TERM EXTRACTION + TERM CLUSTERING: An Integrated Platform for Computer-Aided Terminology

Didier Bourigault
ERSS, UMR 5610 CNRS
Maison de la Recherche
5 allées Antonio Machado
31058 Toulouse cedex, FRANCE
didier.bourigault@wanadoo.fr

Christian Jacquemin
LIMSI-CNRS
BP 133
91403 ORSAY
FRANCE
jacquemin@limsi.fr

Abstract

A novel technique for automatic thesaurus construction is proposed. It is based on the complementary use of two tools: (1) a Term Extraction tool that acquires term candidates from tagged corpora through a shallow grammar of noun phrases, and (2) a Term Clustering tool that groups syntactic variants (insertions). Experiments performed on corpora in three technical domains yield clusters of term candidates with precision rates between 93% and 98%.

1 Computational Terminology

In the domain of corpus-based terminology two types of tools are currently developed: tools for automatic term extraction (Bourigault, 1993; Justeson and Katz, 1995; Daille, 1996; Brun, 1998) and tools for automatic thesaurus construction (Grefenstette, 1994). These tools are expected to be complementary in the sense that the links and clusters proposed in automatic thesaurus construction can be exploited for structuring the term candidates produced by the automatic term extractors. In fact, complementarity is difficult because term extractors provide mainly multi-word terms, while tools for automatic thesaurus construction yield clusters of single-word terms.

On the one hand, term extractors focus on multi-word terms for ontological motivations: single-word terms are too polysemous and too generic and it is therefore necessary to provide the user with multi-word terms that represent finer concepts in a domain. The counterpart of this focus is that automatic term extractors yield important volumes of data that require structuring through a postprocessor. On the other hand, tools for automatic thesaurus construction focus on single-word terms for practical reasons. Since they cluster terms through statistical measures of context similarities, these tools exploit recurring situations. Since single-word terms denote broader concepts than multi-word terms, they appear more frequently in corpora and are therefore more appropriate for statistical clustering.

The contribution of this paper is to propose an integrated platform for computer-aided term extraction and structuring that results from the combination of *LEXTER*, a Term Extraction tool (Bourigault et al., 1996), and *FASTR*[1], a Term Normalization tool (Jacquemin et al., 1997).

2 Components of the Platform for Computer-Aided Terminology

The platform for computer-aided terminology is organized as a chain of four modules and the corresponding flowchart is given by Figure 1. The modules are:

POS tagging First the corpus is processed by *Sylex*, a Part-of-Speech tagger. Each word is unambiguously tagged and receives a single lemma.

Term Extraction *LEXTER*, the term extraction tool acquires term candidates from the tagged corpus. In a first step, *LEXTER* exploits the part-of-speech categories for extracting maximal-length noun phrases. It relies on makers of frontiers together with a shallow grammar of noun phrases. In a second step, *LEXTER* recursively decomposes these maximal-length noun phrases into two syntactic constituents (Head and Expansion).

Term Clustering The term clustering tool groups the term candidates produced at the

[1] *FASTR* can be downloaded from www.limsi.fr/Individu/jacquemi/FASTR.

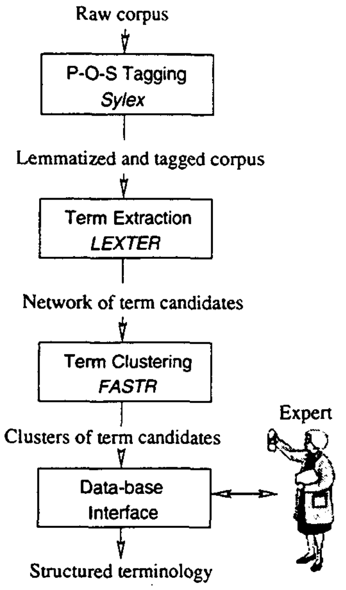


Figure 1: Overview of the platform for computer-aided terminology

preceding step through a self-indexing procedure followed by a graph-based classification. This task is basically performed by *FASTR*, a term normalizer, that has been adapted to the task at hand.

