Discovering Specific Semantic Relationships between Nouns and Verbs in a Specialized French Corpus

Vincent CLAVEAU and **Marie-Claude L'HOMME**
OLST - University of Montreal
C.P. 6128, succ. Centre-Ville
Montréal, QC, H3C 3J7
Canada
{Vincent.Claveau,Marie-Claude.L'Homme}@umontreal.ca

Abstract

Recent literature in computational terminology has shown an increasing interest in identifying various semantic relationships between terms. In this paper, we propose an original strategy to find specific noun-verb combinations in a specialized corpus. We focus on verbs that convey a meaning of realization. To acquire these noun-verb pairs, we use ASARES, a machine learning technique that automatically infers extraction patterns from examples and counter-examples of realization noun-verb pairs. The patterns are then applied to the corpus to retrieve new pairs. Results, measured with a large test set, show that our acquisition technique outperforms classical statistical methods used for collocation acquisition. Moreover, the inferred patterns yield interesting clues on which structures are more likely to convey the target semantic link.

1 Introduction

Recent literature in computational terminology has shown an increasing interest in identifying various semantic relationships between terms. Different strategies have been developed in order to identify pairs of terms that share a specific semantic relationship (such as hyperonymy or meronymy) or to build classes of terms.

However, most strategies are based on "internal" or "external methods" (Grabar and Zweigenbaum, 2002), *i.e.* methods that rely on the form of terms or on the information gathered from contexts. (In some cases, an additional resource, such as a dictionary or a thesaurus, is used during the identification process.) The work reported here infers specific semantic relationships based on sets of examples and counter-examples.

In this paper, the method is applied to a French corpus on computing to find noun-verb combinations in which verbs convey a meaning of realization. The work is carried out in order to assist terminographers in the enrichment of a dictionary on computing that includes collocational information (L'Homme, 2004).

Even though this work is carried out for terminographical and lexicographical purposes, it can certainly be of use in other applications, namely information retrieval. Indeed, such rich semantic links can be used to extend indices or reformulate queries (similar to the work by Voorhees (1994) with WORDNET relations).

2 Objectives

The noun-verb combinations we aim to identify have the following characteristics. They share:

- A syntactic relationship : nouns can be subjects (*e.g.*, "*ordinateur tourne*"; "*computer runs*"); direct objects (*e.g.*, "*configurer l'application*"; "*configure the application*"); or second complement (*e.g.*, "*charger x dans la mémoire*"; "*load x into memory*");
- A valid semantic relationship. The following semantic relationships are sought:
  1. verbs that refer to activities carried out by the nouns (*e.g.*, "*serveur démarre*"; "*server starts*");
  2. verbs that refer to uses of the nouns (*e.g.*, "*activer une option*"; "*activate an option*" or "*naviguer sur Internet*"; "*surf the Internet*");
  3. verbs that refer to activities carried out by means of the nouns (*e.g.*, "*activer l'option au moyen de cette commande*"; "*activate this option with this command*"); and, finally,
  4. verbs that refer to the processes by which nouns are prepared fur use (*e.g.*, "*installer un logiciel*"; "*install an application*").

These noun-verb combinations will hereafter be called *valid N-V pairs*.

The semantic relationships listed above correspond to a set of lexical functions (LFs) defined in (Mel'čuk et al., 1984 1999), *i.e.* LFs used to represent realization (*e.g.*, $Fact_i$, $Real_i$, $Labreal_{ij}$) and to represent the process by which something is prepared ($Prepar$)[1]. (Both of these types of LFs can be combined with others [to create complex lexical functions].) These LFs are opposed to support verbs represented by the LFs $Func_i$, $Oper_i$, $Labor_{ij}$ (*e.g.*, "*créer un fichier*"; "*create a file*"; "*definir une variable*"; "*define a variable*").).

Realization verbs (and verbs denoting the preparation of nouns) were chosen since they are believed to be frequent in technical corpora, such as the corpus of computing used in this experiment. However, this is seen as a first step in order to validate an acquisition process for semantically-related V-N pairs. Other semantic relationships could be sought in the future.

3 Related work

A number of applications have relied on distributional analysis (Harris, 1971) in order to build classes of semantically related terms. This approach, which uses words that appear in the context of terms to formulate hypotheses on their semantic relatedness (Habert et al., 1996, for example), does not specify the relationship itself. Hence, synonyms, co-hyponyms, hyperonyms, *etc.* are not differentiated.

