Searching for and identifying conceptual relationships via a corpus-based approach to a Terminological Knowledge Base (CTKB)

Method and Results

Anne Condamines
CNRS
Equipe de Recherche en Syntaxe et Sémantique
Maison de la Recherche, 5 allées A. Machado
31058 Toulouse cedex, France
acondami@univ-tlse2.fr

Josette Rebeyrolle
Université Toulouse-le Mirail
Equipe de Recherche en Syntaxe et Sémantique
Maison de la Recherche, 5 allées A. Machado
31058 Toulouse cedex, France
rebeyrol@univ-tlse2.fr

Abstract

The aim of this paper is to provide an effective method for constructing a Corpus Based Terminological Knowledge Base –CTKB. The two stages of the method based on linguistic knowledge are outlined : the identification of candidate-terms and the identification of conceptual relationships. This method is applied to a French corpus and the results are assessed from the point of view of various applications.

Introduction

Ever since the concept of a Terminological Knowledge Base (TKB) was first created (Meyer et al., 1992), a number of teams have worked on defining

models, methods and formalisms associated with this new concept. Research is now directed into two distinct fields depending on the interests of the research teams. Those with a background in Artificial Intelligence have worked on issues related to the formalization of terminological data, which has resulted in modeling experiments using conceptual graphs (Gillam and Ahmad, 1996) or using descriptive logic (Biébow and Szulman, 1997). Research teams, including our own, with a background in linguistics or terminology have been more interested in the extraction of data from texts. The concept of the TKB has now been revised and re-examined according to our aims to such an extent that it has become justified to re-organize the work into two categories. This is our conclusion after taking part in a number of multi-disciplinary projects[1]. In this paper, we describe the construction and validation of a TKB as a corpus-based approach to a terminological knowledge base (the CTKB), as opposed to an application-based and formalized TKB (ATKB). The main point of this paper is the presentation of the method for searching for and identifying conceptual relationships which play a very important role both in selecting terms and in constructing a model of corpus.

The results presented here aim to show the feasibility of constructing a CTKB. We show that it is possible, with appropriate systems and linguistic interpretation, to model a text, particularly the conceptual relationships contained in it. This demonstration is based on our most recent experiments, which have been carried out using a corpus supplied by the EDF (French State Electricity Board). However, our involvement with this type of work dates from around five years ago with corpora from Matra Marconi Space (Condamines and Amsili, 1993), Aérospatiale, and the Centre National d'Etudes Spatiales.

We firstly describe the CTKB model that we have set up and we outline a data constitution method. Finally, we will demonstrate the validity of the CTKB concept, both as a corpus model, as well as in its use in a number of different applications.

1. The CTKB

This section presents the main characteristics of a CTKB and the data model that we have proposed.

[1] In the "Terminology and Artificial Intelligence" group as well as in a close collaboration with researchers from IRIT (Institut de Recherche en Informatique de Toulouse) on a Cognitive Sciences GIS project (Groupement d'Intérêt Scientifique), which aimed to evaluate the possibilities of using a TKB for various applications including the constitution of a Knowledge Base System.

1.1. *CTKB characteristics*

Our experience in constructing terminological knowledge bases from corpora enables us to list the following characteristics for a CTKB:

> The construction of the CTKB is not dependent on a given application. Any choice can be justified according to the text studied (cf. 3.), using linguistic analyses.
>
> Such a CTKB is not formalized. There are no coherence or completeness calculations, although such calculations could be envisaged to assist the linguist / terminologist to design the CTKB.
>
> Experts, if required at all, are only brought in for the final stage to guarantee the validity of the CTKB concerning the textual content (and not with respect to their knowledge of either the domain or of a particular field of application).

This CTKB is conceived as a kind of text modeling (which we here contrast with formalization), a kind of pre-processing which can be used by anyone having to read the text for a specific application. The CTKB must be usable in a variety of applications, to create an index or to rewrite the text (as planned in the GIS project) or again to create a knowledge base system for a specific application. The CTKB data can be modified and adapted as required, such modifications being justified by the application and/or the formalization.

We also describe how this CTKB has been used for different applications (cf. 4.).

1.2. *The CTKB Model*

In the last few years, we have developed a methodology for defining a CTKB model which would allow us to understand the functioning of terms and concepts in a given corpus. The four fields in this model contain the following information:

The “terms” contains strictly linguistic data (type and gender, form variants, acronyms, abbreviations, etc.).

