Acquisition of Qualia Elements from Corpora - Evaluation of a Symbolic Learning Method

Pierrette Bouillon[†], Vincent Claveau[*], Cécile Fabre[◇], Pascale Sébillot[*]

[†]TIM/ISSCO - ETI, University of Geneva
40 Bvd du Pont-d'Arve, CH-1205 Geneva, Switzerland
pierrette.bouillon@issco.unige.ch

[*]IRISA
Campus de Beaulieu, 35000 Rennes, France
{vincent.claveau, pascale.sebillot}@irisa.fr

[◇]ERSS, University of Toulouse II
5, allées A. Machado, 31058 Toulouse cedex, France
cecile.fabre@univ-tlse2.fr

Abstract

This paper presents and evaluates a system extracting from a corpus noun-verb pairs whose components are related by a special kind of link: the qualia roles as defined in the Generative Lexicon. This system is based on a symbolic learning method that automatically learns, from noun-verb pairs that are or are not related by a qualia link, rules characterizing positive examples from negative ones in terms of their surrounding part-of-speech or semantic contexts. The qualia noun-verb pair extraction is thus performed by applying the learnt rules on a part-of-speech or semantically tagged text. Stress is put on the quality of the learning when compared with traditional statistical or syntactical-based approaches. The linguistic relevance of the rules is also evaluated through a comparison with manually acquired qualia patterns.

1. Introduction

In the Generative Lexicon (GL) framework (Pustejovsky, 1995), the qualia structure gives access to relational information that prove to be crucial both for linguistic analysis and for NLP applications. In particular, the qualia roles express, in terms of predicative formulae, the basic features of the semantics of nouns (telic, agentive, constitutive, formal). In this model, the noun is linked not only to other nouns via traditional lexical relations (such as meronymy and hyperonymy) but also to verbs. For example, the noun *book* is linked in the telic role to the predicate *read* and in the agentive role to the predicate *write*; hereafter, a noun(N)-verb(V) pair in which V expresses one of the qualia roles of N (like book-read or book-write) is called a qualia pair. Previous works (Fabre and Sébillot, 1999, for example) have demonstrated that these N-V relations provide lexical resources that are very useful for information retrieval applications. Different studies (Grefenstette, 1997; Pustejovsky et al., 1997, for example) also prove that N-V pairs can feed indexes that help a user to select the most interesting occurrences of a given noun in a text. Moreover, a short survey (Vandenbroucke, 2000) at the documentation center of the Banque Bruxelles Lambert (Brussels) shows that verbs that express a qualia relation seem to be more relevant than others for that task. Indeed, in this study, no N-V pairs that are not qualia related were considered as interesting by the documentalists.

Given the lack of lexical resources containing those qualia pairs and the fact that verbs in those pairs may vary considerably from one domain to another (especially in technical domains), methods for corpus-based acquisition of these N-V relations are needed. To simplify matters, two options are usually taken into consideration to acquire N-V links: on one side, statistical approaches extract pairs that are related in a statistically significant way (see (Daille, 1994) for an overview). The problem is that this type of method is not accurate enough to extract precise relations (N-V pairs linked by a qualia relation *versus* other pairs in our case). Another possibility is to use a linguistic approach and to extract the N-V pairs by spotting a set of syntactic structures related to qualia roles (as proposed by Pustejovsky et al. (1993)). In this last case, the advantage is that such patterns can be very precise but a major problem is to define and adapt them to new texts and corpora. In our work, we want to go one step further than the linguistic approach as we have no *a priori* concerning the structures that are likely to convey qualia roles in a given corpus. Thus, we develop and apply a symbolic learning method which automatically produces general rules that explain what, in terms of surrounding context (part-of-speech and semantic tags) in a text, characterizes examples of relevant N-V pairs (*i.e.*, qualia pairs) from irrelevant (*i.e.*, non-qualia) ones. The rules produced this way are then applied to a corpus to exhibit qualia N-V pairs. Therefore, with this system, we aim at combining the precision of linguistic rules (or patterns) in extraction tasks and the flexibility of an automated method.

This paper is divided in three parts: after a presentation of our symbolic learning method, we evaluate the performances of a qualia-pair extraction system based on this learning and compare them with results of other approaches. Lastly, we focus on the linguistic evaluation of the whole system; in other words, what do we learn, and more specifically: (1) what kinds of qualia pairs are or are not retrieved by our system and (2) is it compatible with the

linguistic rules generally proposed by linguists?

2. Learning method description

Our aim is to extract a special kind of semantic relations from a corpus, that is, verbs playing a specific role in the semantic representation of common nouns, as defined in the qualia structure in GL formalism. Trying to infer lexical semantic information from corpora is not new: a lot of work has already been conducted on this subject, especially in the statistical learning domain (see (Habert et al., 1997) or (Pichon and Sébillot, 1997) for surveys of this field). Beside these works, symbolic learning has also led to several studies on the automatic acquisition of semantic lexical elements from corpora (Wermter et al., 1996) during the last years. It is in this last framework that we have chosen to place our project to automatically acquire qualia N-V pairs.