Validation The last step of thesaurus construction is the validation of automatically extracted clusters of term candidates by a terminologist and a domain expert. The validation is performed through a data-base interface. The links are automatically updated through the entire base and a structured thesaurus is progressively constructed.

The following sections provide more details about the components and evaluate the quality of the terms thus extracted.

3 Term Extraction

3.1 Term Extraction for the French Language

Term extraction tools perform statistical or/and syntactical analysis of text corpora in specialized technical or scientific domains. Term candidates correspond to sequences of words (most of the time noun phrases) that are likely to be terminological units. These candidates are ultimately validated as entries by a terminologist in charge of building a thesaurus. *LEXTER*, the term extractor, is applied to the French language. Since French is a Romance language, the syntactic structure of terms and compounds is very similar to the structure of non-compound and non-terminological noun phrases. For instance, in French, terms can contain prepositional phrases with determiners such as: *paroi*$_{\text{Noun}}$ *de*$_{\text{Prep}}$ *l'*$_{\text{Det}}$ *uretère*$_{\text{Noun}}$ (ureteral wall). Because of this similarity, the detection of terms and their variants in French is more difficult than in the English language.

The input of our term extraction tool is an unambiguously tagged corpus. The extraction process is composed of two main steps: Splitting and Parsing.

3.2 Splitting

The techniques of shallow parsing implemented in the Splitting module detect morpho-syntactical patterns that cannot be parts of terminological noun phrases and that are therefore likely to indicate noun phrases boundaries. Splitting techniques are used in other shallow parsers such as (Grefenstette, 1992). In the case of *LEXTER*, the noun phrases which are isolated by splitting are not intermediary data; they are not used by any other automatic module in order to index or classify documents. The extracted noun phrases are term candidates which are proposed to the user. In such a situation, splitting must be performed with high precision.

In order to process correctly some problematic splittings, such as coordinations, attributive past participles and sequences preposition + determiner, the system acquires and uses corpus-based selection restrictions of adjectives and nouns (Bourigault et al., 1996).

For example, in order to disambiguate PP-attachments, the system possesses a corpus-based list of adjectives which accept a prepositional argument built with the preposition *à* (at). These selectional restrictions are acquired through Corpus-Based Endogenous Learning (CBEL) as follows: During a first pass, all the adjectives in a predicative position followed by the preposition *à* are collected. During a second pass, each time a splitting rule has eliminated a sequence beginning with the preposition *à*, the preceding adjective is discarded from the list. Empirical analyses confirm the validity of this procedure. More complex procedures of CBEL are implemented into *LEXTER* in order to acquire nouns sub-categorizing the preposition *à* or the preposition *sur* (on), adjectives sub-categorizing the preposition *de* (of), past participles sub-categorizing the preposition *de* (of), etc.

Ultimately, the Splitting module produces a set of text sequences, mostly noun phrases, which we refer to as Maximal-Length Noun Phrases (henceforth MLNP).

3.3 Parsing

The Parsing module recursively decomposes the maximal-length noun phrases into two syntactic constituents: a constituent in head-position (e.g. *bronchial cell* in the noun phrase *cylindrical bronchial cell*, and *cell* in the noun phrase *bronchial cell*), and a constituent in expansion position (e.g. *cylindrical* in the noun phrase *cylindrical bronchial cell*, and *bronchial* in the noun phrase *bronchial cell*). The Parsing module exploits rules in order to extract two subgroups from each MLNP, one in head-position and the other one in expansion position. Most of MLNP sequences are ambiguous. Two (or more) binary decompositions compete, corresponding to several possibilities of prepositional phrase or adjective attachment. The disambiguation is performed by a corpus-based method which relies on endogenous learning procedures (Bourigault, 1993; Ratnaparkhi, 1998). An example of such a procedure is given in Figure 2.