More recent work on terminology structuring has focussed on formal similarity to develop hypotheses on the semantic relationships between terms: Daille (2003) uses derivational morphology; Grabar and Zweigenbaum (2002) use, as a starting point, a number of identical characters.

Up to now, the focus has been on nouns and adjectives, since these structuring methods have been applied to lists of extracted candidate terms (Habert et al., 1996; Daille, 2003) or to lists of admitted terms (Grabar and Zweigenbaum, 2002). As a consequence, relationships considered have been mostly synonymic or taxonomic, or defined as term variations.

On the other hand, other work has been carried out in order to acquire collocations. Most of these endeavours have focused on purely statistical acquisition techniques (Church and Hanks, 1990), on linguisitic acquisition (by the use of Part-of-Speech filters hand-crafted by a linguist) (Oueslati, 1999) or, more frequently, on a combination of the two (Smadja, 1993; Kilgarriff and Tugwell, 2001, for example). It is worth noting that although these techniques are able to identify N-V pairs, they do not specify the relationship between N and V, nor are they capable of focusing on a subset of N-V pairs. The original acquisition methodology we present in the next section will allow us to overcome this limitation.

4 Methodology for finding valid noun-verb pairs

This section is devoted to the description of the methodology and the data we use to acquire semantically related noun-verb pairs. We first describe the specialized corpus used in this experiment. Then, we briefly present ASARES, a pattern inference tool, on which our acquisition strategy relies. Finally, we explain the different steps of the acquisition process.

4.1 Computer science corpus

The French corpus used in our experiments is composed of more than 50 articles from books or web sites specialized in computer science; all of them were published between 1988 and 2003. It covers different computer science sub-domains (networking, managing Unix computers, webcams...) and comprises 600,000 words.

Segmentation, morpho-syntactic tagging and lemmatization have been carried out using the tool CORDIAL[2]. Each word is accompanied by its lemma and Part-of-Speech tag (noun, verb, adjective). Also, the tool indicates inflection (gender and number for nouns, tense and person for verbs) and gives syntactic information (head-modifier) for noun phrases.

4.2 Overview of ASARES

The method used for the acquisition of N-V pairs relies mainly on ASARES, a pattern inference tool. ASARES is presented in detail in (Claveau et al., 2003). We simply give a short account of its basic principles herein.

ASARES is based on a Machine Learning technique, Inductive Logic Programming (ILP) (Muggleton and De-Raedt, 1994), which infers general morpho-syntactic patterns from a set of examples (this set is noted E^+ hereafter) and counter-examples (E^-) of the elements one

[1] However, our interpretation of LFs in this work is much looser, since we admitted verbs that would not be considered to be members of true collocations as Mel'čuk et al. (1984 1999) define them, *i.e.* groups of lexical units that share a restricted cooccurrence relationship.

[2] CORDIAL is a commercial product of Synapse-Développement.

wants to acquire and their context. The contextual patterns produced can then be applied to the corpus in order to retrieve new elements. The acquisition process can be summarized in 3 steps:

1. construction of the sets of examples (and counter-examples);
2. inference of extraction patterns with ASARES; and
3. extraction of N-V pairs from the corpus with the inferred patterns.

ASARES has been previously applied to the acquisition of word pairs sharing semantic relations defined in the Generative Lexicon framework (Pustejovsky, 1995) and called qualia relations (Bouillon et al., 2001). Here, we propose to use ASARES in a quite similar way to retrieve our valid N-V pairs. However, the N-V combinations sought are more specific than those that were identified in these previous experiments.

Formally, ILP aims at inferring logic programs (sets of Horn clauses, noted H) from a set of facts (examples and counter-examples of the concept to be learnt) and background knowledge (B), such that the program H logically entails the examples with respect to the background knowledge and rejects (most of) the counter-examples. This is transcribed by the two logical formulae $B \wedge H \models E^+$, $B \wedge H \not\models E^-$, which set the aim of an ILP algorithm.

In this framework, ASARES infers clauses expressing morpho-syntactic patterns that generalize the structures of sentences containing the target element (examples) but not the structures of the sentences containing counter-examples. The background knowledge encodes information about each word occurring in the example or counter-example, namely the meaning of its tag (*e.g.*, adjective in plural form, infinitive verb).