This section is devoted to the presentation of the corpus (and its taggings) used for our experiments and to the description of the symbolic learning method which is the core of our qualia-pair extraction system.

2.1. Corpus and its taggings

Our corpus has first undergone a part-of-speech (POS) tagging (see section 2.1.2.) which aims at providing each word of the text with an unambiguous categorial tag (singular common noun, infinitive verb, etc.). Secondly, in order to have some possibilities to learn what distinguishes qualia pairs from non-qualia ones that appear in exactly the same categorial patterns, semantic tags, that is, tags unambiguously describing the semantic class of each word, have been added (see section 2.1.3.).

2.1.1. The MATRA-CCR corpus

The French corpus used in our experiments is a 700 KBytes handbook of helicopter maintenance, provided by MATRA-CCR Aérospatiale, which contains more than 104,000 word occurrences. This technical corpus has some special characteristics that are especially well suited for our task: it is coherent, that is, its vocabulary and syntactic structures are homogeneous; it contains many concrete terms that are frequently used in sentences together with verbs indicating their telic or agentive roles.

2.1.2. Part-Of-Speech tagging

This corpus has been POS-tagged with the help of annotation tools developed in the MULTEXT project (Armstrong, 1996); sentences and words are first segmented with MTSEG; words are analyzed and lemmatized with MMORPH (Petitpierre and Russell, 1998; Bouillon et al., 1998), and finally disambiguated by the TATOO tool, a hidden Markov model tagger (Armstrong et al., 1995). Each word therefore only receives one POS tag which indicates its morpho-syntactic category (and its gender, number, etc.) with a high precision: less than 2% of errors have been detected when compared to a manually tagged 4,000-word test-sample of the corpus.

2.1.3. Semantic tagging

The semantic tagging is performed on the already POS-tagged MATRA-CCR corpus; we therefore benefit from the disambiguation of polyfunctional words (that is, words that have different syntactic categories, as *règle* in French which can be the indicative of the verb *to regulate* and the common noun *rule*) (Wilks and Stevenson, 1996).

To carry out this tagging, the first step is to build a semantic classification which is used as tagset for the semantic tagging. A lexicon containing every word (the lexicon entries) of the MATRA-CCR corpus is created; it associates with each word all its possible semantic tags. The most relevant tagset for each category must be chosen. For example, some WordNet's (Fellbaum, 1998) most generic classes have been used to classify nouns; irrelevant classes (for our corpus) have been withdrawn and, for large classes, a more precise granularity has been chosen. This has led to 33 classes, hierarchically organized as shown in Figure 1 (WordNet classes not used for tagging are in italics and semantic tags are bracketed). Similar tagsets are built for

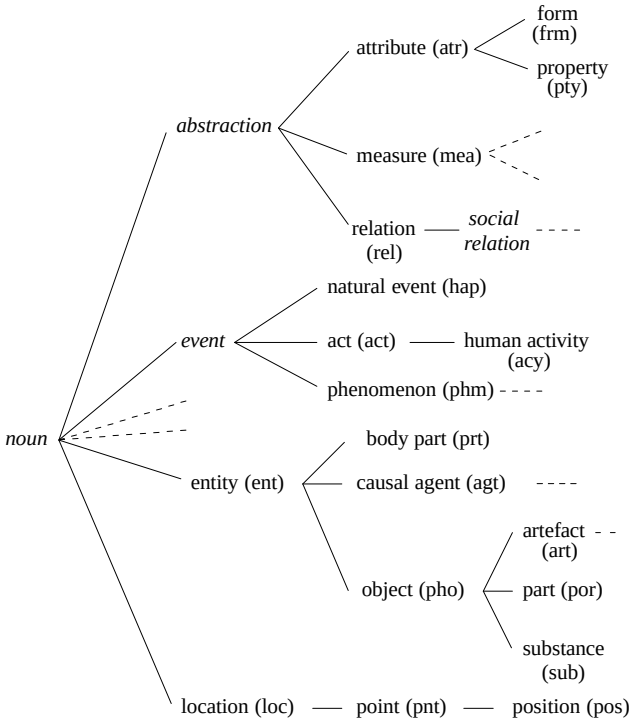

Figure 1: Part of the semantic class hierarchy used for noun tagging

verbs, prepositions, adjectives and other word categories. A more detailed presentation can be found in Bouillon et al. (2001).

In a second step, all those tagsets are used to carry out the semantic tagging of the POS-tagged MATRA-CCR corpus by projecting the semantic tags on the corresponding words. Ambiguities are solved with the help of the probabilistic tagger, following principles described in Bouillon et al. (2000). A 6,000-word sample of the corpus has been chosen to evaluate the semantic tagging precision. It contains 7.78% of ambiguous words; 85% of them have been correctly disambiguated (1.18% of semantic tagging errors).