3.4 Network of term candidates

The sub-groups generated by the Parsing module, together with the maximal-length noun phrases extracted by the Splitting module, are the term candidates produced by the Term extraction tool. This set of term candidates is represented as a network: each multi-word term candidate is connected to its head constituent and to its expansion constituent by syntactic decomposition links. An excerpt of a network of term candidates is given in Figure 3. Vertical and horizontal links are syntactic decomposition links produced by the Term Extraction tool. The oblique link is a syntactic variation link added by the Term Clustering tool.

The building of the network is especially important for the purpose of term acquisition. The average number of multi-word term candidates is 8,000 for a 100,000 word corpus. The feedback of several experiments in which our Term Extraction tool was used shows that the more structured the set of term candidates is, the more efficiently the validation task is performed. For example, the structuring through syntactic decomposition allows the system to underscore lists of terms that share the same term either in head position or in expansion position. Such paradigmatic series are frequent in term banks, and initiating the validation task by analyzing such lists appears to be a very efficient validation strategy.

This paper proposes a novel technique for enriching the network of term candidates through the addition of syntactic variation links to syntactic decomposition links.

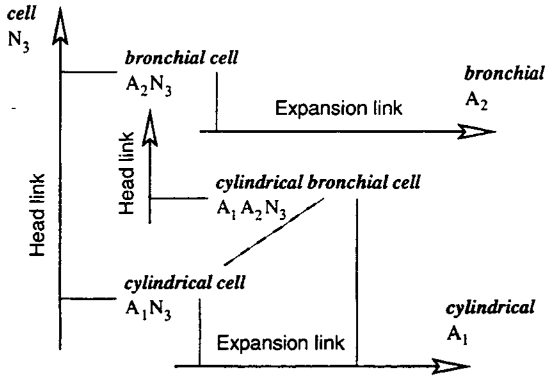


Figure 3: Excerpt of a network of term candidates.

4 Term Clustering

4.1 Adapting a Normalization Tool

Term normalization is a procedure used in automatic indexing for conflating various term occurrences into unique canonical forms. More or less linguistically-oriented techniques are used in the literature for this task. Basic procedures such as (Dillon and Gray, 1983) rely on function word deletion, stemming, and alphabetical word reordering. For example, the index *library catalogs* is transformed into *catalog librar* through such simplification techniques.

In the platform presented in this paper, term normalization is performed by *FASTR*, a shallow transformational parser which uses linguistic knowledge about the possible morpho-syntactic transformations of canonical terms (Jacquemin et al., 1997). Through this technique syntactically and morphologically-related occurrences, such as *stabilisation de prix* (price stabilization) and *stabiliser leurs prix* (stabilize their prices), are conflated.

Term variant extraction in *FASTR* differs from preceding works such as (Evans et al., 1991) because it relies on a shallow syntactic analysis of term variations instead of window-based measures of term overlaps. In (Sparck Jones and Tait, 1984) a knowledge-intensive technique is proposed for extracting term variations. This approach has however never been applied to large scale term extraction because it is based on a full semantic analysis of sentences. Our approach is more realistic because it does not involve large-scale knowledge-intensive interpretation of texts that is known to be unrealistic.

Our approach to the clustering of term can-

Parsing rule

$Noun_1$ Prep $Noun_2$ Adj →

Parse (1)	**Parse (2)**
Head: $Noun_1$	Head: $Noun_1$ Prep $Noun_2$
Exp.: $Noun_2$ Adj	Head: $Noun_1$
Head: $Noun_2$	Exp.: $Noun_2$
Exp.: Adj	Exp.: Adj

Disambiguation procedure:

Look in the corpus for non ambiguous occurrences of the sub-groups:
(a) $Noun_2$ Adj (b) $Noun_1$ Adj (c) $Noun_1$ Prep Noun2