The “concepts” contains data concerning the concept denoted by the term in the form of a definition and of explicit conceptual relationships. The choice of relationships is not restricted, but they are strictly defined.

The “term/concept link” contains information on how valid the term is for the description of a particular concept, that is, its use in a given sub-domain or company, for example.

The "corpus" is used to establish the links between a term and its occurrences in the text which the CTKB models. For example, the text is also used to illustrate each Term - Conceptual relationship - Term triple.

The specification and development of a management and consultation tool (GEDITERM, cf. Aussenac, 1999) based on this model has been carried out by our colleagues at the 'Institut de Recherche en Informatique de Toulouse', Nathalie Aussenac's team.

2. Method for constructing a CTKB

The method we propose for constructing a CTKB is intended to meet the following requirements:

To automate each stage of the research using corpus-processing systems.

To define systematic criteria in order to evaluate, interpret and organize the results provided by these systems.

This method is based on the same assumption as in Pearson (1998). Pearson's goal is to discover explanations of terms as input for the formulations of specialized definitions. Meyer and her team (Meyer, this volume) have exactly the same goal when they identify knowledge-rich contexts in order to develop knowledge-extractions tools. As far as we are concerned, our aim is to build a model of the text, i.e, a CTKB.

We describe here the method for identifying the CTKB model's data. It consists of two major steps:

Identifying Candidate Terms (2.1).

Identifying conceptual relationships. This main step will be described in details below (2.2).

2.1. *Identifying Candidate Terms*

This involves establishing a list of Candidate Terms which will then be used for the identification of conceptual relationships on the basis of the results yielded by the terminological extraction systems.

2.1.1. *Noun and adjective Candidate Terms*

The tools for identifying noun and adjective phrases (cf. Daille et al., 1994; Bourigault et al. 1996; Enguehard and Pantéra, 1995) are used both to identify phrasal variants and to establish an initial list of terms. However, the lists of phrases provided by these systems contain certain word groups which are not real terms such as, *complément du paragraphe (addition to paragraph)*, *mise à jour du présent document* (*update of current document)*, etc. We have therefore defined a set of relatively stable criteria for removing this type of phrase and for delimiting a relevant sub-set of Candidate Terms. To this list are added the term variants, ellipsis and acronyms which are also provided by some terminology extraction systems, in order to search all the conceptual relationships. During the search for conceptual relationships, acronyms and variants are considered as denoting the same concept and can be substituted for the Candidate Terms in order to increase the search coverage.

2.1.2. *Verb Candidate Terms*

The majority of Candidate Terms extraction systems only selects noun phrases and sometimes a few adjectives. It is nevertheless possible to identify verb phrases from the noun phrases (for example, *gestion de configuration (configuration management) / gérer la configuration (managing the configuration*)). It is then necessary to manually look for those effectively used in the corpus. This initial list of verbs is then expanded by studying the conceptual relationships.

2.2. *Searching for and identifying conceptual relationships*

The search for conceptual relationships plays an important role in building a CTKB as long as it is mainly a model of the text content. From this point of view, the most important knowledge within the text is conveyed by conceptual (or semantic) relationships. Conceptual relationships are provided by some parts of the text, by means of certain linguistic patterns, to which we can associate a non-ambiguous interpretation. This interpretation consists in a binary relationship between two elements. We are particularly interested in identifying linguistic patterns to which we can associate a non-ambiguous interpretation, it means to which all the linguistic constraints are described.

Searching for conceptual relationships consists in two complementary steps: searching for taxonomies and searching for syntagmatic relationships.

2.2.1. *Searching for taxonomies*

This first step aims at building classes, i.e. hierarchies of concepts characterized by the same inherited feature(s). Two kinds of linguistics patterns are used for searching concepts:

> those used for expressing hyponymy relationships,
>
> others used for expressing meronymic relationships. These linguistic patterns are the same as for meronymic relationships but with the additional constraint that the heads of the related Candidate Terms have to be the same[2]. This is exemplified by the relationship between *departmental road network*, *national road network* and *communal road network*. These noun phrases may be considered as meronyms of *national network*. They have the same head: *road network* and they occur with meronymic patterns: *A road network is composed of a national road network, and a regional road network*. Note that hyperonymic contexts may also be used: *the national road network is the road network managed by national organisms*. Such hyperonymic contexts are possible because all the features of the national network are inherited by the other kinds of road network.