2.2. Description and settings of the learning method

All those POS and semantic tags in the MATRA-CCR corpus are the contextual key information used by the qualia-pair extraction system that we have developed. This system is built upon an inductive symbolic learning method called inductive logic programming (ILP). The choice of this method is explained in section 2.2.1.; the needed examples and their representations are described in section 2.2.2.; and important settings of the method ensuring the linguistic relevance of the produced rules are given in section 2.2.3..

2.2.1. Learning with ILP

Our selection of a learning method is guided by the fact that this method must not only provide a predictor (this N-V pair is qualia, this one is not), like most of statistical methods, but also infer general rules able to explain the examples, that is, bring linguistically interpretable elements about the predicted qualia relations. This essential explanatory characteristic has motivated our choice of the ILP framework (Muggleton and De-Raedt, 1994) in which programs that are inferred from a set of facts (positive and negative examples of the concept to be learnt) and a background knowledge, are logic programs, that is, sets of Horn clauses. Indeed, ILP relational nature can provide a powerful expressiveness for the still unknown linguistic patterns expressing qualia relations. Moreover, errors inherent in the automatic POS and semantic tagging process previously described make the choice of an error-tolerant learning method essential. The easy handling of data noise in ILP guarantees this robustness.

Most ILP systems provide a way to deal more or less with the form of the generated rules but only some of them enable a total control of this form. Moreover, the particular hierarchical structure of our POS and semantic information makes it essential to use a relational background knowledge processing capable ILP system. For these reasons, we have thus chosen ALEPH[1], an ILP implementation that has already been proven well suited to deal with a large amount of data in multiple domains (mutagenesis, drug structure...) and permits complete and precise customization of all the settings of the learning task.

2.2.2. Example construction

As explained above, ILP algorithms generate rules explaining what characterize positive examples of the concept to be learnt from negative ones. In our case, we want to discriminate qualia N-V pairs from non-qualia ones according to their POS and semantic context. Therefore, our first task consists in building the sets of positive and negative examples, that is, in describing in terms of POS and semantic information the sentences where qualia N-V pairs and non-qualia ones occur. Here is our methodology for their construction.

Given a subset of N-V pairs of our corpus, every occurrence in the text of each pair of this subset is manually annotated as relevant or irrelevant according to Pustejovsky's qualia structure principles. The considered occurrence is then added to the positive example set if it is annotated as relevant, to the negative one otherwise, and the contextual information of this occurrence is added to the background knowledge. The positive and negative examples therefore contain clauses of the form is_qualia(noun_identifier,verb_identifier). where noun_identifier and verb_identifier are the unique identifier of the considered N-V pair occurrence. In the background knowledge, the contextual information is stored in the form of the following clauses:

```
tags(w_1,POS-tag,semantic-tag).
tags(w_2,POS-tag,semantic-tag).
pred(w_2,w_1).
tags(w_3,POS-tag,semantic-tag).
pred(w_3,w_2).
tags(w_4,POS-tag,semantic-tag).
pred(w_4,w_3).
tags(w_5,POS-tag,semantic-tag).
pred(w_5,w_4).
distances(w_4,w_2,distance in words,distance in verbs).
```

where the studied N-V pair w4 w2 occurs in the sentence "w1 w2 w3 w4 w5", pred(x,y) indicates that word y occurs just before word x in the sentence, predicate tags/3 gives the POS and semantic tags of a word, and distances/4 specifies the number of words and the number of verbs between N and V in the sentence. During this step, some word categories (determiners, some adjectives) which are not considered as relevant to bring up information about context of qualia or non-qualia pairs are not taken into account.

3,099 positive examples and 3,176 negative ones are automatically produced this way from the MATRA-CCR corpus. ALEPH's background knowledge is also provided with other information describing the hierarchical relationships among POS and semantic tags. Those relationships encode, for example, the fact that a tag tc_verb_pl indicates a conjugated verb at the plural (conjugated_plural), that can be considered as a conjugated verb (conjugated) or simply a verb (verb).

2.2.3. Hypothesis language

Most ILP systems allow to indicate the form of rules one wants to obtain. Without restricting the expressiveness of the learning process, this important setting, called hypothesis language bias, permits to save computation time and to obtain only well-formed rules with respect to the aimed task, that is, linguistically interpretable rules in our case. For us, a well-formed hypothesis identifying a qualia N-V pair is defined as a clause that gives (semantic and/or POS) information about words (N, V or words occurring in their context) and/or information about respective positions of N and V in the sentence. For example is_qualia(A,B) :- artefact(A), pred(B,C), suc(A,C), auxiliary(C).—which means that a N-V pair (here A-B) is qualia if N is an artefact, V is preceded by an auxiliary verb and N is followed by the same verb—is a well-formed hypothesis. We have therefore to indicate in ALEPH's settings that the predicates artefact/1, pred/2, suc/2, auxiliary/1... can be used to construct a hypothesis.