The main benefits of this acquisition technique lie in the inferred patterns. Indeed, contrary to the more classical statistical methods (Mutual Information, Loglike..., see below) used for collocation acquisition (see (Pearce, 2002) for a review), these patterns allow:

1. understanding of the results, that is, why a specific element has been retrieved or not;
2. highlighting of the corpus-specific structures conveying the target element.

In addition to its explanatory capacity, this symbolic acquisition technique has obtained good results for other acquisition tasks when compared to existing statistical techniques (Bouillon et al., 2002).

4.3 Acquisition process

To infer extraction patterns, ASARES needs a set of examples (E^+) and a set of counter-examples (E^-) of the elements we want to retrieve. In our case, E^+ must thus be composed of (POS-tagged) sentences containing valid N-V pairs; conversely, E^- must be composed of sentences containing non-valid N-V pairs. While this step is tedious and usually carried out manually, the originality of our work lies in the fact that E^+ and E^- are obtained automatically.

To produce the positive examples, we use the existing entries of a terminological database we are currently developing. These entries are thus a kind of *bootstrap* in our acquisition process. More precisely, every N-V pair in which V is defined in the database as a realization verb for N is included. Then, all sentences in our corpus containing this N-V pair are considered as examples and added to E^+. Note that we do not check if each occurrence of the N-V pair actually shares the target semantic link or even a syntactic link in the sentences that are extracted. Some of the examples in E^+ might be incorrect, but ASARES tolerates a certain amount of *noise*.

A totally different technique is needed to produce the E^- set, since no information concerning verbs that are not semantically related is available in the terminological database. To obtain a list of invalid N-V pairs, we acquire them from our corpus using a statistical technique. This produces a list of all N-V pairs that appear in the same sentence, and assigns each a score. Many statistical coefficients exist (Manning and Schütze, 1999); most of them can be easily expressed with the help of a contingency table similar to that reproduced in Table 1 and by noting $S = a + b + c + d$. For example, the

	V_j	V_k, $k \neq j$
N_i	a	b
N_l, $l \neq i$	c	d

Table 1: Contingency table for the pair N_i-V_j

Mutual Information coefficient is defined as:

$$MI = log_2 \frac{a}{(a+b)(a+c)}$$

and the loglike coefficient (Dunning, 1993) as:
$Log = a \log a + b \log b + c \log c + d \log d - (a +$

$b) \log(a+b) - (a+c)\log(a+c) - (b+d)\log(b+d) - (c+d)\log(c+d) + S\log S$.
In the work presented here, we have adopted the Loglike coefficient. From the N-V pair list produced with this method, we have chosen the pairs that obtained the lower Loglike scores. As for the positive examples, we consider that each sentence containing one of the pairs is a counter-example and is added to E^-.

Finally, ASARES is launched with these E^+ and E^- sets; each containing about 2600 sentences. About 80 patterns are then produced; some of them are presented and discussed in section 5.1. These patterns can now be applied to the corpus in order to retrieve valid N-V pairs.

5 Performance evaluation

5.1 Inferred patterns

The inferred patterns give some interesting information about the way the target semantic relationships are expressed in our corpus. While some structures are general, others seem very specific to our corpus.

First of all, proximity is an important factor in valid relationships between nouns and verbs; it can be observed in numerous patterns. For example, the inferred Horn clause:
realization(N,V) :- common_noun(N), contiguous(N,V), N≠V. transcribes the fact that a noun and a verb may be a valid N-V pair if N is a common noun (common_noun(N)) and V is contiguous to N (contiguous(N,V). (Determiners are not taken into account.) This pattern covers the case in which N precedes V, such as "*les utilisateurs lancent tour-à-tour leurs programmes*" ("*users launch in turn their programs*") or in which V precedes N, such as "*il est donc nécessaire d'exécuter la commande* `depmod`" ("*thus, it is necessary to run the* `depmod` *command*". A quite similar clue is given in the following pattern:
realization(N,V) :- near_verb(N,V), suc(V,C), suc(C,N), noun(N), V≠C, N≠C, N≠V. The near_verb(N,V) predicate means that no verb occurs between N and V, and suc(X,Y) that the word X is followed by Y. This clause can be expressed in a more classical way as *V + (anything but a verb) + N* and retrieves pairs like "*C'est un service qui vous permet de vous connecter à Internet*" ("*This is a service that allows you to connect to the Internet*").