Then choose:

```
if the sub-group (a) has been found, then choose Parse (1)
else if the sub-groups (b) or (c) have been found, then choose Parse (2)
else choose Parse (1)
```

Figure 2: An ambiguous parsing rule and associated disambiguation procedure

didates is to group the output of *LEXTER*, by conflating term candidates with other term candidates instead of conflating corpus occurrences with controlled terms. Our technique can be seen as a kind of self-indexing in which term candidates are indexed by themselves through *FASTR*, for the purpose of conflating candidates that are variants of each other. Thus, the term candidate ***cellule bronchique cylindrique*** (cylindrical bronchial cell) is a variant of the other candidate ***cellule cylindrique*** (cylindrical cell) because an adjectival modifier is inserted in the first term. Through the self-indexing procedure these two candidates belong to the same cluster.

4.2 Types of Syntactic Variation Rules

Because of this original framework, specific variations patterns were designed in order to capture inter-term variations. In this study, we restrict ourselves to syntactic variations and ignore morphological modifications. The variations patterns can be classified into the following two families:

Internal insertion of modifiers The insertion of one or more modifiers inside a noun phrase structure. For instance the following transformation NAInsAj:

$$Noun_1\ Adj_2 \\ \rightarrow Noun_1\ ((Adv^?\ Adj)^{1-3}\ Adv^?)\ Adj_2$$

describes the insertion of one to three adjectival modifiers inside a Noun-Adjective structure in French. Through this transformation, the term candidate ***cellule bronchique cylindrique*** (cylindrical bronchial cell) is recognized as a variant of the term candidate ***cellule cylindrique*** (cylindrical cell). Other internal modifications account for adverbial and prepositional modifiers.

Preposition switch & determiner insertion In French, terms, compounds, and noun phrases have comparable structures: generally a head noun followed by adjectival or prepositional modifiers. Such terms may vary through lexical changes without significant structural modifications. For example NPNSynt:

$$Noun_1\ Prep_2\ Noun_3 \\ \rightarrow Noun_1\ ((Prep\ Det^?)^?)\ Noun_3$$

accounts for preposition suppressions such as ***fibre de collagène/fibre collagène*** (collagen fiber), additions of determiners, and/or preposition switches such as ***revêtement de surface/revêtement en surface*** (surface coating).

The complete rule set is shown in Table 1. Each transformation given in the first column conflates the term structure given in the second column and the term structure given in the third column.

4.3 Clustering

The output of *FASTR* is a set of links between pairs of term candidates in which the target candidate is a variant of the source candidate. In order to facilitate the validation of links by the expert, this output is converted into clusters of term candidates. The syntactic variation links can be considered as the edges of an undirected graph $\mathcal{G}$ whose nodes are the term candidates. A node n_1 representing a term t_1 is connected to a node n_2 representing t_2 if and only if there is a transformation $\mathcal{T}$ such that $\mathcal{T}(t_1) = t_2$ or $\mathcal{T}(t_2) = t_1$. Each connected subgraph $\mathcal{G}_i$ of $\mathcal{G}$ is considered as a cluster of term candidates likely to correspond to similar concepts. (A connected subgraph $\mathcal{G}_i$ is

Table 1: Syntactic variation rules exploited by the Term Clustering tool.