Another constraint on the hypothesis language is that there can be at most one POS information and one semantic information about a given word. This mean-

[1] http://web.comlab.ox.ac.uk/oucl/research/areas/machlearn/aleph/aleph_toc.html

s that the hypothesis is_qualia(A,B) :- pred(B,C), participle(C), past_participle(C). is not considered as legal since there are two POS information about the word represented by C. Such a redundant information on one word is indeed superfluous and useless since all our POS and semantic information is hierarchically organized (see figure 1): one of the literals is thus more specific than the others and describes the word in a precise enough way. In our example, there is no need to say that C is a participle (participle(C)) if it is known to be a past participle (past_participle(C)). Conversely, the hypotheses is_qualia(A,B) :- pred(B,C), participle(C), action_verb(C). or is_qualia(A,B) :- pred(B,C), past_participle(C), physical_action_verb(C). or even is_qualia(A,B) :- pred(B,C), suc(A,C). are well-formed with respect to our task.

Several other predicates, in particular those dealing with the distances between N and V and their relative positions, are used in the hypothesis language. More than 100 different predicates can thus occur in a hypothesis.

3. Results and validation

In spite of the accurate settings of our learning system, several steps of evaluation are necessary to ensure that the rules that are learnt will provide good results for our task of qualia-pair extraction.

We first present a theoretical evaluation step of our learning process and its parameter setting. A second kind of evaluation is described in section 3.2., which aims at evaluating empirically the performances of a qualia-pair extraction system built upon the learnt rules. Finally, in order to highlight specificities, advantages and drawbacks of our extraction system, we compare our results with well-known basic co-occurrence extraction techniques.

3.1. Noise rate setting and learning

Before using the generated rules to extract qualia pairs from our corpus, we must ensure that these rules have been correctly learnt, that is, our ILP parameters have been correctly set. One of the most important of these parameters is noise, that is, the number of negative examples allowed to be covered by the learnt rules. Indeed, as explained in section 2., handling noisy data is essential for our learning process as for any corpus-based NLP works. Therefore, different values for the noise parameter are tested and their effects are evaluated. The results of the successive experiments (recall and precision rate of the rules, learning time) are compared using a single performance measure, the Pearson coefficient:

$$Pearson = \frac{(TP * TN) - (FP * FN)}{\sqrt{PrP * PrN * AP * AN}}$$

where A = *actual*, Pr = *predicated*, P = *positive*, N = *negative*, T = *true*, F = *false*; a value close to 1 indicates a good learning.

In order to precisely estimate these different characteristics of the learning stage, we perform a 10-fold cross-validation (Kohavi, 1995). We split up therefore the initial set of 3,099 positive examples and 3,176 negative ones into ten subsets. Each subset is alternatively used as testing set while the nine others are used to train the ILP algorithm. The computing time[2], precision, recall and Pearson coefficient averages and standard deviations obtained through these ten learning experiments for the best noise rate (*i.e.*, maximizing the Pearson coefficient) are summarized into table 1.

	Time (seconds)	Precision (%)	Recall (%)	Pearson coeff.
Average	10285	81.3	89.0	0.693
Standard deviation	1440	2.8	2.4	0.047

Table 1: Cross-validation results

A final learning experiment is then conducted with the entire set of examples as training set. A total of nine rules is obtained (see section 4.2. for a detailed presentation of these rules).

3.2. Empirical validation

Beside the theoretical validation described above which only aims at evaluating the learning step performances, we want to know how a system built upon those learnt rules performs in our qualia-pair extraction task. We therefore construct a real-condition test set and compare our ILP-based system results with experts' ones on this set.

3.2.1. Empirical test set construction

The test corpus on which the qualia-pair extraction test is performed is a subset of about 32,000 words of the MATRA-CCR corpus. Despite this relative small size, examining every N-V pair in this subset to see if it is a qualia or non-qualia pair is impossible. We have therefore focused our attention on seven domain relevant nouns: *vis, ´ecrou, porte, voyant, prise, capot, bouchon* (screw, nut, door, indicator signal, plug, cowl, cap). None of these common nouns has been used either as part of a positive or negative N-V pair examples during the learning process.

A Perl program retrieves all N-V pair occurrences including one of the seven studied common nouns and any verb occurring in the same sentence. Then, four GL experts manually tag each pair as relevant or not relevant. Divergences are discussed until complete agreement is reached. Finally, 286 different pairs containing one of the seven nouns are found, 66 of which are qualia pairs.

3.2.2. Empirical validation results

The learnt rules produced by our ILP learning method are applied to the sub-corpus. That is, each N-V pair containing one of the seven test nouns and any verb co-occurring with it within a sentence is tested to see whether it is accepted by one of the learnt rules.