Another frequent clue in the patterns produced is, unsurprisingly, that N must be the head of a noun phrase. For example, realization(N,V) :- near_word(N,V), near_verb(N,V), precedes(V,N), noun_ph_head(N), pred(N,C), preposition(C), V≠C, N≠C, N≠V. This pattern means *V + (anything but a verb)? + (preposition) + N head of a noun phrase* and retrieves pairs such as: "*les dispositifs d'impression n'étaient pas contrôlés par ordinateur*" ("*the printing devices were not controlled by computer*").

Prepositions also play an important part and appear frequently in the patterns: realization(N,V) :- near_verb(N,V), pred(N,C), lemma(C,"sur"), V≠C, N≠C, N≠V. (that is *V + (anything but a verb)* + sur + N*) covers for example "*un terminal (...) permet de travailler sur l'ordinateur*" ("*a terminal (...) allows [one/the user] to work on the computer*").

The pattern realization realization(N,V) :- near_verb(N,V), precedes(V,N), pred(N,C), lemma(C,"à"), V≠C, N≠C, N≠V., that is *V + (anything but a verb)* + à + N*, covers "*comment vous connecter à Internet*" ("*How to connect to the Internet*").

realization(N,V) :- near_verb(N,V), precedes(N,V), pred(V,C), lemma(C,"à"), common_noun(N), V≠C, N≠C, N≠V., that is *N + (anything but a verb)* + à + V*, covers "*(...) mode de traitement des données suivant lequel les programmes à exécuter (...)*" ("*...mode of data processing in which the programs to execute...*").

A certain number of patterns express structures more specific to our corpus. For example, the clause:
realization(N,V) :- near_verb(N,V), precedes(V,N), suc(N,C), proper_noun(C), common_noun(N), V≠C, N≠C, N≠V. (that is, *V + (anything but a verb)* + common noun N + (proper noun)*) is very specific to structures including a proper noun, such as the sentence: "*(...) Internet utilise le protocole TCP/IP (...)*" ("*the Internet uses the TCP/IP protocol*").

5.2 Methodology for evaluation

In order to evaluate the quality of the extracted N-V pairs, we are interested in two different measures. The first one expresses the completeness of the set of retrieved N-V pairs, that is, how many valid pairs are found with respect to the total number of pairs which should have been found; this is the recall rate. The second measure indicates the reliability of the set of retrieved N-V pairs, that is, how many valid pairs are found with respect to the total number of retrieved pairs; this is the precision rate (defined below). These two rates were evaluated using a test sample containing all this information.

To construct this test set, we have focused our attention on ten domain-specific terms: *commande* (*command*), *configuration*, *fichier* (*file*), *Internet*, *logiciel* (*software*), *option*, *ordinateur* (*computer*), *serveur* (*server*), *système* (*system*), *utilisateur* (*user*). The terms have been identified as the most specific to our corpus by a program developed by Drouin (2003) and called TERMOSTAT. The ten most specific nouns have been produced by comparing our corpus of computing to the French corpus *Le Monde*, composed of newspaper articles (Lemay et al., 2004). Note that to prevent any bias in the results, none of these terms were used as positive examples during the pattern inference step. (They were removed from the example set.)

For each of these 10 nouns, a manual identification of valid and invalid pairs was carried out. Linguists were asked to analyze the sentences and decide whether the highlighted pairs were valid N-V pairs. Examples of the sentences produced are given in Table 2 for the term *utilisateur* (*user*). A pair is considered as valid if at least one of its occurrences has the desired semantic (and syntactic) relationship (*cf.* section 2).

Finally, 603 of the N-V pairs examined are valid and 4446 are considered not to be valid. The results for each noun are detailed in Table 3.

	valid	non-valid	total
commande	114	346	460
configuration	4	361	365
fichier	21	604	625
option	69	324	393
système	82	605	687
Internet	9	535	544
ordinateur	85	432	517
utilisateur	96	435	531
logiciel	64	440	504
serveur	59	364	423
Total	603	4446	5049

Table 3: Summary of the test set

5.3 Results

To compare the results obtained by our technique with the analysis carried out manually, we use the traditional precision/recall approach. Thus we applied the patterns to the corpus and kept all the pairs retrieved when N is one of the ten specific nouns. The results of the comparison are summarized with the help of confusion matrices like the one presented in Table 4[3].

	actual valid	actual non-valid	Total
predicated valid	TP	FP	PrP
predicated non-valid	FN	TN	PrN
Total	AP	AN	S

Table 4: Confusion matrix

It is important to note here that the values in this confusion matrix depend on a parameter: a detection threshold. Indeed, a single occurrence of a N-V pair with the patterns is not sufficient for a pair to be considered as valid. The threshold, called s, represents the minimal number of occurrences to be detected to consider a pair as valid. The recall and precision rates (respectively R and P), measured on our test set, are thus defined according to s.

