Authors' addresses

Jakob Halskov
Research assistant
Danish Language Council, Copenhagen
Njalsgade 136
Bygn. 27,3.sal
2300 København S
Denmark

jhalskov@dsn.dk

Caroline Barrière
Research officer
Interactive Language Technology Group
National Research Council of Canada
283, Boulevard Alexandre, Taché
Édifice CRTL, pièce FI-040
Gatineau, Canada J8X 3X7

Caroline.Barriere@cnrc-nrc.gc.ca

About the authors

Jakob Halskov has a doctorate in Computational Linguistics from Copenhagen Business School and a master's degree in English from Copenhagen University. He has worked as research assistant at the Danish Language Council since 2007. His main research interests are corpus linguistics and computational terminology. He is particularly interested in the task of automatic knowledge extraction from natural language text, but also in the semantic fuzziness which arises when non-experts make use of terminology from specialized fields in their everyday communication.

Caroline Barrière has a doctorate in Computational Linguistics from Simon Fraser University (Vancouver, Canada), as well as a master's degree in Electrical Engineering and a bachelor's degree in Computer Engineering from École Polytechnique de Montréal. She worked as Assistant Professor at University of Ottawa's School of Information and Technology Engineering (SITE) from 1997 to 2003, and then became a research officer at the Interactive Language Technology Group of the National Research Council of Canada. Her work focuses on computational terminology and lexicography. She is particularly interested in the automatic extraction of knowledge from dictionaries and corpora, as well as the conceptual representation of this knowledge. She further aims at applying her research in the development of tools for terminologists, translators and language teachers.

Designing and evaluating patterns for relation acquisition from texts with Caméléon

Nathalie Aussenac-Gilles and Marie-Paule Jacques

Pattern-based approaches for knowledge identification in texts assume that linguistic regularities always characterise the same kind of knowledge, such as semantic relations. In this paper, we report the experimental evaluation of a large set of patterns using an ontology enrichment tool: Caméléon. Results emphasize the strong influence of the corpus on pattern efficiency and on their meaning. This influence confirms two of the hypotheses that motivated to define Caméléon as a support used in a human-driven process: (1) patterns and relations must be adapted to each project; (2) human interpretation is required to decide how to report the pieces of knowledge identified with patterns in the ontology.

Keywords: knowledge patterns for French, ontology engineering from text, knowledge engineering, conceptual relation, semi-automatic information extraction, terminology structuring, dependence on textual genre and domain, tool evaluation

1. Introduction

Relation extraction is one of the major issues in knowledge acquisition from texts. Relations can be an efficient means to rapidly structure a conceptual model. Moreover, relations may help to identify significant domain concepts. Various complementary approaches may be used to identify relations: searching for co-occurring terms and identifying the possible semantics of their relation thanks to existing relations in identified lexical and terminological resources like WordNet (Staab and Maedche 2001; Cimiano et al. 2005; Velardi et al. 2006); matching lexico-syntactic patterns in domain corpora (Girju and Moldovan 2002; Hearst 1992; Séguéla 2001) or in very large and general corpora (like Le Monde for French[1] or BNC for English);[2] learning dependencies between phrases through the distribution analysis of terms in domain corpora (Bourigault 2002); statistical learning of term clusters and their relations (Cimiano 2007).

Terminology 14:1 (2008), 45–73. DOI 10.1075/term.14.1.04aus
ISSN 0929–9971 / E-ISSN 1569–9994 © John Benjamins Publishing Company

Pattern-based approaches for knowledge identification in texts assume that linguistic regularities always characterise the same kind of knowledge, such as semantic relations. We present here the new version of a tool, Caméléon, that implements pattern matching in corpora to identify relations and concepts for ontology engineering. (Séguéla 2001) In this paper, we focus on the process of building and evaluating a set of patterns that must fill in a database of patterns which is provided with the tool.

Caméléon is based on two hypotheses: (1) patterns and relations may vary with the domain and corpus under study so they must be defined for and/or adapted to each project; (2) human interpretation is required to decide how to report the pieces of knowledge identified with patterns in an ontology. For these reasons, Caméléon is a support to be used in a human-driven process. Although the process is not completely automated, Caméléon contributes to ontology learning from texts.

Following the first hypothesis, we made the assumption that domain-specific patterns would be easier to define if some "generic" patterns were already available. Such a set of "generic" patterns had been made available in a previous version of Caméléon (Séguéla 2001) which processed untagged texts. The main idea was to provide users with an "almost ready-to-use" set of patterns they could adapt to a particular domain or even use "as is".

The new version of Caméléon presented here processes tagged texts. (Aussenac-Gilles and Jacques 2006) The previous set of "generic" patterns had to be modified. This paper reports on how we built and evaluated a new set of 70 patterns for the French language. The results prove that rather than aiming at being generic, patterns should be adaptable and reusable. The experiment reported here led to a shift from the notion of 'generic' to the notion of 'reusable'. To be 'reusable' means that each pattern is a product of previous work, it has been used for a given project for which it has been judged valuable, but its use within the framework of a new project may be conditional to modification, and even major modifications, depending of the texts that make up the corpus.

The aim of the provided database then becomes twofold: to store valuable patterns which can be considered a bootstrap for any project, to give examples of knowledge patterns that can help to understand how the tool supports relation extraction.

Beyond pattern definitions, this experiment contributes to validate our choice of proposing a human-driven process in Caméléon. Human interpretation is indispensable both to adapt patterns to domain and corpus characteristics and to evaluate results of pattern-matching before enriching a conceptual model that could become an ontology.

However, we must emphasize the fact that pattern-matching does not produce a huge amount of contexts to analyze. Pattern-matching focuses on quality and on fine-grained analyses, unlike co-occurrence context clustering or other statistical approaches that, instead, produce more quantity and coarse-grained analyses. It may yield less quantity but more confident results. As a matter of fact, it remains really difficult to evaluate the overall relation extraction process and to compare it with completely different methods because different methods produce heterogeneous results.

The rest of the paper is structured as follows. After sketching the state of the art in Section 2, Section 3 offers an overview of Caméléon as a method for ontology enrichment and presents the two steps of that process: defining domain specific patterns and then enriching the ontology. Section 4 describes the tool itself, and focuses particularly on the way patterns may be defined and evaluated during the first step. In Section 5, we present the method used for pattern definition and evaluation. Results about pattern evaluation are detailed in Section 6, and discussed in Section 7. We conclude by emphasizing the dependence to the corpus and we propose new functionalities in Caméléon to account for this dependence.

2. The state of the art: Pattern-based knowledge identification

2.1 Pattern-based approaches: A cross-disciplinary matter

Patterns are lexical, semantic and/or syntactic characterizations of linguistic contexts in which one expects to find some specific piece of information. The literature about patterns gathers contributions from linguists, researchers in natural language processing and, recently, in ontology engineering. Each of research work refers to investigations with different *foci*.

Linguistic grounding of patterns: Linguists consider patterns as a means to explore language regularity in corpora. They evaluate the ability of patterns to reveal grammatical, syntactical or even semantic dependencies between words or phrases. Their goal is to list and identify patterns, clarify their meaning or possible interpretation, characterize their occurring contexts and their modalities of use. For a typology of such patterns, read Marshman and L'Homme (2006). Syntactic patterns have been defined to characterize noun phrases or noun terms for instance. Semantic patterns revealing causal relations in French have been identified in Garcia (1998), Barrière (2001) and Marshman and L'Homme (2006). Rebeyrolle (2000) studied definition patterns in French. Condamines (2002) has emphasized the variability of some unusual patterns for part-of relations like *chez* in very specific types of scientific corpora. After Hearst's early paper (1992), many linguists like Feliu and Cabré Castellvi (2002) have proposed lists of patterns for hypernymy relation in various languages. These patterns are often defined or checked by manual text browsing, although many linguists tend to use simple and efficient tools like concordancers (SATO (Daoust 1996), *Concord* by *WordSmith*), KeyWordsInContext like SystemQuick (Ahmad and Holmes-Higgin 1995), or basic text browsing functions in text editors.

Work on pattern implementation: Researchers in corpus linguistics collaborate with natural language processing specialists in order to optimize pattern implementation. Issues raised by this research include: Should patterns apply to tagged corpora or untagged one? How to get the most efficient operational mark up that corresponds to a pattern? How can a pattern matching algorithm be optimized for a given corpus? As shown by Rebeyrolle and Tanguy (2000), tuning pattern definitions in a particular tool has a significant impact on their recall and precision. In other words, one cannot say that a pattern is relevant regardless of its encoding. Patterns may be implemented with the help of finite automata (Velardi et al. 2006), with regular forms used by concordancers (Feliu and Cabré Castellvi 2002; Marshman and L'Homme 2006) or with rules like in Gate (Bontcheva et al. 2004). In those cases, each item in the pattern has an equal weight. Alternative approaches, like contextual exploration defined by Desclés (Desclés 1997; Garcia 1998; Jackiewicz 1996) distinguish more or less important items in a pattern. For instance, in a definition pattern, the verb to define may be considered as the focus, and the fact that it is followed by a determiner and a noun could be the context. Then, the search relies firstly on the most significant part of the pattern, and then the remaining of the pattern is searched in the sentence including the focus.