Ident.	Base term	Variant
NAInsAv	Noun$_1$ Adj$_2$	Noun$_1$ ((Adv$^?$ Adj)$^{0-3}$ Adv) Adj$_2$
NAInsAj	Noun$_1$ Adj$_2$	Noun$_1$ ((Adv$^?$ Adj)$^{1-3}$ Adv$^?$) Adj$_2$
NAInsN	Noun$_1$ Adj$_2$	Noun$_1$ ((Adv$^?$ Adj)$^?$
		(Prep$^?$ Det$^?$ (Adv$^?$ Adj)$^?$ Noun) (Adv$^?$ Adj)$^?$ Adv$^?$) Adj$_2$
ANInsAv	Adj$_1$ Noun$_2$	(Adv) Adj$_1$ Noun$_2$
NPNSynt	Noun$_1$ Prep$_2$ Noun$_3$	Noun$_1$ ((Prep Det$^?$)$^?$) Noun$_3$
NPNInsAj	Noun$_1$ Prep$_2$ Noun$_3$	Noun$_1$ ((Adv$^?$ Adj)$^{0-3}$ Prep Det$^?$ (Adv$^?$ Adj)0-3) Noun$_3$
NPNInsN	Noun$_1$ Prep$_2$ Noun$_3$	Noun$_1$ ((Adv$^?$ Adj)$^{0-3}$ (Prep Det$^?$)$^?$ (Adv$^?$ Adj)$^{0-3}$ Noun
		(Adv$^?$ Adj)$^{0-3}$ (Prep Det$^?$)$^?$ (Adv$^?$ Adj)0-3) Noun$_3$
NPDNSynt	Noun$_1$ Prep$_2$ Det4 Noun$_3$	Noun$_1$ ((Prep Det$^?$)$^?$) Noun$_3$
NPDNInsAj	Noun$_1$ Prep$_2$ Det4 Noun$_3$	Noun$_1$ ((Adv$^?$ Adj)$^{0-3}$ Prep Det$^?$ (Adv$^?$ Adj)0-3) Noun$_3$
NPDNInsN	Noun$_1$ Prep$_2$ Det4 Noun$_3$	Noun$_1$ ((Adv$^?$ Adj)$^{0-3}$ (Prep Det$^?$)$^?$ (Adv$^?$ Adj)$^{0-3}$ Noun
		(Adv$^?$ Adj)$^{0-3}$ (Prep Det$^?$)$^?$ (Adv$^?$ Adj)0-3) Noun$_3$

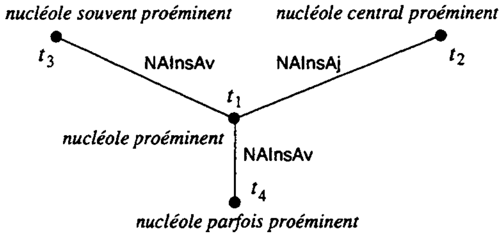

nucléole souvent proéminent
nucléole central proéminent
t3
NAInsAv
NAInsAj
t2
t1
nucléole proéminent
NAInsAv
t4
nucléole parfois proéminent

Figure 4: A sample 4-term cluster.

such that for every pair of nodes (n_1, n_2) in $\mathcal{G}_i$, there exists a path from n_1 to n_2.)

For example, t_1 =*nucléole proéminent* (prominent nucleolus), t_2 =*nucléole central proéminent* (prominent central nucleolus), t_3 =*nucléole souvent proéminent* (frequently prominent nucleolus), and t_4 =*nucléole parfois proéminent* (sometimes prominent nucleolus) are four term candidates that build a star-shaped 4-word cluster illustrated by Figure 4. Each edge is labelled with the syntactic transformation $\mathcal{T}$ that maps one of the nodes to the other.

5 Experiments

Experiments were made on three different corpora described in Table 2. The first two lines of Table 2 report the size of the corpora and the number of term candidates extracted by *LEXTER* from these corpora. The third and fourth lines show the number of links between term candidates extracted by *FASTR* and the number of connected subgraphs corresponding to these links. Finally, the last two lines report statistics on the size of the clusters and the ratio of term candidates that belong to one of the subgraphs produced by the clustering algorithm. Although the variation rules implemented in the Term Structuring tool are rather restrictive (only syntactic insertion has been taken into account), the number of links added to the network of term candidates is noticeably high. An average rate of 10% of multi-word term candidates produced by *LEXTER* belong to one of the clusters resulting from the recognition of term variants by *FASTR*.

Table 3: Frequencies of syntactic variations.