In order to represent every possible value of R and P according to s, we draw a recall-precision graph in which each value of P is related to its corresponding value of R. Figure 1 gives the graph obtained when applying the inferred patterns to the test set. For comparison purposes, Figure 1 also indicates the recall-precision graphs obtained by two common statistical techniques for collocation acquisition (the Loglike et Mutual Information coefficients presented in section 4.3). As a baseline, this graph also gives the density, computed as AP/S, which represents the precision that would be obtained by a system deciding randomly if a pair is valid or not.

The recall-precision graph shows that our symbolic technique outperforms the two statistical ones; for a fixed recall, the precision gain is up to 45% with respect to the Loglike results and is even higher with respect the MI coefficient. Thus, our acquisition technique meets the objective of offering assistance to the terminographer, since it provides many reliable N-V pair candidates. However, results show that invalid pairs are also retrieved; thus a manual analysis remains unavoidable.

5.4 Discussion of the results

When examining the retrieved pairs, it appears that most invalid N-V pairs can be classified in

[3]The meaning of the variables is given by the combination of the letters: A means actual, Pr predicated, T true, F false, P positive and N negative.

Examples	Comment
utilisateur___exécuter 162.823 13.907792621292 * MemCheckBoxInRunDlg Autorise les utilisateurs#N# à exécuter#V# un programme 16 bits dans un processus VDM (Virtual DOS Machine) dédié (non partagé) .	Valid syntactic and semantic relationship: "user runs a program"
utilisateur___formater 162.823 0.27168791461736 * AllocateDasd Détermine quels utilisateurs#N# peuvent formater#V# et éjecter les disques_durs amovibles .	Valid syntactic and semantic relationship: "user formats a hard disk"
utilisateur___lancer 162.823 17.9010990569368 * Il utilise donc le numéro de l_ utilisateur#N# réel qui a lancé#V# la commande pour savoir si c_ est bien l_ utilisateur#N# root qui l_ a lancé#V# .	Valid syntactic and semantic relationship: "user launches a command"
Counter-examples	**Comment**
utilisateur___accepter 162.823 0.0059043737128377 * Ces commandes sont absolument essentielles pour pouvoir utiliser le système , mais elles sont assez rébarbatives et peu_d_ utilisateurs#N# acceptent#V# de s_ en contenter .	Syntactic relationship is valid; Semantic relationship is not valid
utilisateur___entourer 162.823 0.41335402801632 * Mais , côté utilisateur#N# , plus on a entouré#V# ou rempli son Macintosh de périphériques , plus grands sont les risques de rencontrer des blocages ou des sautes_d_humeur .	Syntactic relationship is not valid

Table 2: Positive and negative examples with *utilisateur* (Engl. *user*)