So when either hyperonymic patterns or meronymic patterns with repeated heads occur, we consider that the connected Candidate Terms may constitute a taxonomy. We start with these patterns for two main reasons.

a) These patterns are well known. We can think that most of them are context-free and may be used in any corpus (even if they are not always used). We derived this assumption from the experiment proposed by Morin (1999) for French patterns. By searching automatically recurrent contexts combined with pairs of hyperonym terms in an agricultural corpus, he identified almost the same patterns as the ones identified by Borillo (1997) intuitively. So, we can conclude that, for the two relationships, some specific patterns may appear in some corpora but the core set of patterns is stable and can be used as a starting point. We have compiled and used hyperonymic and meronymic French linguistic patterns proposed by Borillo (1997) and Jackiewicz (1996).

b) The taxonomies we obtain by using theses patterns are characterized by the fact that all members have one or more shared features. So, we think that, in some contexts, these members are substitutable. In the second part of the research (section 3), we will show how we exploit this paradigmatic characteristic in order to identify syntagmatic relationships.

---

[2] When the heads of the related terms are different, there is no automatic way to decide if it is a meronymic or a hyponymic relationship while no formal mark appears. Nevertheless, with human interpretation, it is possible to take into account general linguistic competence to conclude that there is a common feature between the related terms.

Taxonomies are built following two steps:

1- Looking in the corpus for contexts where one of the patterns appears in association with two Candidate Terms, for example:

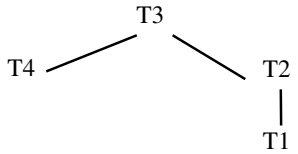
> a CT1 is a CT2 which
>
> a CT1 is split into CT2 and CT3 (CT1, CT2 and CT3 have the same head).

Several types of results are obtained:

> bad contexts; i.e. contexts not showing the expected relationship or showing the good relationship but within a specific or subjective point of view, as in the following example: *Le plan de développement est le document le plus difficile à réaliser*[3]. Such contexts are abandoned.
>
> good contexts, i.e. contexts showing the expected relationship with a generic point of view. In this case, lists of term pairs are created.

2- Building taxonomies: term pairs are combined in order to build hierarchies. All pairs of terms that share a common term may be combined.

If we have T1 R T2, T2 R T3, T4 R T3, we can build the following hierarchy:



Then, we build several taxonomies with at least one common feature (see below examples of such taxonomies).

Note that when a Candidate Term has been identified, combined with a linguistic pattern, it becomes a term.

2.2.2. *Searching for syntagmatic relationships*

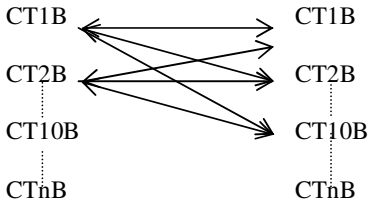
In order to search for syntagmatic relationships, we start from the hypothesis that in corpora, these relations may be identified by means of linguistic patterns including elements belonging to taxonomies. So, we try to identify syntagmatic relationships from paradigmatic relationships; these paradigmatic relationships are the ones used for building taxonomies in the previous step. Then, the aim is to identify both conceptual relationships and their associated linguistic patterns; the method consists in combining taxonomies two by two.

---

[3] *The development plan is the most difficult document to write.*

1— In the corpus, each Candidate Term is marked up with the feature(s) of the taxonomy to which it belongs.

2— An extract of each taxonomy is combined with an extract of all the other taxonomies in order to bring out contexts where the same relationship is potentially expressed. Consider the following taxonomies A and B and their members $CT^{4}_{1}A$ to $CT_{n}A$, for the first one and $CT_{1}B$ to $CT_{m}B$, for the other one. Data to be analyzed would be all the combinations between all the CT of the A taxonymy with all the CT of the second (B).