	[Menel.]	[Brouss.]	[DER]
NAInsAv	21%	30%	1%
NAInsAj	33%	25%	5%
NAInsN	23%	21%	13%
ANInsAv	3%	3%	0%
NPNSynt	2%	2%	18%
NPNInsAj	6%	11%	8%
NPNInsN	1%	2%	11%
NPDNSynt	1%	2%	22%
NPDNInsAj	8%	2%	11%
NPDNInsN	2%	2%	11%
Total	100%	100%	100%

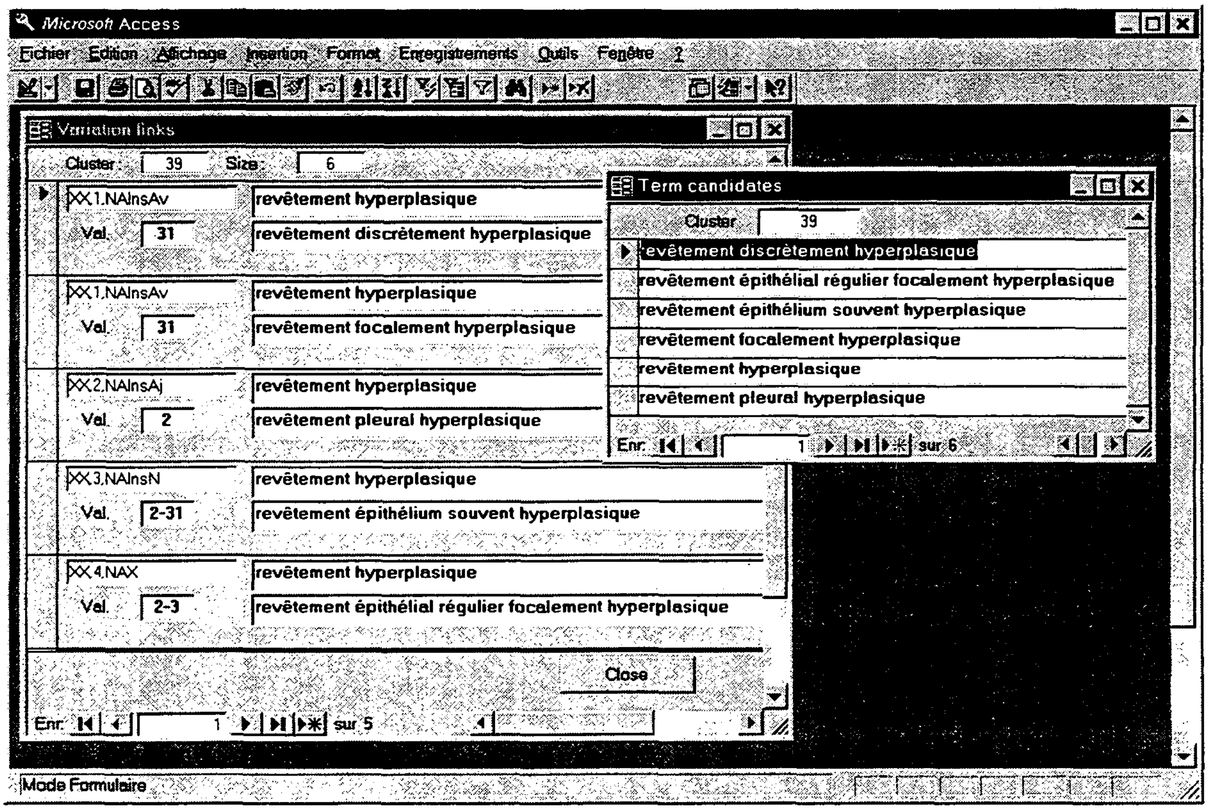
Frequencies of syntactic variations are reported in Table 3. A screen-shot showing the type of validation that is proposed to the expert is given by Figure 5.

6 Expert Evaluation

Evaluation was performed by three experts, one in each domain represented by each corpus. These experts had already been involved in the con-

Table 2: The three corpora exploited in the experiments.