We can decide to consider a N-V pair as relevant if x occurrences of this pair are detected in the test corpus by the learnt rules, that is, if the context of the x occurrences correspond to the general patterns defined by the rules. Of course, if x is high, the precision rate is higher than if x is

[2]Experiments were conducted on a 966MHz PC running Linux.

small, and conversely, for a small x, the recall rate is higher than for a high x. The value of the threshold x is chosen to give the best results for our extraction task according to a certain quality criterion, that is, a single performance measure. In the information retrieval context, in order to easily compare different system performances, the weighted harmonic mean of the recall rate (R) and the precision rate (P), called F-measure and defined as follows is often used:

$$F = \frac{PR}{(1-\alpha)P + \alpha R},\ 0 \leq \alpha \leq 1.$$

The most popular value for α is 0.5 and therefore, the F-measure we use is defined by:

$$F = \frac{2PR}{P+R}.$$

However, in order to compare exhaustively the performances of the different methods, we also use the Pearson coefficient (see section 3.1.) which, unlike F-measure, integrates the fallout rate. The number of detections needed to consider a N-V pair as qualia is therefore chosen to maximize the Pearson coefficient. Finally, this number is found to be 1, that is, a N-V pair is considered as qualia as soon as one occurrence of this pair is covered by one of the learnt rules. Table 2 sums up the results obtained on our empirical test set.

	recall (%)	precision (%)	F-measure	Pearson coeff.
ILP-based system	92.4	62.9	0.748	0.677

Table 2: ILP-based method empirical results

The results show a very good recall rate and a quite good precision rate. Thus, the learnt rules seem to describe precisely enough the qualia concept. Such an ILP-based qualia-pair extraction system can therefore be used on the whole corpus. A detailed discussion about the kinds of N-V pairs correctly retrieved, forgotten or incorrectly found is done in section 4.1..

3.3. Comparison with other approaches

A lot of work has been done in the co-occurrence extraction framework. Most of the studies use either predefined linguistic knowledge such as morpho-syntactic patterns or statistical tools (association criteria, distance measures...). In order to compare the results obtained by our ILP-based system with these different approaches, we have used on the same test set basic statistical methods as well as an entirely manual syntactical-based method to perform our qualia-pair extraction task. We present here the results.

3.3.1. Statistical models

A lot of statistical measures exist and have been applied in numerous domains including biology, sociology and of course lexical analysis. We have used ten well-known of these measures to carry out the qualia-pair extraction task in order to construct an evaluation basis for our ILP method.

All of the statistical indexes we use can be expressed with the help of occurrences of N-V pairs in the corpus. Note that the co-occurrences of nouns and verbs are calculated in the scope of sentences and are based on the lemmas of words. With each N-V pair of the corpus, we can associate a contingency table summing up these co-occurrences as it is shown in table 3, where a is the number of occurrences of the N-V pair (N_j,V_i), b of N-V pairs where the noun is N_j but the verb is not V_i, c of N-V pairs where the verb is V_i but the noun is not N_j, and d of N-V pairs where the noun is not N_j and the verb is not V_i. Let us call S the total number of N-V pair occurrences, that is, S = a + b + c + d.

	V_i	$V_{i'}, i' \neq i$
N_j	a	b
$N_{j'}, j' \neq j$	c	d

Table 3: Contingency table of the N-V pair (N_j,V_i)

We can now easily express some well-known statistical association criteria such as:

- the Kulczinsky coefficient: $Kul = \frac{a}{2}\left(\frac{1}{a+b} + \frac{1}{a+c}\right)$
- the Ochiai coefficient: $Ochiai = \frac{a}{\sqrt{(a+b)(a+c)}}$
- the mutual information coefficient: $MI = log_2 \frac{a}{(a+b)(a+c)}$
- the cubed mutual information coefficient (Daille, 1994): $MI^3 = log_2 \frac{a^3}{(a+b)(a+c)}$
- the McConnoughy coefficient: $MC = \frac{a^2 - bc}{(a+b)(a+c)}$
- the χ^2 test of association: $\chi^2 = \frac{\left(a - \frac{(a+b)(a+c)}{S}\right)^2}{\frac{(a+b)(a+c)}{S}}$
- the loglike coefficient (Dunning, 1993): $loglike = a\,log\,a + b\,log\,b + c\,log\,c + d\,log\,d - (a+b)\,log(a+b) - (a+c)\,log(a+c) - (b+d)\,log(b+d) - (c+d)\,log(c+d) + S\,log\,S$
- the simple matching coefficient: $SMC = \frac{a+d}{S}$
- the Yule coefficient: $Yule = \frac{ad-bc}{ad+bc}$
- the Φ^2 test (Church and Gale, 1991): $\Phi^2 = \frac{(ad-bc)^2}{(a+b)(a+c)(b+c)(b+d)}$

All these statistical measures are then evaluated for each of the 286 N-V pairs containing one of the seven nouns. Similarly to what we do for our ILP method (see section 3.2.2.), we also try to find the coefficient threshold value which maximizes the Pearson coefficient for each of these statistical coefficients. Table 4 indicates the best results obtained.