Note that for each combination, several contexts may be rpovided. Finally, contexts to be analyze may be very numerous. So, we reduce the number of the examined combinations by choosing arbitrarily to analyze only the combinations between the 10 first members of A with the 10 first members of B. Even with this selection the number of contexts to be analyzed may be very high. If this is the case, we can first examine the data where terms from A are combined with terms from B and then, the data where terms from B occur with terms from A. Actually, the combinations are oriented within the sentence, they are not always reversible. In some cases (but not all), the reversible combination corresponds to the passive voice and there is only one triple CT_1 R CT_2, but in other cases, there are two triples CT_1 R_1 CT_2 and CT_2 R_2 CT_1.

The aim of the analysis is to identify recurrent contexts that may be interpreted as referring to the same relationship. At this point we use our linguistic knowledge allowing us to interpret these contexts as referring to the same relationship. These contexts may either be equal, contain similar elements, or be completely different. What is important is that they can refer to the same relationship.

With such an approach, we get a lot of noise because it is not frequent that the context between two terms, even if they potentially label connected concepts, refers to the concerned relationship.

---

[4] where CT = Candidate Term

3— For each relationship, all the linguistic patterns are strictly specified with all the restrictions (morphologic, semantic and syntactic) in order that by applying them, we get as little noise as possible.

4—The linguistic patterns identified are projected onto the corpus, with all the Candidate Terms (not just with the CT from the classes concerned). The aim is to identify new pairs of terms. At the same time, each term is added to the relevant taxonomy.

5— These new pairs of terms are projected onto the corpus so as to bring out new patterns.

Steps 3, 4 and 5 are repeated until the results are saturated.

Notes:

a) The combination of classes concerns also the combination within a single class. There are two reasons for this:

On the one hand, some artificial cuts within taxonomies may appear in the corpus. For example, the depth level may not be the same within all the taxonomies built from the corpus. Moreover, in some cases, just one taxonomy is extracted from the corpus whereas it seems that it could be split up into two or more sub-taxonomies. In some other cases, it seems that taxonomies could be combined, even if no superordinate concept appears in the corpus.

On the other hand, the unity of each taxonomy is very often given just by one feature (see 3.3 for examples). It is possible for elements having just one common feature to be linked by a conceptual relationship. Take the case of the human taxonomy. It is very common to meet nouns with human feature linked by a conceptual relationship within corpora. For example, a *human may (cure, be in charge of, communicate with...) an other human*. So, the fact that two Candidate Terms may be in the same taxonomy does not prevent themselves from being linked by a relationship other than the hyperonymic one.

b) In the case of a small corpus, the second stage may be applied to all the elements of the classes, not just to a subset of them. In such case, steps 4- and 5- are omitted.

c) One of our aims is to amass linguistic patterns associated both with relationships and corpus type (domain, text genre). So, in the future, the second step will begin with the application of known linguistic patterns in order to accelerate the identification of conceptual relationships. Pairs of identified terms will then be used for bringing out new linguistic patterns either generic, or specific to the corpus.

This could be the case here with meronymic patterns. While we used them (with the constraint on heads) for searching for hyperonymic relationships, we could

use them for identifying meronymic relationships; however, in order to avoid complicating the presentation, we have not developed this step

2.2.3. *Tools for helping the search and identification of relationships*

A new generation of tools are devoted to identifying relationships in French corpora; they use either a top-down or a bottom-up method. The former apply predefined linguistic patterns. This is the case for Seek for hyperonymic relationships (Jouis, 1994), Coatis for causal relationships (Garcia, 1997). The latter help in the search for and identification of relationships in texts, without any *a priori* linguistic knowledge. This is the case for Mantex (Rousselot et al., 1996) and Prométhée (Morin, 1999). Some tools combine top-down and bottom-up methods such as Caméléon (Séguéla, 1999). But, our work has a more linguistic aim on the one hand, and a more application-oriented aim on the other hand (we build CTKB for firms). Our main linguistic issues are the following:

> What does one mean by linguistic patterns for relationships?
>
> What are the links between a corpus (domain and genre) and some linguistic patterns?
>
> Is the hypothesis that it is possible to build just one model from the corpus valid?

We think that, in order to re-use linguistic patterns, it is necessary to specify as much as possible two kinds of elements:

> which linguistic constraints are necessary so that the linguistic patterns bring out only good contexts - that is to say contexts associated with just one relationship, without ambiguity.
>
> which characterizations of corpora must be added to combine with the description of linguistic patterns in order to use only the good ones according to the nature of the corpus studied.