	[Broussais]	[DER]	[Menelas]
Domain	anatomy pathology	nuclear engineering	coronarian diseases
Type of documents	medical reports	technical reports	medical files
Number of words	40,000	230,000	110,000
Number of multi-word term candidates	3,439	14,037	10,155
Number of variation links	240	785	634
Number of clusters	168	556	448
Maximal size of the clusters	10	13	13
Number of term candidates belonging to one cluster	438 (12.7%)	1,349 (9.6%)	1,173 (11.6%)


Microsoft Access
Fichier Edition Affichage Insertion Format Enregistrements Outils Fenêtre ?
Variation links
Term candidates
Close
Mode Formulaire

Variation links

Cluster: 39 Size: 6

XX1.NAInsAv revêtement hyperplasique
Val. 31 revêtement discrètement hyperplasique

XX1.NAInsAv revêtement hyperplasique
Val. 31 revêtement focalement hyperplasique

XX2.NAInsAj revêtement hyperplasique
Val. 2 revêtement pleural hyperplasique

XX3.NAInsN revêtement hyperplasique
Val. 2-31 revêtement épithélium souvent hyperplasique

XX4.NAX revêtement hyperplasique
Val. 2-3 revêtement épithélial régulier focalement hyperplasique

Enr: 1 sur 5

Term candidates

Cluster 39

revêtement discrètement hyperplasique
revêtement épithélial régulier focalement hyperplasique
revêtement épithélium souvent hyperplasique
revêtement focalement hyperplasique
revêtement hyperplasique
revêtement pleural hyperplasique

Enr: 1 sur 6

Figure 5: The expert interface for cluster validation

struction of terminological products through the analysis of the three corpora used in our experiments: an ontology for a case-memory system dedicated to the diagnosis support in pathology ([Broussais]), a semantic dictionary for the Menelas Natural Language Understanding system ([Menelas]), and a structured thesaurus for a computer-assisted technical writing tool ([DER]).

The precision rates are very satisfactory (from 93% to 98% corresponding to error rates of 7% and 2% given in the last line of Table 4), and show that the proposed method must be considered as an important progress in corpus-based terminology. Only few links are judged as conceptually irrelevant by the experts. For example, *image d'embole tumorale* (image of a tumorous embolus) is not considered as a correct variant of *image tumorale* (image of a tumor) because the first occurrence refers to an embolus while the second one refers to a tumor.

The experts were required to assess the proposed links and, in case of positive reply, they were required to provide a judgment about the actual conceptual relation between the connected terms. Although they performed the validation independently, the three experts have proposed very similar types of conceptual relations between term candidates connected by syntactic variation links. At a coarse-grained level, they proposed the same three types of conceptual relations:

Synonymy Both connected terms are considered as equivalent by the expert: *embole tumorale* (tumorous embolus)/*embole vasculaire tumorale* (vascular tumorous embolus). The preceding example corresponds to a frequent situation of elliptic synonymy: the notion of *integrated metonymy* (Kleiber, 1989). In the medical domain, it is a common knowledge that an *embole tumorale* is an *embole vasculaire tumorale*, as everyone knows that *sunflower oil* is a synonym of *sunflower seed oil*.

Generic/specific relation One of the two terms denotes a concept that is finer than the other one: *cellule épithéliale cylindrique* (cylindrical epithelial cell) is a specific type of *cellule cylindrique* (cylindrical cell).

Attributive relation As in the preceding case, there is a non-synonymous semantic relation between the two terms. One of them denotes a concept richer than the other one because it carries an additional attributes: a *noyau volumineux irrégulier* (large irregular nucleus) is a *noyau irrégulier* (irregular nucleus) that is additionally *volumineux* (large).

7 Future Work

This study shows that the clustering of term candidates through term normalization is a powerful technique for enriching the network of term candidates produced by a Term Extraction tool such as *LEXTER*.