Only a few statistical measures have good enough results to be used for automatic qualia-pair extraction, and none of them matches the results obtained by our ILP-based system. Of course, the differences between our ILP-based

	recall (%)	precision (%)	F-measure	Pearson coeff.
Kul	36.4	70.6	0.48	0.414
$Ochiai$	42.4	82.4	0.56	0.517
MI	51.5	40	0.45	0.261
MI^3	36.4	92.3	0.522	0.52
MC	36.4	70.6	0.48	0.414
χ^2	37.9	78.1	0.51	0.464
$loglike$	42.4	80	0.554	0.505
SMC	100	25.3	0.385	0.17
$Yule$	53	41.2	0.464	0.279
Φ^2	51.5	47.9	0.496	0.338

Table 4: Statistical method results

and the statistical-based method results can be easily explained by the differences of knowledge used by these two kinds of techniques. Indeed, while statistical models only use word lemma occurrences, our inductive learning process makes the most of categorial and semantic tags but also needs (positive and negatives) examples which is a way to implicitly add linguistic knowledge to the extraction system. Nevertheless, this comparison remains interesting from a pragmatic point of view, more particularly in the balance between the choice of a supervised or unsupervised method and the resulting performances.

3.3.2. Syntactic linkedness

We have also compared our qualia extraction system with an entirely manual approach: a syntactic annotation of the studied text. Each N-V pair occurring within a sentence of the corpus is tagged as syntactically linked (that is, the noun is subject or object of the verb) or not. The underlying idea of this method is to say that a frequent syntactic link between a noun and a verb in a text may indicate a semantic link between this noun and this verb, for example a qualia link.

Therefore, a N-V pair is considered qualia if more than a certain number of its occurrences are detected syntactically linked. This threshold, as for the ILP-based and statistical methods, is chosen to maximize the Pearson coefficient; the value found is 1. Table 5 gives the performances of such a system for our test set.

	recall (%)	precision (%)	F-measure	Pearson coeff.
syntactic linkedness	86.4	79.2	0.826	0.772

Table 5: Syntactic linkedness method results

These results indicate a slightly lower recall rate but a better precision rate than our ILP-based method. This would tend to show that a qualia link is more than a basic syntactic link, but also that our ILP-based method could improve its results, especially its precision rate, by considering syntactic information. However, automatic syntactic annotation remains too noisy to be used without human supervision, and a manual annotation cannot be foreseen for a huge amount of texts. Here again, one should choose between high quality results and automatic or quasi-automatic extraction methods, accordingly with one's goals.

A comparison between the N-V pairs retrieved by this approach and our inductive learning approach is made in section 4.1..

4. Linguistic evaluation of the results

This section is devoted to a linguistically orientated discussion of the different results presented above. More precisely, we first examine the causes of undetected or misdetected qualia N-V pairs during the empirical test of our extraction system. Secondly, we focus our attention on the rules learnt by the ILP learning method, examine the patterns they describe and compare them with observations made manually on the same corpus.

4.1. Retrieved N-V pairs

The results of the ILP-based extraction from the test set are quite promising. On one side, our system detects most of the qualia N-V couples. The five non-detected pairs appear in very rare constructions in our test subcorpus, like *prise-relier* (plug-connect) in *la citerne est reliée à l'appareil par des prises* (the tank is connected to the machine by plugs) where a prepositional phrase (PP) *à l'appareil* (to the machine) is inserted between the verb and the *par*-PP (by-PP). These are clearly too rare to be taken into consideration by the learning method. On the other side, only 8 pairs from the 36 non-qualia pairs detected qualia are not linked syntactically. That means that the ILP algorithm can already reliably distinguish between syntactically and not syntactically linked pairs.

If we compare these results with those obtained by the different statistical methods, the conclusion is obvious: the main problem for statistical methods is silence (good qualia pairs that are not retrieved) while the main problem for the ILP algorithm is precision (non-qualia pairs that are retrieved). But here we should carefully distinguish between two types of errors made by the ILP method. The first ones are caused by constructions that are ambiguous and where N and V can be or cannot be syntactically related, as *enlever-prises* (remove-plugs) in *enlever les shunts sur les prises* (remove the shunts from the plugs). These couples cannot be disambiguated by superficial contextual clues (that is, word tags) and thus show the limitation of learning only from POS and semantic information. However they are very rare in our corpus (8 pairs). On the contrary, all remaining errors seem more related to the parameter settings of the learning method. For example, taking into consideration the number of nouns between the V and the N could avoid a lot of wrong pairs like *poser-capot* (put-up-cover) in *poser les obturateurs capots* (put up cover stopcocks) or *assurer-voyant* (make sure-warning light) in *s'assurer de l'allumage du voyant* (make sure that the warning light is switched on).

The empirical validation can be therefore considered as positive and we can now focus on the last step of the evaluation that consists in assessing the linguistic validity of the generalized clauses.