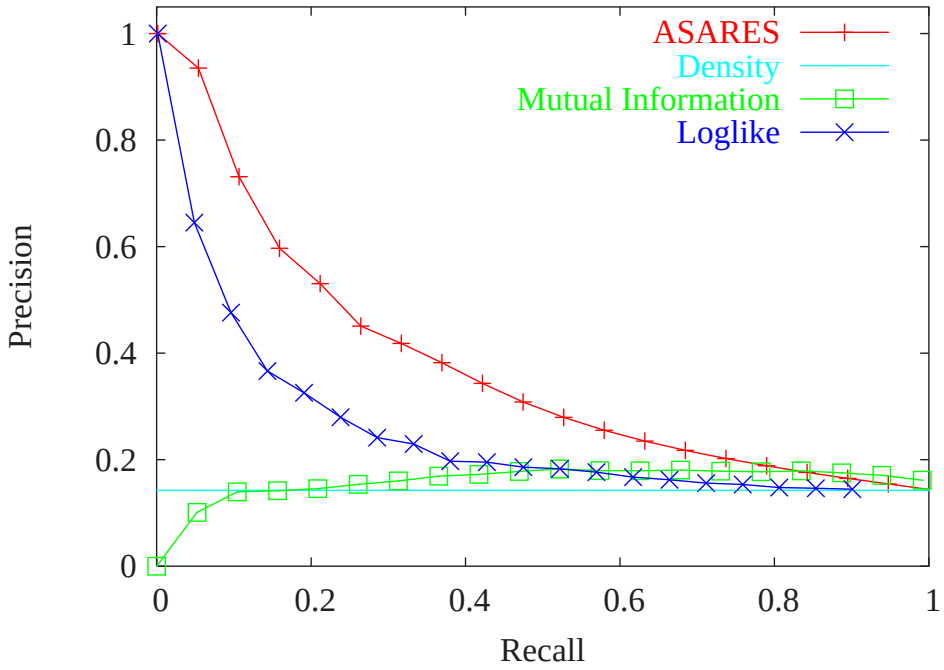

Figure 1: Recall-Precision graph

one of the four following categories.

First, some errors are due to tagging mistakes. For example, in the sentence "*la première solution est d'utiliser la commande* `date`" ("*the first solution is to use the* `date` *command*") the noun `date` was incorrectly tagged as a verb, and the N-V pair *commande-dater* was retrieved by one of the inferred patterns. Even if this kind of error does not come directly from our acquisition technique and does not call into question our approach, it is still a factor that should be taken into account, especially when considering the choice of

the tagger and the quality of the texts composing our corpus.

Secondly, in a few cases, there is no syntactic link between N and V, as is the case with *logiciel-garantir* (*software-guarantee*) in "*cette phase garantit la vérification du logiciel*" ("*this step guarantees the software verification*"). Even if these errors are rare in the retrieved pairs examined, our symbolic extraction system could certainly be enhanced with information about the syntactic function of nouns (subject, direct object...). The learning algorithm could then incorporate this information and produce more relevant patterns.

Thirdly, some N-V pairs are retrieved, although there is no semantic link between N and V, or at least, not a semantic link that would be encoded by a terminographer in a dictionary. This is the case for *ordinateur-savoir* (*computer-know*) in "*l'ordinateur ne sait pas où chercher*" ("*the computer does not know where to look*"). Again, morpho-syntactic information is not always sufficient to distinguish between semantically related pairs and non-semantically (but syntactically) related ones.

Finally, other very frequent errors are caused by the fact that there actually is an interesting semantic link between N and V in a retrieved pair, but not the realization link we are looking for. Indeed, some nouns belonging to specific semantic classes will often cooccur with verbs expressing realization meanings (and thus often appear in valid N-V pairs) while others do not. For example, nouns like *ordinateur* (*computer*) and *utilisateur* (*user*) often appear in valid N-V pairs. On the other hand, other nouns clearly do not appear in combinations with realization verbs (*e.g.*, *configuration*; *Internet*, refer to Table 3).

These last two kinds of error clearly illustrate the limitations of our symbolic approach (but are also very frequent errors in statistical approaches since cooccurrence is not enough to capture subtle semantic distinctions). In fact, they tend to show that our method could be enhanced if it could incorporate richer linguistic information. Indeed, morpho-syntactic information is not always sufficient to operate fine-grained sense distinctions. For example, the two sentences below have the same combination of Part-of-Speech tags: "*vous pouvez utiliser la commande exit*" ("*you can use the exit command*") and "*vous devez choisir l'option 64MB*" ("*you must choose the 64MB option*"); in the first one the underlined N-V pair is valid whereas this is not the case in the second one. Realization verbs for *option* would be *valider* (*validate*) and *activer* (*activate*), for example. Here again, these subtle distinctions could be handled by our symbolic method provided that some semantic information is given on nouns This could be supplied by a semantic tagger.

6 Concluding remarks and future work

We have presented an original acquisition process for noun-verb pairs in which verbs convey a realization meaning. These noun-verb pairs are acquired from a domain-specific corpus of computing. Our acquisition method, which relies on ASARES, a extraction pattern inference technique, produces results that are quite good and could improve manual terminographical work. In particular, our method outperforms classical statistical techniques often used for collocation acquisition. Moreover, the inferred patterns give interesting clues about the structures that are likely to convey the target semantic link.

Many possibilities for future work are suggested by this experiment. Concerning our acquisition process, some adaptations could certainly improve the results, currently limited by the sole use of Part-of-Speech tags and noun-phrase information. As was previously mentioned, syntactic and semantic information could be added to the corpus through a tagging and parsing process. These two enrichments could help to overcome some limitations of our symbolic approach to capture the nature of N-V relationships. In terms of applications, it would be interesting to use a similar technique for the acquisition of other more specific semantic links between nouns and verbs and even between nouns and nouns or other categories of words. These semantic relationships would allow us to complete the description of the terminological units contained in our dictionary of computing. The comparison of the acquisition results and of the inferred patterns could lead to interesting insights.

