Society and Media, EC study, 2006.
3. Organisation mondiale de la Santé, Genève. International Classification of Diseases for Oncology (ICD-O), 2000.
4. Burnage G. CELEX – A Guide for Users. Centre for Lexical Information, University of Nijmegen, 1990.
5. Hathout N, Namer F, Dal G. An experimental constructional database: the MorTAL project. In: Boucher P (ed). Morphology book. Cambridge, MA: Cascadilla Press; 2001.
6. NLM. UMLS Knowledge Sources Manual. National Library of Medicine, Bethesda, Maryland, 2007. www.nlm.nih.gov/research/umls/.
7. Schulz S, Romacker M, Franz P, et al. Towards a multilingual morpheme thesaurus for medical free-text retrieval. In: Medical Informatics in Europe (MIE), 1999.
8. Zweigenbaum P, Baud R, Burgun A, et al. Towards a Unified Medical Lexicon for French. In: Medical Informatics in Europe (MIE), 2003.
9. Fellbaum C. A semantic network of English: the mother of all WordNets. Computers and Humanities. EuroWordNet: a multilingual database with lexical semantic network 1998; 32 (2–3): 209–220.
10. Hamon T, Nazarenko A. Detection of synonymy links between terms: experiment and results. In: Recent Advances in Computational Terminology. John Benjamins; 2001. pp 185–208.
11. Smith B, Fellbaum C. Medical WordNet: a new methodology for the Construction and Validation of Information. In: Proc of 20th CoLing. Geneva, Switzerland. 2004. pp 371–382.
12. Poprat M., Beisswanger E, Hahn U. Building a BioWordNet Using WordNet Data Structures and WordNet's Software Infrastructure – A Failure Story. ACL 2008 workshop "Software Engineering, Testing, and Quality Assurance for Natural Language Processing", 2008.
13. Robert.Le nouveau petit Robert. Dictionnaires Le Robert, Paris, 1993.
14. Partee BH. Compositionality. F Landman and F Veltman; 1984.
15. Berroyer JF. Tagen, un analyseur d'entités nommées: conception, développement et évaluation. Mémoire de D.E.A. d'intelligence artificielle, Université Paris-Nord, 2004.
16. Schmid H. Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK; 1994. pp 44–49.
17. Hole W, Srinivasan S. Discovering missed synonymy in a large concept-oriented metathesaurus. In: AMIA 2000. pp 354–358.
18. Verspoor CM, Joslyn C, Papcun GJ. The gene ontology as a source of lexical semantic knowledge for a biological natural language processing application. In: SIGIR workshop on Text Analysis and Search for Bioinformatics; 2003. pp 51–56.
19. Gene Ontology Consortium . Gene Ontology: tool for the unification of biology. Nature genetics 2000; 25: 25–29.
20. Grabar N, Jaulent MC, Hamon T. Combination of endogenous clues for profiling inferred semantic relations: experiments with gene ontology. In: AMIA 2008, Washington, USA. 2008. pp 252–256.
21. Ogren P, Cohen K, Acquaah-Mensah G, Eberlein J, Hunter L. The compositional structure of Gene Ontology terms. In: Pacific Symposium of Biocomputing; 2004. pp 214–225.

© Schattauer 2009