Patterns in ontology engineering from text: When building ontologies from text, patterns are considered as one of the possible tools to identify either terms (that will contribute to define concepts) or lexical relations (that could reveal semantic relations and concepts). (Byrd and Ravin 1999) In early work like Prométhée (Morin 1999) and Caméléon (Séguéla 2001), the problem was to define high quality and accurate patterns that would lead to relevant domain relations, with high recall and precision. Because defining good patterns is time consuming and little productive, the objective has shifted to get a more automatic process, and to gain efficiency in the overall approach. Several methods take advantage of the combination of pattern based approaches with other techniques (statistical text mining, reuse of resources, etc.). Among these methods are RelExt by Schutz and Buitelaar (2005), OntoLearn by Velardi et al. (2006), the process defined by Sabou (2004) or the method proposed by Gillam et al. (2005) to name but a few.

2.2 Pattern-based identification of semantic relations

Hearst (1992) was the first to experiment a pattern-based approach for lexical relation identification. The underlying hypothesis assumes that terms that share similar linguistic contexts may belong to the same semantic class or that similar semantic relations may connect them. Various tools implement Hearst's patterns. (Reinberger and Spyns 2004; Cimiano et al. 2005; Velardi et al. 2007) Hearst tested some general patterns mainly expressing definitions or hypernymy. She noticed that linguistic patterns had to be tuned for each corpus and domain. Over the last ten years, patterns were widely used with success for information extraction or relation extraction. (Reinberger and Spyns 2004)

To gain efficiency, research has investigated two main tracks. Firstly, to reduce the cost of pattern definition and tuning, patterns may be learned from manually-tagged corpora (Cimiano et al. 2005; Faure and Poibeau 2000; Staab and Maedche 2001); they may refer to named entities and known semantic classes (Girju and Moldovan 2002); they may be learned by mining the contexts of co-occurring terms (Cimiano 2007). Secondly, to reduce the time required to select valid pattern instances and the noise of the overall process, various statistical text analyses have been tested to sort the matched sentences according to their possible validity. Additional processing can help determine the exact label for automatically learned relations. (Kavalec and Staték 2005) Like Girju and Moldovan (2002), we consider that an alternative contribution would be to store robust patterns and know-how about their use, together with information about their semantics, their precision and recall in various types of domains and documents. A third issue is to improve the identification of the right concepts and semantic relations from linguistic indices. Reusing available taxonomies, thesaurus or ontologies can help to focus on domain relevant entities.

3. A pattern-based method for a semantic relation identification

3.1 Background and motivations

Caméléon is a method and tool to extend an existing network of concepts with new terms, concepts and semantic relations by applying a pattern-based approach. (Séguéla 2001) A conceptual model built up with Caméléon is a semantic network where concepts are associated with a set of terms (synonym terms that label this concept). This model may be the starting point for the design of an ontology or it may be considered as a result by itself. This tool can be one of the components of a natural language processing (NLP) and modelling chain from texts to ontologies or structured terminologies, such as the ones proposed in Kaon (Staab and Maedche 2001), in Terminae (Aussenac-Gilles et al. 2000) or in (Gillam et al. 2005). The Caméléon tool provides means to manually define and evaluate patterns for various types of semantic relations on corpora, and to define new concepts and relations in a conceptual model. So, in addition to ontology and terminology engineering, it may be also relevant for corpus linguistic studies.

The theoretical background of Caméléon comes from the ideas promoted in France by the French TIA[3] special interest group about knowledge modelling from text. This group emphasized the necessity of a deep understanding of the fuzzy, flexible and complex nature of the term-concept connection in order to define relevant tools and processes for identifying knowledge from text. It promoted a textual semantics, where a meaning can be assigned to terms from observing their use, and where the concepts and terms of a domain are not definitive but continuously evolving. From this point of view, modelling knowledge from text may gain from NLP tools for text processing, but it requires the selection of relevant texts and a human supervision to build application and domain specific models. Concerning pattern-based text analysis, this trend assumes the following:

- Patterns and semantic relations are domain and corpus dependant: each new domain will lead to a new set of relations and for each of them, to new associated patterns;
- Matching patterns on a corpus provides sentences where several or none possible concepts or conceptual relations could be identified: human interpretation is required to identify the right terms, concepts and relations to be added to the model if any;
- Conceptual models in general, and even consensual domain ontologies, will not be used in a software system unless they are relevant for this system. The target system influences both their content and their structure. Human supervision during conceptual modelling is a means to take into account these relevance criteria.

3.2 Overview of the approach

For this reason, Caméléon suggests two steps:

1. Defining project-specific patterns relevant for the corpus to be analysed and for the objective of the model;
2. Matching these patterns to the corpus and extending the conceptual model with new terms, concepts and relations.

Hence the tool contains two modules: one supports pattern definition, matching and testing, the other one helps to interpret the sentences that match the patterns and to improve the conceptual model. The current version of the tool processes tagged texts. Any tagger can be used, for the tagset has to be given in a parameter file. So patterns may be formed with words or lemmas, wildcards, Part-of-Speech word characterizations or semantic classes defined by the user.

We now describe the two-step approach in Caméléon, and thereafter we present the issue of evaluating the method and tool.

3.3 Step 1: Project pattern definition

For a given project and corpus of texts, the user is expected to define a specific set of domain relations together with valid patterns that would identify them. Corpus specific patterns may be obtained by one of the following means:

1. Adapting some of the patterns already available in Caméléon;
2. Manually defining new patterns for already identified domain relations;
3. Defining new relations and patterns after observing the contexts in which related terms are used.

Each of them requires defining or fixing patterns, searching them in the corpus and finally evaluating their efficiency. The evaluation of the patterns relies on the validation or rejection of the corpus sentences identified. A valid pattern is the result of an incremental process: by checking the contexts it matches, the original pattern is modified as needed in order to reduce noise and enhance recall. To carry out (1), Caméléon proposes a set of patterns stored in a "generic" database. It is the design of these knowledge patterns which is described in the following sections. These patterns have as a starting point patterns that have already been identified by linguists as indices of well known semantic relations in 'general language'.[4]

To carry out (3), new patterns and lexical relation types can be identified following Riloff's suggestion (1996): pairs of related terms are searched for in the corpus; their shared contexts are browsed to identify semantic relations from linguistic regularities. Patterns are then abstractions of these regularities.

Once the final set of patterns for a given project has been fixed, the second step is conceptual model enrichment.

3.4 Step 2: Enriching the conceptual model

As explained above, a new piece of information (such as term, concept or relation) is added to the conceptual model inasmuch as it is relevant with regard to the final application. A term, concept or relation bears no relevance by itself but only when taking into consideration the final use of the model. So the relevance of a term, concept or relation is narrowly linked to the way the conceptual model is thought of. As a methodological consequence of this statement, the user must decide whether each of the sentences that match each of the patterns provides something that is worth being added to the conceptual model. These decisions must be taken having in mind both the intended model and the current state of the model.

We can then say that extending the model results from a compromise between its current content, the information available in the source text, the role of the model in the target application and ontological structuring guidelines like concept differentiation. At this time, no heuristic has been made explicit to report how such a compromise is reached; therefore, human management is still required. Human intervention is needed to analyze one by one the sentences in the corpus which match project patterns and to assess whether the matched sentences suggest new concepts and relations.

Suggestions of relations are then presented in the Caméléon ontology browser, when the user edits one of the related concepts. The user must decide whether to define a new relation or not, and whether the related concepts are those suggested or other ones.

This process is quite complex and time-consuming. It requires know-how in knowledge modelling and a correct appreciation of the intended role of the ontology. To save time, matched sentences are presented in a list that can be either checked one by one, or overviewed in a glance by a human who determines the validity of each pattern found.

3.5 Towards an evaluation of Caméléon

An approach like the one implemented in Caméléon, where human interpretation plays an important part, is difficult to evaluate. A full evaluation should include the design of a real ontology for a well-determined system. For instance, we could measure if ontology relevance increases after enriching it with new concepts and relations identified in texts with the help of Caméléon.

Since the tool integrates two modules and a set of supposedly generic patterns, we have decided to carry out first an evaluation of the pattern definition module. This evaluation requires an evaluation of the tool functionalities when adapting pre-defined patterns, defining new patterns and evaluating them. It also requires the evaluation of the quality of the base of "generic" patterns.

This article reports on the work of defining the bootstrap set of knowledge patterns. In doing so, we have used and tested the functions devoted to pattern definition and evaluation. This experiment contributed to the validation of the two foundational hypotheses (the need for pattern adaptation and human interpretation). We first present how the tool supports pattern definition and evaluation before providing some details about the corpora we used and the method we followed for this experiment.