4.2. Linguistic validation of the learnt rules

For a linguist, the issue is not only to find good examples of qualia relations but also to identify in texts the linguistic patterns that are used to express them. Consequently, the question is: what do the learnt clauses tell us about the linguistic structures that are likely to convey qualia relations between a noun and a verb? We know from previous research (Morin, 1999) made on other types of semantic relations, that a given relation can be instantiated in a large variety of structures, and that this set of structures may greatly vary from one corpus to another. Such research generally focuses on hyperonymy (is-a) and meronymy (part-of) relations, which provide the basic structure of ontologies. Our aim is thus similar, with the additional difficulty that some of the relations we focus on (such as telic or agentive ones) have never been studied extensively on corpora, and are more difficult to identify than more conventional semantic relations.

We are thus faced with a set of nine clauses that we now try to interpret in terms of linguistic rules:

(1) is_qualia(A,B) :- precedes(B,A), near_verb(A,B), infinitive(B), action_verb(B).

(2) is_qualia(A,B) :- contiguous(A,B).

(3) is_qualia(A,B) :- precedes(B,A), near_word(A,B), near_verb(A,B), suc(B,C), preposition(C).

(4) is_qualia(A,B) :- near_word(A,B), pred(A,C), void(C).

(5) is_qualia(A,B) :- precedes(B,A), suc(B,C), colon(C), pred(A,D), punctuation(D), singular_common_noun(A).

(6) is_qualia(A,B) :- near_word(A,B), suc(B,C), suc(C,D), action_verb(D).

(7) is_qualia(A,B) :- precedes(A,B), near_word(A,B), pred(A,C), punctuation(C).

(8) is_qualia(A,B) :- near_verb(A,B), pred(B,C), pred(C,D), pred(D,E), preposition(E), pred(A,F), void(F).

(9) is_qualia(A,B) :- precedes(A,B), near_verb(A,B), pred(A,C), subordinating_conjunction(C).

Predicates must be read as follows : precedes(X,Y) means that X occurs somewhere in a sentence before Y. pred(X,Y) means that Y occurs immediately before X and conversely suc(Y,X) means that X occurs immediately after Y. near_word(X,Y) means that X and Y are separated by at least one word and at most 2 words, and near_verb(X,Y) that there is no verb between X and Y.

What is first striking is the fact that, at this level of generalization, few linguistic features are retained. The clauses seem to provide very general indications and tell us very little about types of verbs (action verb is the only information we get), nouns (common noun) or prepositions that are likely to fit into such structures. But the clauses contain other information, related to several aspects of linguistic descriptions, like:

- proximity: this is a major criterion. Most clauses indicate that the noun and the verb must be either contiguous (clause 2) or separated by at most one element (clauses 3, 4, 6, 7) and that no verb must appear between N and V (clauses 1, 3, 8, 9).

- position: clauses 4 and 7 indicate that the one of the two elements is found at the beginning of a sentence or right after a punctuation mark, whereas the relative position of N and V (precede/2) is given in clauses 1, 3, 5, 7, 9.

- punctuation: punctuation marks, and more specifically colons, are mentioned in clauses 5 and 7.

- morpho-syntactic categorization: the first clause detects a very important structure in the text, corresponding to action verbs in the infinitive form.

These features bring to light linguistic patterns that are very specific to the corpus—a text falling within the instructional genre. We find in this text many examples in which a verb at the infinitive form occurs at the beginning of a proposition and is followed by a noun phrase. Such lists of instructions are very typical of the corpus:

– *d´ebrancher la prise* (disconnect the plug)

– *enclencher le disjoncteur* (engage the circuit breaker)

– *d´eposer les obturateurs* (remove the obturators)

To further evaluate these findings, we have compared what is obtained by the automatic learning process to linguistic observations made manually on the same corpus (Galy, 2000). Galy has listed a set of canonical verbal structures that convey telic information:

– infinitive verb + det + noun (*visser le bouchon*) (to tighten the cap)

– verb + det + noun (*ferment le circuit*) (close the circuit)

– noun + past_participle (*bouchon maintenu*) (held cap)

– noun + be + past_participle (*circuits sont raccordés*) (circuits are connected)

– noun + verb (*un bouchon obture*) (a cap blocks up)

– be + past_participle + par + det + noun (*sont obturées par les bouchons*) (are blocked up by caps)

The two types of results show some overlap: both experiments demonstrate the significance of infinitive structures and bring to light patterns in which verb and noun are very close to each other. Yet the results are quite different since the learning method proposes a generalization of the structures discovered by Galy. In particular, the opposition between passive and active constructions is merged in clause 2 by the indication of mere contiguity (V can occur before or after N). Conversely, some clues have not been observed by manual analysis because they are related to levels of linguistic information that are usually neglected by linguistic observation (punctuation marks and position in the sentence).

Consequently, when we look at the results of the learning process from a linguistic point of view, it appears that the clauses give very general surface clues about the structures that are favored in the corpus for the expression of qualia relations. Yet, these clues are sufficient to give access to some corpus-specific patterns, which is a very interesting result.