In our approach, term normalization is performed through the conflation of specific term variants. We have focused on syntactic variants that involve structural modifications (mainly modifier insertions). As reported in (Jacquemin, 1999), morphological and semantic variations are two other important families of term variations which can also be extracted by *FASTR*. They will be accounted for in order to enhance the number of clustered term candidates. It is our purpose to focus on these two types of variants in the near future.

Acknowledgement

The authors would like to thank the experts for their comments and their evaluations of our results: Pierre Zweigenbaum (AP/HP) on [Menelas], Christel Le Bozec and Marie-Christine Jaulent (AP/HP) on [Broussais], and Henry Boccon-Gibod (DER-EDF) on [DER]. We are also grateful to Henry Boccon-Gibod (DER-EDF) for his support to this work. This work was partially funded by Électricitié de France.

References

Didier Bourigault, Isabelle Gonzalez-Mullier, and Cécile Gros. 1996. Lexter, a natural language processing tool for terminology extraction. In *Seventh EURALEX International Congress on Lexicography (EURALEX96), Part II*, pages 771–779.

Didier Bourigault. 1993. An endogeneous corpus-based method for structural noun phrase disambiguation. In *Proceedings, 6th Conference of the European Chapter of the Association for Computational Linguistics (EACL'93)*, pages 81–86, Utrecht.

Caroline Brun. 1998. Terminology finite-state preprocessing for computational lfg. In *Proceedings, 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 196–200, Montreal.

Table 4: Results of the validation.

	[Broussais]	[Menelas]	[DER]
Number of variation links proposed by the system	240	634	785
Number of variation links validated by the expert	240	227	344
Types of conceptual relation given by the expert			
synonymy	44 (18%)	14 (6%)	136 (40%)
generic/specific	96 (40%)	147 (65%)	121 (35%)
attributive	96 (40%)	61 (27%)	62 (18%)
non relevant	4 (2%)	5 (2%)	25 (7%)

Béatrice Daille. 1996. Study and implementation of combined techniques for automatic extraction of terminology. In Judith L. Klavans and Philip Resnik, editors, *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, pages 49-66. MIT Press, Cambridge, MA.

Martin Dillon and Ann S. Gray. 1983. FASIT: A fully automatic syntactically based indexing system. *Journal of the American Society for Information Science*, 34(2):99–108.

David A. Evans, Kimberly Ginther-Webster, Mary Hart, Robert G. Lefferts, and Ira A. Monarch. 1991. Automatic indexing using selective NLP and first-order thesauri. In *Proceedings, Intelligent Multimedia Information Retrieval Systems and Management (RIAO'91)*, pages 624–643, Barcelona.

Gregory Grefenstette. 1992. A knowledge-poor technique for knowledge extraction from large corpora. In *Proceedings, 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'92)*, Copenhagen.

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery.* Kluwer Academic Publisher, Boston, MA.

Christian Jacquemin, Judith L. Klavans, and Evelyne Tzoukermann. 1997. Expansion of multiword terms for indexing and retrieval using morphology and syntax. In *Proceedings, 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL - EACL'97)*, pages 24–31, Madrid.

Christian Jacquemin. 1999. Syntagmatic and paradigmatic representations of term variation. In *Proceedings, 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, University of Maryland.

John S. Justeson and Slava M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9–27.

George Kleiber. 1989. *Paul est bronzé* versus *la peau de paul est bronzée.* Contre une approche référentielle analytique. In Harro Stammerjohann, editor, *Proceedings, Ve colloque international de linguistique slavo-romane*, pages 109–134, Tübingen. Gunter Narr Verlag. Reprinted in *Nominales*, A. Colin, Paris, 1995.

Adwait Ratnaparkhi. 1998. Statistical models for unsupervised prepositional phrase attachment. In *Proceedings, 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 1079–1085, Montreal.

Karen Sparck Jones and John I. Tait. 1984. Automatic search term variant generation. *Journal of Documentation*, 40(1):50–66.