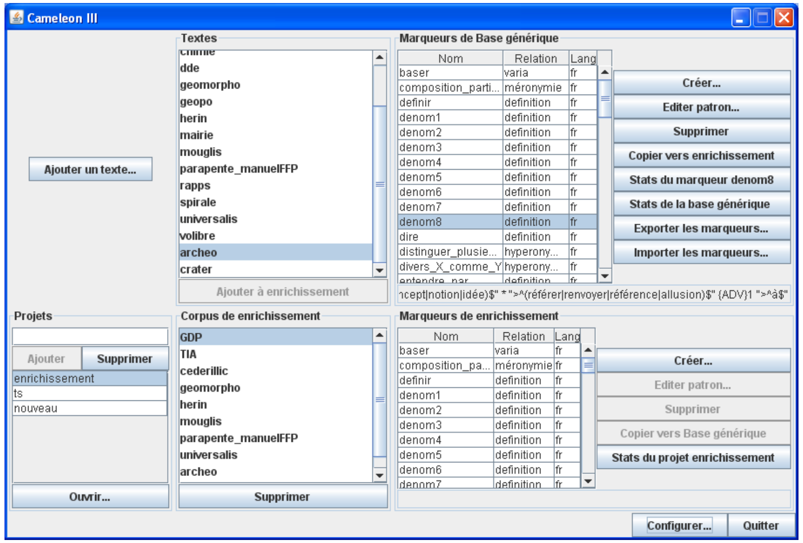
4. How the tool supports relation extraction

Since the main idea behind Caméléon is to assist terminologists or knowledge engineers when building terminologies and ontologies from texts, Caméléon is designed to process texts and to automatically retrieve knowledge patterns from texts. So the current version of the tool includes the use of a concordancer, the Keskya concordancer, which matches the patterns to the texts tagged with the help of a POS tagger like TreeTagger. Since the design and adaptation of knowledge patterns is very time-consuming, the tool aims to propose as a bootstrap a set of patterns that can be used and adapted for any new project. These patterns and the corresponding relations form two knowledge bases that must be adapted and enriched for any new project. The pattern database obtained at the end of the current experiment includes about 70 patterns for the French language, and we plan to build up another base for English.

The documents to be analysed for knowledge extraction are supposed to be tagged and then stored in the Caméléon database. We present here some of the screens that illustrate the functionalities designed to guide the define process of a new knowledge model from texts.

4.1 Setting up a new project

A project in Caméléon entails a set of texts — the project corpus, a set of patterns to be designed or adapted from available ones, and a conceptual model built up after analysing sentences identified by matching patterns in the corpus. Project parameters are set up on the tool main screen which is two-fold (Figure 1): the upper part lists available texts (*textes*), patterns (*marqueurs*) and corresponding relations (*relation*) in the tool data-base; the lower part is used to set up or to select the undergoing project (*projets*). When defining a new project (here, *archeo*), a set of texts is selected among available ones; it forms the *corpus* of a project. Then a set of patterns can be selected from the list of available patterns (*copier vers archeo* button) and evaluated for this corpus after selecting *ouvrir* (open). New patterns may be created (*créer*) opening the Pattern Editor (Figure 2).



Marqueurs de Base générique

Nom	Relation	Lang
baser	varia	fr
composition_parti...	méronymie	fr
definir	definition	fr
denom1	definition	fr
denom2	definition	fr
denom3	definition	fr
denom4	definition	fr
denom5	definition	fr
denom6	definition	fr
denom7	definition	fr
denom8	definition	fr
dire	definition	fr
distinguer_plusie...	hyperony...	fr
divers_X_comme_Y	hyperony...	fr
entendre_par	definition	fr

```
ncept|notion|idée)$" * ">^(référer|renvoyer|référence|allusion)$" {ADV}1 ">^à$"
```

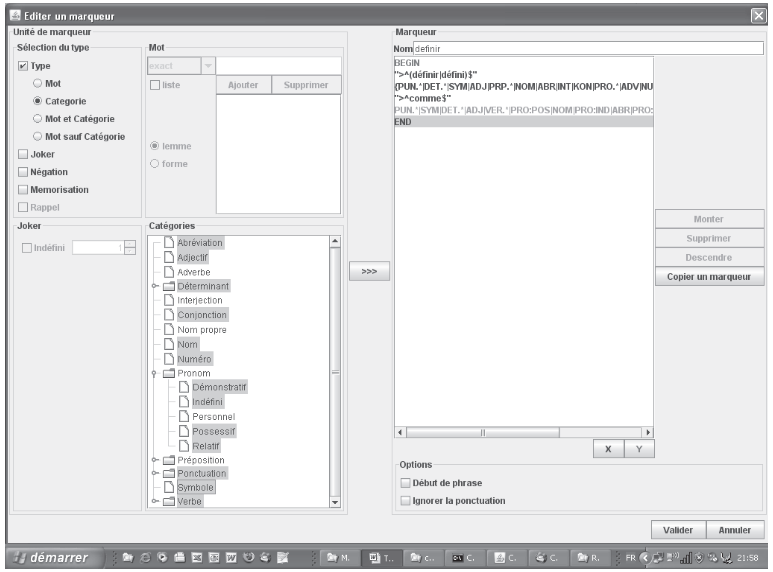
Marqueurs de enrichissement

Nom	Relation	Lang
baser	varia	fr
composition_pa...	méronymie	fr
definir	definition	fr
denom1	definition	fr
denom2	definition	fr
denom3	definition	fr
denom4	definition	fr
denom5	definition	fr
denom6	definition	fr
denom7	definition	fr

Figure 1. Caméléon main screen. On this screen dump, the Enrichissement project relies on eight texts. Several patterns have already been pasted from the "generic" base (from the upper pattern list). The *denom8* pattern that identifies *definition* relations is selected and will be evaluated for all the corpora of the Enrichissement project.

4.2 Pattern design and adaptation

The internal representation of patterns is the one required by the Keskya concordancer. Patterns are meant to be included in a single sentence. They are expressed mainly with lemmas combined with Part-of-Speech (POS) tags, and a set of operators like 'or' (represented by '|'), negation or iterations (*joker*). We call this list of items the pattern definition. Because writing a pattern with this representation would require specific skills, the interface proposes a pattern editor (Figure 2) that makes it easier to define (or modify) each pattern chunk after chunk. The user selects one of the options on the left part of the window and adds new components to the pattern presented on the right part of the window. Because patterns characterise linguistic contexts where semantic relations between concepts may appear in texts, the knowledge engineer must specify which parts of the pattern will refer to the related concepts (*X* and *Y*). Each of these chunks is turned into a particular colour that will be used later on to colour the words that may correspond to the related concepts, when parsing the sentences that match the pattern (Figure 3).



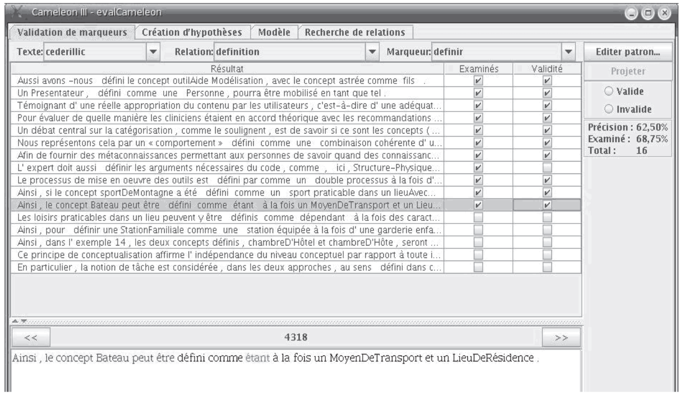
Nom definir

```
BEGIN
">^(définir|défini)$"
{PUN.*|DET.*|SYM|ADJ|PRP.*|NOM|ABR|INT|KON|PRO.*|ADV|NU
">^comme$"
PUN.*|SYM|DET.*|ADJ|VER.*|PRO:POS|NOM|PRO:IND|ABR|PRO:
END
```

Figure 2. Caméléon pattern editor. The pattern presented here, called *definir*, mainly searches for forms like "X is defined as Y". The user preferred not to specify where the defined concept could appear in a sentence (*BEGIN* is in the *X* colour), but gave several constraints (list above END) on how Y (the defining term) could be formulated.

4.3 Pattern evaluation

Patterns are then evaluated one by one (Figure 3). Evaluating a pattern means checking some of the sentences where the pattern appears in each of the corpus texts. The goal is to decide whether the pattern is to be rejected, modified or kept "as is" as a relevant pattern for this project. To influence this decision, a precision score is displayed after checking a set of sentences extracted from the corpus.