Aknowlegements

The authors would like to thank Sahara Iveth Carreño Cruz and Léonie Demers-Dion for their help in analyzing the data and Elizabeth Marshman for her comments on a previous version of this paper.

References

Pierrette Bouillon, Vincent Claveau, Cécile Fabre, and Pascale Sébillot. 2001. Using Part-of-Speech and Semantic Tagging for the Corpus-Based Learning of Qualia Structure Elements. In *First International Workshop on Generative Approaches to the Lexicon, GL'2001*, Geneva, Switzerland.

Pierrette Bouillon, Vincent Claveau, Cécile Fabre, and Pascale Sébillot. 2002. Acquisition of Qualia Elements from Corpora – Evaluation of a Symbolic Learning Method. In *3rd International Conference on Language Resources and Evaluation, LREC 02*, Las Palmas de Gran Canaria, Spain.

Kenneth W. Church and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29.

Vincent Claveau, Pascale Sébillot, Cécile Fabre, and Pierrette Bouillon. 2003. Learning Semantic Lexicons from a Part-of-Speech and Semantically Tagged Corpus using Inductive Logic Programming. *Journal of Machine Learning Research, special issue on ILP*, 4:493–525.

Béatrice Daille. 2003. Conceptual structuring through term variation. In *Workshop on Multiword Expressions. Analysis, Acquisition and Treatment. Proceedings of the ACL'03*, Sapporo, Japan.

Patrick Drouin. 2003. Term-extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99–115.

Ted E. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Natalia Grabar and Pierre Zweigenbaum. 2002. Lexically-based terminology structuring. some inherent limits. In *Second Workshop on Computational Terminology, Computerm 2002. Coling 2002*, Taipei, Taiwan.

Benoît Habert, Ellie Naulleau, and Adeline Nazarenko. 1996. Symbolic word clustering for medium-sized corpora. In *Proceedings of the 16th Conference on Computational Linguistics, Coling'96*, Copenhagen, Denmark.

Zellig Harris. 1971. *Structures mathématiques du langage*. Paris: Dunod.

Adam Kilgarriff and David Tugwell. 2001. WORD-SKETCH: Extraction and Display of Significant Collocations for Lexicography. In *Workshop on Collocation: Computational Extraction, Analysis and Exploitation, 39th ACL and 10th EACL Conference*, Toulouse, France.

Chantal Lemay, Marie-Claude L'Homme, and Patrick Drouin. 2004. Two methods for extracting "specific" single-word terms from specialized corpora. Forthcoming.

Marie-Claude L'Homme. 2004. Sélection de termes dans un dictionnaire d'informatique : Comparaison de corpus et critères lexico-sémantiques. In *Euralex 2004. Proceedings*, Lorient, France. Forthcoming.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA, USA.

Igor Mel'čuk, Nadia Arbatchewsky-Jumarie, Léo Elnitsky, Lidija Iordanskaja, Adèle Lessard, Louise Dagenais, Marie-Noëlle Lefebvre, Suzanne Mantha, and Alain Polguère. 1984–1999. *Dictionnaire explicatif et combinatoire du français contemporain, Recherches lexico-sémantiques*, volumes I-IV. Les Presses de l'Université de Montréal, Montréal, QC, Canada.

Stephen Muggleton and Luc De-Raedt. 1994. Inductive Logic Programming: Theory and Methods. *Journal of Logic Programming*, 19-20:629–679.

Rochdi Oueslati. 1999. *Aide à l'acquisition de connaissances à partir de corpus*. Ph.D. thesis, Univesité Louis Pasteur, Strasbourg, France.

Darren Pearce. 2002. A Comparative Evaluation of Collocation Extraction Techniques. In *3rd International Conference on Language Resources and Evaluation, LREC 02*, Las Palmas de Gran Canaria, Spain.

James Pustejovsky. 1995. *The Generative Lexicon*. The MIT Press, Cambridge, MA, USA.

Frank Smadja. 1993. Retrieving Collocations from Text: XTRACT. *Computational Linguistics*, 19(1):143–178.

Ellen M. Voorhees. 1994. Query Expansion Using Lexical-Semantic Relations. In *Proceedings of ACM SIGIR'94*, Dublin, Ireland.