5. Conclusion and future work

We have presented a system extracting from a POS and semantically tagged corpus N-V pairs such that N and V are linked by a qualia relation. This system is based upon contextual rules that are automatically learnt by ILP from examples provided by an expert. These rules use the POS and semantic tags of the N-V pair context to characterize what distinguish qualia N-V pairs from non-qualia ones. This semi-automatic system is compared with two different

co-occurrence extraction approaches on a test set:

– statistical models, entirely automatic, do not perform well enough to be used without enhancements or *a posteriori* human supervision;

– a manual syntactic annotation of the N-V pairs, gives high quality results but is too costly to be used on a big amount of texts.

In this respect, our symbolic learning approach is a good compromise, combining good results and a modest human intervention. Moreover, the rules generated by ILP provide interesting linguistic patterns to describe the qualia relation from a theoretical point of view.

As regards the symbolic learning approach, one next step of this work is to repeat the experiment on new corpora and other languages in order to help to identify specific structures carrying qualia relations. Another future work is to apply similar methods to extract other kinds of co-occurrences and more generally to any information extraction task.

Concerning the qualia N-V pairs, future studies will be undertaken to use them to reformulate or extend indexes in a real information retrieval system such as a textual search engine.

6. References

Susan Armstrong, Pierrette Bouillon, and Gilbert Robert. 1995. Tagger Overview. Technical report, ISSCO.

Susan Armstrong. 1996. Multext: Multilingual Text Tools and Corpora. In H. Feldweg and W. Hinrichs, editors, *Lexikon und Text*. Tübingen: Niemeyer.

Pierrette Bouillon, Sabine Lehmann, Sandra Manzi, and Dominique Petitpierre. 1998. Développement de lexiques à grande échelle. In *Proceedings of colloque de Tunis 1997 "La mémoire des mots"*, Tunis, Tunisie.

Pierrette Bouillon, Robert H. Baud, Gilbert Robert, and Patrick Ruch. 2000. Indexing by Statistical Tagging. In *Proceedings of Journées d'Analyse statistique des Données Textuelles, JADT2000*, Lausanne, Switzerland.

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 *Proceedings of First International Workshop on Generative Approaches to the Lexicon, GL'2001*, Genève, Suisse.

Kenneth W. Church and William A. Gale. 1991. Concordances for parallel text. In *Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research*.

Béatrice Daille. 1994. *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*. Ph.D. thesis, Université Paris 7.

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

Cécile Fabre and Pascale Sébillot. 1999. Semantic Interpretation of Binominal Sequences and Information Retrieval. In *Proceedings of International ICSC Congress on Computational Intelligence: Methods and Applications,CIMA'99, Symposium on Advances in Intelligent Data Analysis AIDA'99*, Rochester, N.Y., USA.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Edith Galy. 2000. Repérer en corpus les associations sémantiques privilégiées entre le nom et le verbe : le cas de la fonction dénotée par le nom. Master's thesis, Université de Toulouse - Le Mirail.

Gregory Grefenstette. 1997. SQLET: Short Query Linguistic Expansion Techniques, Palliating One-Word Queries by Providing Intermediate Structure to Text. In McGill-University, editor, *Proceedings of Recherche d'Informations Assistée par Ordinateur, RIAO'97*, Montréal, Québec, Canada.

Benoît Habert, Adeline Nazarenko, and André Salem. 1997. *Les linguistiques de corpus*. Armand Collin/Masson, Paris.

Ron Kohavi. 1995. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence, IJCAI 95*, Montréal, Québec, Canada.

Emmanuel Morin. 1999. *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques*. Ph.D. thesis, Université de Nantes.

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

Dominique Petitpierre and Graham Russell. 1998. M-morph - the Multext Morphology Program. Technical report, ISSCO.

Ronan Pichon and Pascale Sébillot. 1997. Acquisition automatique d'informations lexicales à partir de corpus : un bilan. Research report 3321, INRIA, Rennes.

James Pustejovsky, Peter Anick, and Sabine Bergler. 1993. Lexical Semantic Techniques for Corpus Analysis. *Computational Linguistics*, 19(2):331–358.

James Pustejovsky, Branimir Boguraev, Marc Verhagen, Paul Buitelaar, and Michael Johnston. 1997. Semantic indexing and typed hyperlinking. In *Proceedings of the American Association for Artifical Intelligence Conference, Spring Symposium, NLP for WWW, 120-128. Stanford University, CA.*

James Pustejovsky. 1995. *The Generative Lexicon*. Cambridge:MIT Press.

Laurence Vandenbroucke. 2000. Indexation automatique par couples nom-verbes pertinents, Mémoire de DES en information et documentation. Master's thesis, Université Libre de Bruxelles.

Stefan Wermter, Ellen Riloff, and Gabriele Scheler, editors. 1996. *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*. Lecture Notes in Computer Science, Vol. 1040, Springer Verlag.

Yorick Wilks and Mark Stevenson. 1996. The Grammar of Sense: is Word-Sense Tagging much more than Part-of-Speech Tagging? Technical report, University of Sheffield, UK.