Texte: cederillic

Relation: definition

Marqueur: definir

Résultat	Examinés	Validité
Aussi avons -nous défini le concept outilAide Modélisation , avec le concept astrée comme fils .	☑	☑
Un Presentateur , défini comme une Personne , pourra être mobilisé en tant que tel .	☑	☑
Témoignant d' une réelle appropriation du contenu par les utilisateurs , c'est-à-dire d' une adéquat...	☑	☑
Pour évaluer de quelle manière les cliniciens étaient en accord théorique avec les recommandations ...	☑	☑
Un débat central sur la catégorisation , comme le soulignent , est de savoir si ce sont les concepts (...	☑	☑
Nous représentons cela par un « comportement » défini comme une combinaison cohérente d' u...	☑	☑
Afin de fournir des métaconnaissances permettant aux personnes de savoir quand des connaissanc...	☑	☑
L' expert doit aussi définir les arguments nécessaires du code , comme , ici , Structure-Physique...	☑	□
Le processus de mise en oeuvre des outils est défini par comme un double processus à la fois d'...	☑	☑
Ainsi , si le concept sportDeMontagne a été défini comme un sport praticable dans un lieuAvec...	☑	☑
Ainsi , le concept Bateau peut être défini comme étant à la fois un MoyenDeTransport et un Lieu...	☑	☑
Les loisirs praticables dans un lieu peuvent y être définis comme dépendant à la fois des caract...	□	□
Ainsi , pour définir une StationFamiliale comme une station équipée à la fois d' une garderie enfa...	□	□
Ainsi , dans l' exemple 14 , les deux concepts définis , chambreD'Hôtel et chambreD'Hôte , seront ...	□	□
Ce principe de conceptualisation affirme l' indépendance du niveau conceptuel par rapport à toute i...	□	□
En particulier , la notion de tâche est considérée , dans les deux approches , au sens défini dans c...	□	□

Précision : 62,50%
Examiné : 68,75%
Total : 16

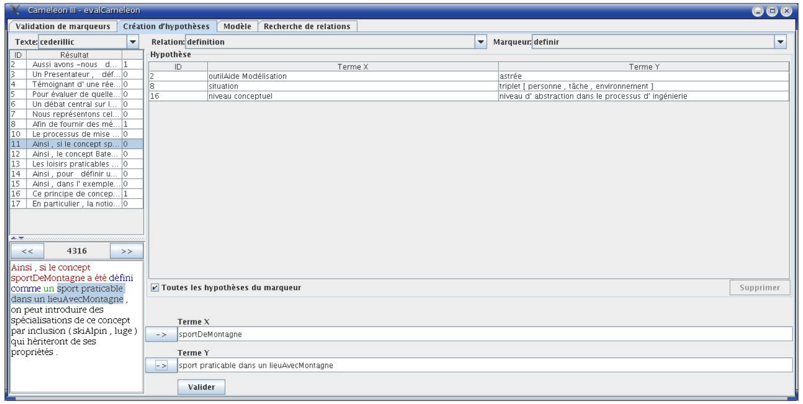
4318

Ainsi , le concept Bateau peut être défini comme étant à la fois un MoyenDeTransport et un LieuDeRésidence .

Figure 3. Caméléon pattern evaluation Screen. Given a text (*texte*), a relation type (*relation*) and a pattern, the user may ask to match the pattern to the text (*projeter*). Results are text excerpts (sentences) listed for checking. Selecting one of these sentences (the highlighted one) makes it possible to read the whole sentence that matches the pattern in the lower frame. Coloured words correspond to those identified as possible related concepts (X and Y). After validating or rejecting as many sentences as requested (*validate* check boxes), the user may decide to modify the pattern (*Editer patron*), reject or validate it (*invalide* or *valide* radio-button on the right). The precision score displayed on the right may guide this decision.

4.4 Text fragment selection

Once a set of patterns has been tuned to the project corpus, the user checks each of the sentences matched in the corpus. The user is supposed to precisely identify potential relations (relation hypotheses) in each of them (Figure 4). The user must decide whether a relation between concepts can be identified, and whether it is relevant or not to insert the relation and concepts in the conceptual model to be built. If the sentence (lower left box in Figure 4) is relevant for the target model, the user cuts and pastes the words that may correspond to related concepts (X and Y). Coloured words may guide him.



Texte: cederillic

Relation: definition

Marqueur: definir

ID	Résultat	
2	Aussi avons -nous d...	1
3	Un Presentateur , déf...	0
4	Témoignant d' une rée...	0
5	Pour évaluer de quelle...	0
6	Un débat central sur l...	0
7	Nous représentons cel...	0
8	Afin de fournir des mé...	1
10	Le processus de mise ...	0
11	Ainsi , si le concept sp...	0
12	Ainsi , le concept Bate...	0
13	Les loisirs praticables ...	0
14	Ainsi , pour définir u...	0
15	Ainsi , dans l' exemple...	0
16	Ce principe de concep...	1
17	En particulier , la notio...	0

4316

Ainsi , si le concept sportDeMontagne a été défini comme un sport praticable dans un lieuAvecMontagne , on peut introduire des spécialisations de ce concept par inclusion (skiAlpin , luge) qui hériteront de ses propriétés .

Hypothèse

ID	Terme X	Terme Y
2	outilAide Modélisation	astrée
8	situation	triplet [personne , tâche , environnement]
16	niveau conceptuel	niveau d' abstraction dans le processus d' ingénierie

☑ Toutes les hypothèses du marqueur

Terme X: sportDeMontagne

Terme Y: sport praticable dans un lieuAvecMontagne

Figure 4. Text fragment extraction. Checking the *definir* pattern (for *definition* relations).

4.5 Model enrichment

The conceptual model enriched from relation extraction and concept identification may be empty at first. New concepts may be identified from domain terms. All the terms identified as possible concept labels in relation hypotheses are available. When selecting one of these terms to define a concept, all the available relation hypotheses are presented. They may lead to define new concepts and conceptual relations. We will not go into details about this part of the tool because it is not used much in our evaluation. It would require another kind of evaluation including the design of an application specific ontology.

5. Method and Corpora for evaluating the set of knowledge patterns

In this section, we present the method for evaluating the patterns we designed for filling in the "generic" database of Caméléon. The criteria presented for this evaluation are always involved in the process of establishing the "final" version of the patterns (i.e., the one which will be provided with the tool).

We first explain how the patterns were designed, we then indicate the corpora against which the patterns have been evaluated and finally we present the criteria we used for evaluation. Results are presented in the following section.

5.1 Pattern design by using previous work

Since a lot of work has been done concerning knowledge patterns, the idea was not to begin from scratch but to design the patterns that would be stored in the "generic" database by using previous work as far as possible. This implied collecting already available patterns and adapting them to the specific constraints of Caméléon, if necessary.

The available patterns were first the patterns implemented in the previous version of Caméléon. (Séguéla 2001) We also had at our disposal a set of patterns that formed the basis of a previous experiment on semi-automatic retrieval of definitions (Rebeyrolle and Tanguy 2000), also processing tagged texts. Rebeyrolle and Tanguy provided us with the lexico-syntactic patterns intended for the retrieval of definitions together with their corpora and a set of reference sentences (i.e., the sentences which contain definitions) taken from these corpora.

Although it may seem irrelevant to search for definitions when aiming at building an ontology, definitions are useful because they may link a hypernym and a hyponym (Marshman et al. 2002), for example:[5]

> Les objets appartenant à une classe **sont appelés** instances… (Engl. The objects belonging to a class **are called** instances …)

This context corresponds to a pattern with *appeler* (*to call*) as a key term and it can be used to link *instances* and *objets*. In addition to this, definitions may serve to enrich the final ontology.

These two sets of patterns were available but had to be adapted in order to comply with Caméléon's representation of patterns.

At this point, we must recall that patterns in this new version of Caméléon are made of a combination of words, lemmas and POS tags (as can be seen in Figure 2 above). The patterns that came from the first version were designed to retrieve relations from untagged texts, so they listed lexical forms as pieces of the patterns. For instance, a pattern devoted to the relation of inclusion lists the different forms of the verbs bearing such a relation (the symbol | in the pattern means *or*):

> inclut|incluent|incluant|intègre|intègrent|intégrant

The challenge here was to design new Caméléon patterns so as to benefit from tagging, e.g. replacing lists of forms by lemmas combined with POS tags.

As for the definition patterns, they already used lemmas and POS tags, so we had to "translate" them from their original format into the one compliant with Caméléon. This means adapting the patterns from one tagset to another, for the tagger used by Rebeyrolle and Tanguy is Cordial Université,[6] which is based on a set of about 200 tags, while the one used by Caméléon is TreeTagger,[7] based on a set of 33 tags. This implies finding the corresponding tags for a given pattern. For instance, a pattern such as:

> <ce> <être>+(Vi|Vpp) (D|Mc)

which means 'lemma of *ce* followed by lemma of *être* in the indicative or present participle, followed by a determiner or a cardinal number', must be transformed, due to the fact there is no TreeTagger tag for "in the indicative", neither for "cardinal number" (*vs.* "ordinal number"), but differentiated tags for each verb tense (present, imperfect, future and so on) and a unique tag for numbers.

In the Discussion section, we will return to the consequences of adaptation and "translation" concerning the efficiency of the patterns — especially for definition patterns, in that they adopt a comparable conception.

5.2 Varied corpora

Since Caméléon is intended to retrieve semantic relations within specific domains, our corpora are all made up of specialized texts. They can be divided into two sets: (i) a set of 5 corpora together with 1617 reference sentences taken from them, provided as already mentioned by Rebeyrolle and Tanguy; (ii) another set of 3 corpora for which we had no reference sentences. The texts and the domains covered are the following.

First set:

1. A guide for planning electric networks (GDP, 187,800 words);
2. Scientific papers from the French conference Ingénierie des Connaissances (knowledge engineering), published in Charlet et al. (2000) (IC, 198,500 words);
3. A handbook of geomorphology (GEO, 260,000 words);
4. A handbook for software engineering specification, in the domain of electricity (MOU, 57,500 words);
5. Articles taken from Encyclopaedia Universalis, mainly regarding geomorphology (ENC, 200,500).

Second set:

6. A handbook on paragliding (PAR, 23,300 words);
7. PhD theses in archaeology (ARCH, 95,000 words);
8. Texts from the domain of telecommunications[8] (CRAT, 1,000,000 words).

It is worth specifying that one of the authors is a specialist in knowledge engineering, the other being a linguist and a specialist in paragliding. The other domains are not familiar to us and we will see later what difficulties arose from this.

The corpora do not correspond to any pre-established hypothesis. Apart from the CRAT corpus, for which we already knew that it was rich in "knowledge rich contexts" for it has been analyzed by other scholars (e.g., Pearson 1998), we did not know how productive the different corpora would prove. Since our perspective was to carry out the evaluation of a set of patterns, it did not matter if an ontology was effectively built or not. Indeed, if one wants to build an ontology from texts, one has to choose with great care the texts and keep in mind that the pattern-based approach may be one among others to retrieve relations within texts (Condamines and Jacques 2006).

We now turn to the framework within which the evaluation was carried out.

5.3 Evaluating the patterns

The patterns were evaluated using measures of precision and recall.

Since we did not have reference sentences for the whole set of texts, the final evaluation of the patterns is based for some of them (the definition patterns) on measures of precision and recall while it is only based on a measure of precision for the other ones. For this reason, the Results section is divided in two subsections: one devoted to the definition patterns, the other one being devoted to the taxonomic and other patterns.

For the sake of simplicity, the measures of precision and recall that we present in the Results section are those obtained at the end of the process of pattern design, i.e., measures of efficiency of the versions that have been fixed after doing the modifications that the matching sentences suggest. It would make sense to give the intermediate measures of recall and precision but only if our purpose was to explain how to adapt patterns for a specific corpus. This is for example the view adopted in (Rebeyrolle and Tanguy 2000), who explain step by step how to refine knowledge patterns to increase their performance.

6. Results

The next two subsections give both an overview of the patterns and the measures of their performance. Comments on these results will be developed in the Discussion section. (For further information about results and variability of performance, see Jacques and Aussenac (2006).)

6.1 Definition patterns

Concerning the retrieval of definitions, we adapted 19 patterns, some being relevant both for definitions and for hypernymy, most of them being specifically devoted to the expression of definitions. For example, the following pattern means 'lemma of verb *définir* (to define) followed by a wildcard followed by lemma of *comme* (as)'.

> <définir> 1 <comme>

This yields a context such as:

> Un Projet Logiciel peut **se définir comme** un Processus de Développement. (Engl. A software project may be defined as a development process.)

Table 1 presents the results of some of the patterns used, for the 5 corpora of the first set. N is the number of contexts yielded; R stands for Recall and P for Precision. R and P are expressed in percentages.

Table 1. Results of the evaluation of the definition patterns

	GDP			**IC**			**GEO**			**MOU**			**ENC**		
	N	**R**	**P**	**N**	**R**	**P**	**N**	**R**	**P**	**N**	**R**	**P**	**N**	**R**	**P**
définir	3	*100*	*100*	43	*88*	*98*	0			2	*100*	*100*	2	*100*	*100*
dénommer	7	*100*	*29*	10	*100*	*10*	57	*96*	*89*	0			23	*100*	*100*
entendre par	0			7	*100*	*71*	3	*100*	*33*	2	*100*	*100*	0		
signifier	0			13	*100*	*38*	29	*96*	*76*	0			7	*50*	*14*
être-un	258	*86*	*17*	489	*83*	*18*	641	*82*	*23*	120	*83*	*8*	375	*84*	*15*

The main comment on Table 1 is that patterns differ considerably from each other regarding Recall and Precision. Furthermore, the results of a given pattern may vary to a great extent with the corpus. To give but one example, the *être-un* (*is-a*) pattern, usually considered THE generic pattern, ranges from 120 to 641 contexts yielded and from 8% to 23% in terms of precision (and even to 40% for the PAR corpus of the second set).

We will return to this point at greater length in the Discussion section.

6.2 Taxonomic and other patterns

For the second step, we entered 52 more patterns: 35 for hypernymy, 14 for meronymy, 1 for reformulation, 2 'varia'. Table 2 below gives the results for a sample of patterns.

Table 2. Results of the second step of evaluation (N= Number of contexts; P= Precision)

	GDP		IC		GEO		MOU	
	N	P	N	P	N	P	N	P
et Adv	10	*10*	15	*7*	56	*30*	6	*17*
sorte de	0		7	*57*	3	*67*	0	
Inclure	75	*51*	32	*41*	16	*50*	18	*61*
partie de	0		0		7	*0*	0	
situé dans	40	*53*	63	*38*	38	*24*	4	*50*
c-à-dire	6	*67*	37	*54*	40	*80*	3	*100*
	ENC		PAR		ARCH		CRAT	
	N	P	N	P	N	P	N	P
et Adv	66	*5*	2	*0*	13	*38*	19	*58*
sorte de	1	*100*	0		0		4	*100*
Inclure	29	*62*	2	*100*	27	*19*	267	*48*
partie de	1	*100*	1	*0*	1	*0*	11	*18*
situé dans	55	*24*	4	*75*	36	*56*	291	*59*
c-à-dire	14	*29*	2	*100*	8	*63*	11	*64*

The first two belong to the "hypernymy" field. '*et Adv*' picks up contexts where the hyperonym is introduced with the adverbs *notamment, notablement, spécialement, particulièrement*, which mean *specially, particularly*; *'sorte de'* (*kind of*) retrieves contexts where the hypernym is part of a NP whose head is *sorte, type, genre, style, variété, espèce*:

> En ce qui concerne les grandes stations et **particulièrement** les stations Intelsat de type A… (Engl. For «large stations» and **particularly** the Intelsat Standard A stations…)

> les amines, qui sont des **sortes de** substances chimiques ; (Engl. amines, which are **kinds of** chemical substances;)

'*Inclure*' and *'partie-de'* capture converse relations within the field of meronymy, the former from the whole to the part, the latter from the part to the whole:

> Les services de base à fournir dans le Rmtp **comprennent** les téléservices et les services support… (Engl. The basic services to be provided in the Plmn **include** teleservices and bearer services […])

> l'ontologie **est un composant de** la mémoire d'entreprise… (Engl. an ontology **is a component** of corporate memory…)

The last two patterns are less conventional. '*Situé dans*' is a pattern related to what is often called a "specific" or "transversal" relation, insofar as such a relation does not take part in taxonomy. The pattern picks up contexts that express localisation:

> Le Ccm interroge l'Elv à chaque fois qu'il a besoin d'informations relatives à une station mobile donnée **située** à ce moment **dans** la zone du Ccm. (Engl. The Msc interrogates the Vlr whenever it needs information relating to a given mobile station currently **located in** the Msc area.)

'*C'est-à-dire*' does not really count as a semantic relation, but it yields contexts where a term is paraphrased. It may be of some interest in order to enrich the final ontology with definitions or explanations:

> la résolution, **c'est-à-dire** la taille des objets qui se distinguent, est de 100 m. (Engl. the resolution, **i.e.,** the size of objects that can be distinguished, is 100 m.)

Just as in Table 1, the different patterns produced heterogeneous results and the results of a given pattern depend on the corpus.

The measures shown by Table 1 and by Table 2 tend to challenge the notion of "generic pattern", insofar as the knowledge patterns have been identified by linguists and other scholars as indices of well known and widely used semantic relations in the general language. If one considers that "generic" means "equivalent performance", these results imply either revise the meaning of "generic" or consider that these so-called "generic patterns" are not so generic. This will be the second point of the discussion.

7. Discussion

The experiment we carried out gives rise to several issues: (1) issues related to pattern elaboration itself; (2) issues related to the results; (3) difficulties that are inherent to the task (our evaluation of the tool). The latter highlights the skills required to efficiently use the tool.

7.1 Pattern elaboration

Two aspects of pattern elaboration deserve to be taken into account: firstly, the construction of an abstract formulation that combines lexical units with POS tags; secondly, the adaptation of the resulting pattern to a specific corpus in order to enhance results.

Lexico-syntactic formulation of the patterns

In our experiment, the starting point was a set of existing patterns. We previously mentioned the problem we had to solve when "translating" the patterns from one tagset into another, less fine-grained tagset. Surprisingly, we did not notice an increase of noise due to tagset differences.

As for the patterns made up of lexical forms, the adaptation consisted in replacing a list of lexical forms with a lemma combined with tags. Due to the small number of available tags (33), we might expect a loss of information. For example, TreeTagger does not differentiate between the different persons of the verb. Here again, the number of contexts does not increase at the expense of precision, because the text itself filters out the undesirable contexts: nothing else than "3 Singular" is used in our corpora, for they contain mainly technical documents (cf. §5.2). If the texts were more general (e.g., news or novels), the patterns would probably have to be more constrained.

We can conclude that a tagset offers a convenient method of designing patterns in that it facilitates the expression of more abstract features while avoiding tedious entry of lists of forms. These remarks lead to the issue of the choice of the tagger for such a task. The accuracy of the tagset must represent a trade-off between need for precision and manageability: the more accurate the tagset, the more difficult the understanding of the tags — especially when the user is not attuned to dealing with morpho-syntactic categories — and the more difficult the handling of the tagset. In this sense, what could seem a loss when impoverishing the tagset is actually not.

Adaptation of the patterns to the different corpora

A given pattern is seldom convenient for every corpus, it is therefore necessary to modify it, generally to reduce irrelevant contexts. For instance, one of the hypernymy patterns is:

> NP1 <être> 1 ART_DEF NP2 ART_DEF (plus|moins)

That is 'a noun phrase — the hyponym — followed by verb *être* (*be*) followed by a definite article, followed by a noun phrase — the hypernym — followed by a definite article followed by *plus* or *moins*' (*the most*… or *the less*…). Here is an example of target context:

> La lave des coulées **est** la roche volcanique **la plus** résistante. (Engl. The lava of lava flow **is the most** resistant volcanic rock.)

Depending on corpora, a slight constraint may be put on the pattern. When the corpus is made up of scientific texts or consists of handbooks, the structure described is often used not to express hypernymy but to point to some striking example, or to mention some typical case of what is under discussion:

> La méthode KOD en **est** l'exemple **le plus** frappant: (Engl. KOD method **is the most** striking example of this)

In order to avoid this kind of context, we need to specify that NP2 must not have *exemple, cas* or *résultat* as its head. Notice that this is an *ad-hoc* constraint, since if one works with other types of texts, one may never face such contexts or may wish to retrieve them.

Generally, it must be kept in mind that the so-called generic patterns capture the most frequent or the most widespread constructions for a given relation. To a certain extent, it would be unrealistic to hope to take such a pattern and to use it without modification, because genuine texts reveal that even the most reliable construction can express meanings or relations of no interest with regard to the intended taxonomy or ontology. In this sense, no one can consider the elaboration of a pattern as definitive, since each new text may challenge its formulation.

7.2 What is a "generic pattern"?

The results presented in Section 6, together with the above observations about the "portability" of the patterns, challenge the notion of "generic pattern". If a generic pattern is the lexico-syntactic formulation of a semantic relation, which is said to invariably retrieve the same amount of relevant contexts, whatever the corpus, then we can conclude from our experiments that none of our patterns is generic.

Even the *is-a* pattern shows a huge difference between corpora, although it is acknowledged as being as generic as possible, in the sense it "occurs frequently and in many text genres". (Hearst 1992: 540) If one tests this pattern only on the PAR corpus, one will conclude this pattern is worth keeping since it has 40% precision; while if the same pattern is tested only on the MOU corpus, it is likely to be rejected, for its precision is 8%.

Furthermore, in addition to performance variability, one and the same pattern may correspond to slightly different meanings, depending on the corpus. For example, a meronymy pattern retrieves contexts expressing a "Component / Integral object" relation:

> Le Comité Stratégique **est constitué du** Conseil de Direction et des Hauts Responsables. (Engl. The Strategic Committee **is constituted by** Direction Council and Officials)

Due to the polysemy of the key term *constituer* (*to constitute*), there is another meaning associated with this structure, illustrated by:

> La plupart des roches détritiques **sont constituées** essentiellement **de** grains séparés… (Engl. Most of detritic rocks **are formed by** separated grains…)

In this case, *constitué* denotes a "Stuff / Object" relation, the criterion for distinguishing it from the previous one being that one can say that *detritic rocks* are made of *grains* while one cannot say that the *Strategic Committee* is made of *Direction Council*. The "Stuff / Object" sense occurs mainly in the GEO corpus. This polysemy does not justify discarding the pattern but highlights the role of the corpus regarding the semantics of the pattern.

As a consequence, one could ask whether any generic pattern do exist. The point concerns both the semantics of the pattern and its performance, conceived of as a trade-off between productivity (the pattern yields a lot of contexts) and accuracy (the contexts it yields are relevant with regard to the task). In this sense, none of the patterns collected maintains a constant performance throughout different corpora, as shown by Table 1 and Table 2 in the Results section.

However, from our point of view, even if a pattern proves to be unproductive (it yields a few or no contexts), noisy (it yields a great number of irrelevant contexts) and/or polysemous, it deserves to be included in the tool base of generic relations, for no one can say in advance what the results will be with a new corpus. So the notion of "generic database" evolves from a base that stores generic patterns to a base that stores well-known and probably reusable patterns, together with pieces of information about their performance in different corpora. As a result of this experiment, the content of the base is to be viewed more as a bootstrap for constructing patterns by reusing already tested ones than as a set of confident and "ready-to-use" patterns.

7.3 Needed skills to perform the evaluation task

Although we are able to give results in terms of precision, we must point to the major difficulty that comes with the evaluation task. Since we wanted to test the patterns on several texts, we assembled corpora from various domains and this entails that we are not able to really understand some of the contexts we had to check. This is especially true for the "CRATER" corpus, whose subject matter is telecommunications. We decided to use this corpus because we already knew that it contained several occurrences of target sentences, and because it is a multilingual corpus, which fits the purpose of pattern elaboration for English, but actually, it proved difficult to understand for a non-specialist.

Nevertheless, the context may sometimes be so clear that one can recognize the intended relation, even if the sentence to be checked is not completely understood, for instance:

> La Lme **est la partie de** l'entité de couche qui gère les ressources… (Engl. The Lme **is that part of** a Layer Entity which manages resources…)

Even if one does not know what *the Lme* and *a Layer Entity* are, the sentence clearly asserts that the former is part of the latter.

However, in real use, this problem may be avoided if the user of Caméléon is a specialist in the domain in which the ontology is built, or can rely on a domain expert's knowledge.

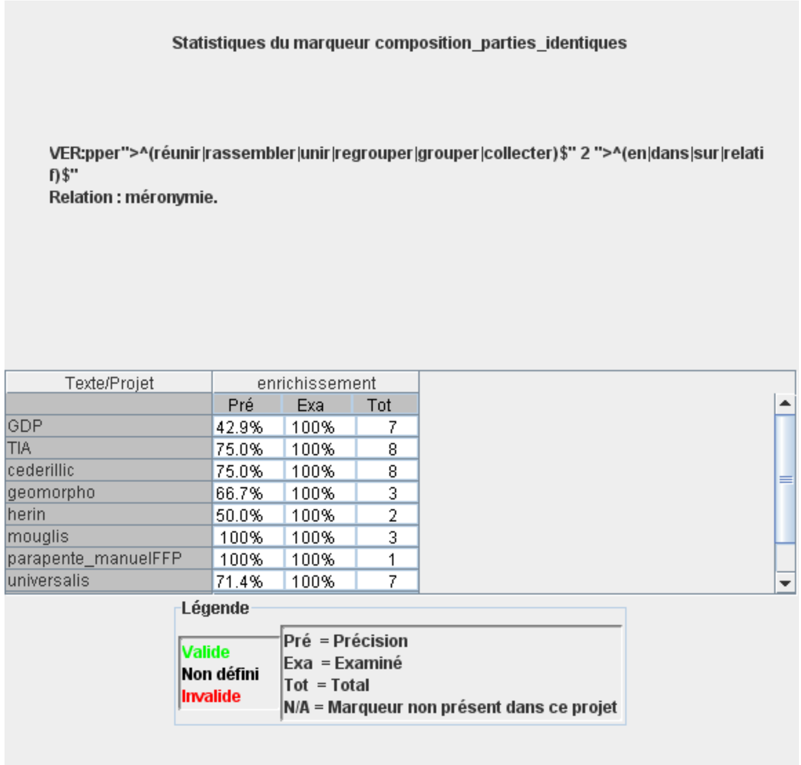
On the other hand, the intended user may encounter another kind of difficulty, related to tagset management, as we mentioned in 7.2. In order to elaborate or to adapt lexico-syntactic patterns, the user must know at least a little about morpho-syntactic categories (e.g., what a determiner or a preposition is).

A third kind of skill is needed when integrating identified concepts and relations in the ontology. Here again, some know-how in knowledge modelling and ontology engineering is required to efficiently define concepts and relations in the model.

To sum up, the ideal user must be a team who knows both about the specific domain under investigation, about language itself and about ontology engineering!

7.4 Need for documenting patterns

Due to the relative complexity of pattern reuse, we decided to determine new features to provide reusable patterns in Caméléon. As no one can predict how relevant a given pattern will be when used with new texts, the best thing to do is to report pieces of information about its elaboration and its previous use with other texts. Hence pattern storage in Caméléon has been enriched with: (a) examples of intended contexts; (b) the results of previous tests, in terms of productivity and precision; (c) a description of the corpora it has been evaluated against; and (d) any additional piece of information that could help to evaluate the pattern. Then a new functionality has been developed in the tool so that this information could be presented before the selection or rejection of reusable patterns for a new project. Tables, such as those shown in Figures 5 and 6, offer a synthetic view of this information, which can be shown either separately for each pattern (Figure 5) or globally for a list of selected patterns, such as the whole base, for all stored projects (Figure 6).



Statistiques du marqueur composition_parties_identiques

```
VER:pper">^(réunir|rassembler|unir|regrouper|grouper|collecter)$" 2 ">^(en|dans|sur|relati
f)$"
```

Relation : méronymie.

Texte/Projet	enrichissement		
	Pré	Exa	Tot
GDP	42.9%	100%	7
TIA	75.0%	100%	8
cederillic	75.0%	100%	8
geomorpho	66.7%	100%	3
herin	50.0%	100%	2
mouglis	100%	100%	3
parapente_manuelFFP	100%	100%	1
universalis	71.4%	100%	7

Légende

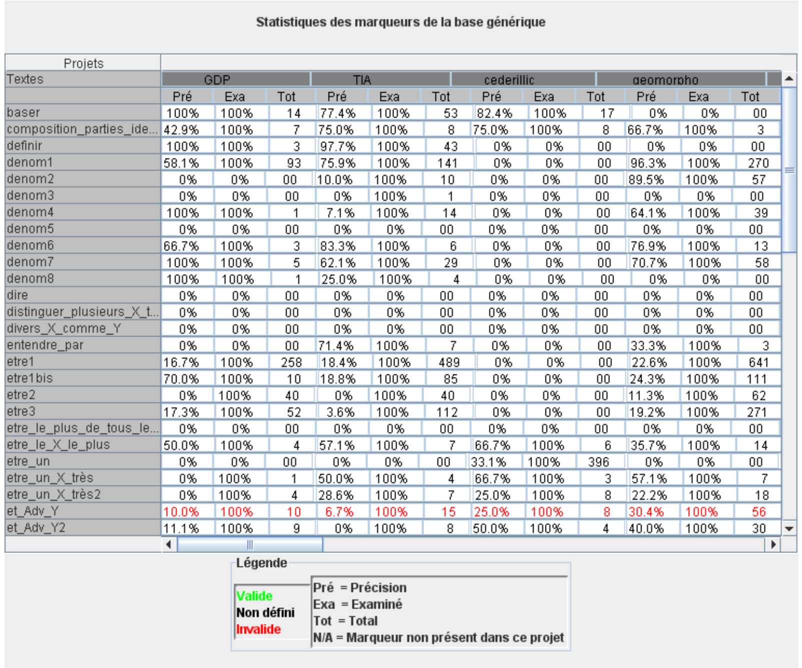
Valide
Non défini
Invalide
	Pré = Précision
	Exa = Examiné
	Tot = Total
	N/A = Marqueur non présent dans ce projet

Figure 5. Example of statistical data for the pattern "composition_parties_identiques". For each corpus it has been evaluated against (list in the left hand column) in a particular project (here, "enrichissement"), a measure of precision (Pré) is given, together with the percentage of analysed (Exa) sentence matches and with the absolute quantity of sentence matches (Tot).

8. Conclusion

We have presented here a tool and an approach for human-driven relation and concept identification. We have focussed on the elaboration and evaluation of the lexico-syntactic patterns of the tool base of generic relations and associated patterns. This evaluation challenges the notion of generic pattern. Patterns cannot be said to be generic, for the number of contexts they yield and their precision may vary considerably, depending on the corpus.

A first consequence of these statements was to define a new functionality in the Caméléon tool that would take advantage of the storage of the previous use



Statistiques des marqueurs de la base générique

Projets												
Textes	GDP			TIA			cederillic			geomorpho		
	Pré	Exa	Tot	Pré	Exa	Tot	Pré	Exa	Tot	Pré	Exa	Tot
baser	100%	100%	14	77.4%	100%	53	82.4%	100%	17	0%	0%	00
composition_parties_ide...	42.9%	100%	7	75.0%	100%	8	75.0%	100%	8	66.7%	100%	3
definir	100%	100%	3	97.7%	100%	43	0%	0%	00	0%	0%	00
denom1	58.1%	100%	93	75.9%	100%	141	0%	0%	00	96.3%	100%	270
denom2	0%	0%	00	10.0%	100%	10	0%	0%	00	89.5%	100%	57
denom3	0%	0%	00	0%	100%	1	0%	0%	00	0%	0%	00
denom4	100%	100%	1	7.1%	100%	14	0%	0%	00	64.1%	100%	39
denom5	0%	0%	00	0%	0%	00	0%	0%	00	0%	0%	00
denom6	66.7%	100%	3	83.3%	100%	6	0%	0%	00	76.9%	100%	13
denom7	100%	100%	5	62.1%	100%	29	0%	0%	00	70.7%	100%	58
denom8	100%	100%	1	25.0%	100%	4	0%	0%	00	0%	0%	00
dire	0%	0%	00	0%	0%	00	0%	0%	00	0%	0%	00
distinguer_plusieurs_X_t...	0%	0%	00	0%	0%	00	0%	0%	00	0%	0%	00
divers_X_comme_Y	0%	0%	00	0%	0%	00	0%	0%	00	0%	0%	00
entendre_par	0%	0%	00	71.4%	100%	7	0%	0%	00	33.3%	100%	3
etre1	16.7%	100%	258	18.4%	100%	489	0%	0%	00	22.6%	100%	641
etre1bis	70.0%	100%	10	18.8%	100%	85	0%	0%	00	24.3%	100%	111
etre2	0%	100%	40	0%	100%	40	0%	0%	00	11.3%	100%	62
etre3	17.3%	100%	52	3.6%	100%	112	0%	0%	00	19.2%	100%	271
etre_le_plus_de_tous_le...	0%	0%	00	0%	0%	00	0%	0%	00	0%	0%	00
etre_le_X_le_plus	50.0%	100%	4	57.1%	100%	7	66.7%	100%	6	35.7%	100%	14
etre_un	0%	0%	00	0%	0%	00	33.1%	100%	396	0%	0%	00
etre_un_X_très	0%	100%	1	50.0%	100%	4	66.7%	100%	3	57.1%	100%	7
etre_un_X_très2	0%	100%	4	28.6%	100%	7	25.0%	100%	8	22.2%	100%	18
et_Adv_Y	10.0%	100%	10	6.7%	100%	15	25.0%	100%	8	30.4%	100%	56
et_Adv_Y2	11.1%	100%	9	0%	100%	8	50.0%	100%	4	40.0%	100%	30

Légende

Valide
Non défini
Invalide
	Pré = Précision
	Exa = Examiné
	Tot = Total
	N/A = Marqueur non présent dans ce projet

Figure 6. Example of statistical analysis of the "generic" database. For each pattern listed on the left part of the table, for each corpus it has been evaluated against (columns of the table), a measure of precision (Pré) is given, together with the percentage of analysed sentence matches (Exa) and with the absolute quantity of sentence matches (Tot).

of each pattern. This device makes it possible to know how a pattern or a set of patterns behave when applied to different types of corpus. Such knowledge may guide the analyst when he decides to either select or reject this pattern in a new project. Another piece of feedback from this experiment concerns the complexity of the pattern definition task. This complexity justifies pattern learning from semantically tagged corpora in the literature. (Faure and Poibeau 2000) By providing a set of possible well-documented knowledge patterns, Caméléon makes it easier to define new patterns by the adaptation of existing ones or by analogy. Another way for simplifying this task could be to promote modularity when elaborating a pattern. It could be convenient to define new patterns by selecting and combining chunks of existing patterns.

Another lesson learned from this work is the high cost of sentence evaluation. Many other approaches (Reinberger and Spyns 2004) (Staab and Maedche 2001) prefer to avoid human intervention as much as possible: each sentence matched with a pattern provides a possible relation between concepts labelled with related terms. This new relation is systematically added to the conceptual model. The knowledge engineer checks the model as a whole and decides which of the learned relations are relevant or not. The alternative considered in Caméléon privileges quality and precision in the identification of the relation type and the related concepts. Because this option is very expensive and time-consuming, a more flexible alternative seems to be required. To reduce manual effort should become a priority. First, we plan to postpone human evaluation later on in the process, after the identification of candidate related terms. Then, this evaluation should not rely on all possible relations but rather on the "best" one. This assumes that some parameters could be defined by the user to rank the suggestions made by the system.

In spite of this cost, this experiment provides additional evidence in favour of human supervision of ontology learning from texts. (Aussenac-Gilles and Soergel 2005) Human understanding is mandatory at two different moments: firstly, when a set of domain specific patterns is being defined; secondly, when deciding how to integrate the knowledge identified in texts in the ontology. When a fine-grained decision has to be made about concept- or relation-definition in the ontology, the system may only provide proposals to be validated, modified or rejected by the knowledge engineer.

Notes

[1] Distributed by ELRA http://www.elra.info/ . Accessed December 2007.

[2] http://www.natcorp.ox.ac.uk/. Accessed December 2007.

[3] TIA: Terminologie et Intelligence Artificielle, special interest group of the French AFIA and GDR-I3 scientific associations, http://tia.loria.fr/. Accessed November 2007.

[4] It is the reason why, at the very beginning of the CAMÉLÉON project, these patterns were viewed as 'generic patterns'. As a matter of fact, the experiment we present in this paper changed this point of view, it is why the term 'generic' is temporarily used but this notion is now replaced with the notion of "reusable pattern".

[5] The original sentences are in French, but we give a translation below them. We put bold on the part of the sentence that matches the pattern.

[6] http://www.synapse-fr.com/. Accessed March 2008.

[7] http://www.ims.unistuttgart.de/projekte/corplex/TreeTagger/. Accessed March 2008.

[8] Which forms the CRATER corpus http://www.comp.lancs.ac.uk/ucrel/corpora.html#crater. Accessed 14 December 2007.

References

Ahmad, K. and P.R., Holmes-Higgin. 1995. "SystemQuirk: A unified approach to text and terminology." In *Terminology in Advanced Microcomputer Applications. Proceedings of the 3rd TermNet Symposium: Recent advances and user reports*. 181–194. Vienna, Austria.

Aussenac-Gilles, N., B. Biébow and S. Szulman. 2000. "Revisiting ontology design: A method based on corpus analysis." In Dieng, R. and O. Corby (eds.). *Knowledge Engineering and Knowledge Management: Methods, models and tools*. Lecture Notes in Artificial Intelligence 1937. 172–188. Berlin: Springer Verlag.

Aussenac-Gilles, N. and D. Soergel. 2005. "Text analysis for ontology and terminology Engineering." *Applied Ontology* 1(1): 35–46.

Aussenac-Gilles N., M.-P. Jacques. 2006. "Designing and evaluating patterns for ontology enrichment from texts." In Staab S., V. Svatek (Eds.), *International Conference on Knowledge Engineering and Knowledge Management EKAW 2006*, Lecture Notes in Artificial Intelligence 4248, 158–165. Springer-Verlag, ftp://ftp.irit.fr/IRIT/CSC/EKAW2006defin-LNCS4248-0158.pdf.

Barrière, C. 2001. "Investigating the causal relation in informative texts." *Terminology* 7(2): 135–154.

Bontcheva, K., V. Tablan, D. Maynard and H. Cunningham 2004. "Evolving Gate to meet new challenges in language Engineering." *Natural Language Engineering* 10(3–4): 349–373.

Bourigault, D. 2002. "Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus." In *Actes de Traitement Automatique des Langues Naturelles (TALN 2002)*. 75–84, Nancy France.

Byrd, R. and Y. Ravin. 1999. "Identifying and extracting relations in text." In *Proceedings of NLDB (Natural Language for Information Systems*) 99. Klagenfurt, Austria. http://citeseer.ist.psu.edu/403130.html. Accessed December 2007.

Charlet, J., M. Zacklad, G. Kassel and D. Bourigault (eds.). 2000. *Ingénierie des connaissances Evolutions récentes et nouveaux défis*. Paris : Eyrolles.

Cimiano, P., A. Pivk, L. Schmidt-Thieme and S. Staab. 2005. "Learning taxonomic relations from heterogeneous evidence." In Buitelaar, P., P. Cimiano and B. Navigli (eds.). *Ontology Learning from Text: Methods, evaluation and applications*. 59–73. Amsterdam: IOS Press.

Cimiano, P. 2007. *Ontology Learning and Population from Text. Algorithms, evaluation and applications*. Berlin: Springer.

Condamines, A. 2002. "Corpus analysis and conceptual relation patterns." *Terminology* 8(1): 141–162.

Condamines, A. and M.-P. Jacques. 2006. "Le repérage de l'hyperonymie par un faisceau d'indices : mise en question de la notion de « marqueur »." In *Actes de la Journée « Textes et connaissances » de la Semaine de la Connaissance SdC2006*. Nantes, France. http://www.sdc2006.org/cdrom/contributions/Condamines.pdf. Accessed December 2007.

Daoust, F. 1996. *SATO (Système d'analyse de texte par ordinateur), Version 4.0, Manuel de référence*, Service d'analyse de texte par ordinateur (ATO), Montréal : Université du Québec à Montréal (http://www.ling.uqam.ca/sato. Accessed December 2007.

Desclés, J.-P. 1997. « Systèmes d'exploration contextuelle », In Guimier C. (ed.). *Co-texte et calcul du sens*. 215–232. Caen: Presses Universitaires de Caen.

Faure, D. and T. Poibeau. 2000. “First experiments of using semantic knowledge learned by ASIUM for information extraction task using INTEX.” In *ECAI-2000 Workshop on Ontology Learning*. 7–12. Berlin Germany.

Feliu, J. and M.T. Cabré Castellví. 2002. “Conceptual relations in specialized texts: New typology and extraction system proposal.” In *TKE’02, 6th International Conference in Terminology and Knowledge Engineering*. 45–49. Nancy France.

Garcia, D. 1998. *Analyse automatique des textes pour l’organisation causale des actions, Réalisation du système Coatis*. Thèse de Doctorat en informatique, Université Paris IV-Sorbonne.

Gillam, L., M. Tariq and K. Ahmad. 2005. “Terminology and the construction of ontology.” *Terminology* 11(1): 55–81.

Girju, R. and D. Moldovan. 2002. “Text mining for causal relations.” In *Proceedings of FLAIRS. 2002*. 360–364. Pensacola Beach, Florida.

Hearst, M. 1992. “Automatic acquisition of hyponyms from large text corpora.” In *Proceedings. of the 15th International Conference on Computational Linguistics (COLING-92)*. 539–545. Nantes, France.

Jackiewicz, A. 1996. “ L’expression lexicale de la relation d’ingrédience (partie-tout).” *Faits de langues* 7. 53–62.

Jacques, M.-P. and N. Aussenac-Gilles. 2006. “Variabilité des performances des outils de TAL et genre textuel.” *Traitement automatique des langues* 47(1) : 11–32.

Kavalec, M. and V. Svaték. 2005. “A study on automated relation labelling in ontology learning.” In Buitelaar, P., P. Cimiano and B. Magnini (eds.). *Ontology Learning from Text: Methods, evaluation and applications*. 44–58. Amsterdam: IOS Press.

Marshman, E., T. Morgan and I. Meyer. 2002. “French patterns for expressing concept relations.” *Terminology* 8(1): 1–30.

Marshman, E. and M.-C. L’Homme. 2006. “Disambiguating lexical markers of cause and effect using actantial structures and actant classes.” In *Proceedings of the 15th European Symposium on Language for Special Purposes, LSP 2005*. 261–285. Bergamo, Italy.

Morin, E. 1999. “Des patrons lexico-syntaxiques pour aider au dépouillement terminologique.” *Traitement automatique des langues* 40(1) : 143–166.

Pearson, J., 1998, *Terms in Context*, Amsterdam/Philadelphia: John Benjamins.

Rebeyrolle, J., 2000, *Forme et fonction de la définition en discours*. Thèse de Doctorat en Sciences du Langage, Université Toulouse II-Le Mirail.

Rebeyrolle, J. and L. Tanguy. 2000. “Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires.” *Cahiers de grammaire* 25 : 153–174.

Reinberger, M.-L. and P. Spyns. 2004. “Discovering knowledge in texts for the learning of DOGMA-inspired ontologies.” In *ECAI-2004 Workshop on Ontology Learning and Population*. 19–24. Valencia, Spain.

Riloff, E. 1996. “Automatically generating extraction patterns from untagged text.” In *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI-96)*. 1044–1049. Portland, Oregon.

Sabou, M. 2004. “Extracting ontologies from software documentation: A semi-automatic method and its evaluation.” In *ECAI-2004 Workshop on Ontology Learning and Population*. Valencia, Spain. http://olp.dfki.de/ecai04/OLP-Proceedings.zip. Accessed December 2007.

Séguéla, P. 2001. *Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques*. Thèse de Doctorat en Informatique, Université Toulouse III Paul Sabatier.

Schutz, A. and P. Buitelaar. 2005. “RelExt: A tool for relation extraction from text in ontology extension.” In Gil, Y., E. Motta, V.R. Benjamins and M. Musen (eds.). *The Semantic Web — Proceedings of ISWC 2005: 4th International Semantic Web Conference, ISWC 2005*, Galway, Ireland. Lecture Notes in Computer Science 3729. 593–606. Berlin: Springer Verlag.

Staab, S. and A. Maedche. 2001. “Ontology learning for the semantic Web.” In *IEEE Intelligent Systems*. Special Issue on the Semantic Web. 16(2): 72–79.

Velardi, P., R. Navigli, A. Cuchiarelli and R. Neri. 2006. “Evaluation of Ontolearn, a methodology for automatic learning of domain ontologies.” In Buitelaar, P., P. Cimiano and B. Magnini (eds.). *Ontology Learning from Text: Methods, evaluation and applications*. 92–106. Amsterdam: IOS Press.

Velardi, P., A. Cucchiarelli and M. Petit. 2007. “A taxonomy learning method and its application to characterize a scientific Web community.” In *IEEE Transactions on Knowledge and Data Engineering* 19(2): 180–191.

Authors’ addresses

Nathalie Aussenac-Gilles
Institut de Recherche en Informatique de Toulouse (IRIT) — CNRS UPS
118, route de Narbonne
31062 Toulouse Cedex 9, France

aussenac@irit.fr

Marie-Paule Jacques
Université Marc Bloch Strasbourg 2
UFR LSHA, 22 rue René Descartes
BP 80010
67084 Strasbourg Cedex

marie-paule.jacques@umb.u-strasbg.fr

About the authors

Nathalie Aussenac-Gilles graduated in 1986 from an engineering school in computer science and obtained her PhD. in 1989. She became a CNRS (French National Research Agency) researcher in 1991. Since then, she is a member of the IRIT laboratory in Computer Science at University “Paul Sabatier” of Toulouse. Her research interests include knowledge engineering, natural language processing and terminology based approaches for ontology engineering from text. Her work is influenced by long term cross-disciplinary collaborations with linguists and researchers in human factors.

Marie-Paule Jacques obtained her PhD in Linguistics in 2003. After temporary positions as assistant professor at Toulouse le Mirail University and post-doc researcher at IRIT (Paul Sabatier University, Toulouse) and at LIPN (Laboratoire d’Informatique de Paris-Nord), Paris 13 University, she is lecturer at Marc Bloch University, Strasbourg, since 2007. Her research interests focus on corpus linguistics, Natural Language Processing and linguistic studies that aim at improving NLP (such as discourse organisation, anaphora processing, tagging and parsing), terminology and retrieval of conceptual relations from texts. She has been involved in projects that implied collaborations with other linguists, with researchers in informatics and with psychologists.