Then, any kind of tools helping us in our plan is welcome but all of these tools can only propose candidates (terms or relationships); none of them can interpret the output results. However, interpreting the output is time consuming and requires detailed analysis of contexts; these results may just constitute a starting point for us.

2.3. *Results obtained*

As a result of the 5— step procedure described in 2.2.2, the following data are available and accessible by means of the CTKB system:

> the list of terms finally selected;

the list of pairs related by a given conceptual relationship;

for each conceptual relationship, its characteristic patterns;

for each "term-relationship-term" triple, the section(s) in the corpus which contains it.

3. Experiment on a French corpus

The experiment was carried out on a French handbook for software engineering specification (478 KB) used by EDF (Electricité De France). Note that this corpus is highly descriptive and explanatory because its main aim is to provide recommendations. Our hypothesis was that this corpus, relatively to its genre, includes a large number of knowledge rich contexts.

Our aim here was twofold: first, to provide some quantified data using a corpus; second, to illustrate how we carried out the different steps of the method for constructing a CTKB.

3.1. *Identifying Candidate Terms*

To this end, we used the LEXTER (Logiciel d'EXtraction de TERminologies) software, developed at the EDF by Didier Bourigault (Bourigault, 1996). A set of linguistic criteria was applied to reduce the 5878 Candidate Terms proposed by LEXTER to 1516 Candidate Terms, that is 22% of the initial list, which were to be included for the identification of conceptual relationships (Condamines and Rebeyrolle, 1997).

The identification of verbs is not discussed here at any length. We just stress that we have identified 27 verbs, as *to distribute, to develop, to modify, to test* from noun phrase such as: *distribution activity, development activity, modification activity, test activity*.

We are therefore concerned here with the method used for searching conceptual relationships.

3.2. *Searching for conceptual relationships*

In our experiment, we used the "Système d'Analyse de Textes par Ordinateur" (SATO) developed at the ATO center in Montreal[5] because it has a number of interesting functions for implementing our method. The most significant one is

[5] http://www.ling.uqam.ca/sato/outils/sato.htm

enables the user to characterize text elements (from the morpheme to the textual segments) in order to search for structures indicating conceptual relationships in the corpus. This function compensates for the absence of syntactic categorisation. Moreover, consulting the text is a very flexible process that makes it possible to modify or add structures to take into account linguistic knowledge which is specific to a sub-domain and/or a text genre.

The application of the method described in 2.2. led to the following results: 404 concepts and 455 C-terms. The difference between these two figures can be explained by the fact that 51 of the terms are synonymous, that is to say, that two or more terms can denote the same concept.

Table 1 summarizes the main types of conceptual relationships found in the corpus:

Conceptual relationships	Number of markers	Number of related pairs
is a	21	204
is composed of	41	128
occurs in	1	8
occurs during	1	59
starts during	1	5
ends during	1	8
precedes	1	16
conditions the start of	1	11
conditions the end of	1	14
is the result of	1	8
is updated during	1	7
is responsible for	1	3
has responsibility for	2	50
plays the role of	1	5

Table 1: Number of patterns and pairs connected by each of the CTKB conceptual relationships

It is not surprising that for the most specific structures we only have a single characteristic structure, because even if there are several verbs, the structure itself is unchanged (except where a transitive verb becomes an intransitive verb, for example). Furthermore, the total number of concepts greatly exceeds the 404 specified since a given concept has relationships with many other concepts in the corpus.

Hitherto, we have concentrated on describing the linguistic patterns of the different kinds of relationships. The next section details with examples the steps described in 2.2.

3.3. *Searching for taxonomies*

The hyponymy relationship is marked by generic phrases such as: *Le/un/les* $N_{hyponyme}$ *est un/des* $N_{hyperonyme}$. However, these phrases are not precise enough to relate only hyperonymic pairs. Therefore, we think that some constraints can play an essential role in their identification. We have proposed a broader conception of marker, in terms of configurations of lexical, syntactic, typographical and layout features (Pery-Woodley and Rebeyrolle, 1998).

From the 21 structures indicating hyperonymy, we can bring out:

§[6] + def_det + CT1 + Vtobe (present) + undef_det + {kind, type, etc. of} CT2 + {relative clause, past participle, present participle, adjective}

to quote an example from this structure: *§ Le guide d'élaboration de la documentation de spécification est un guide méthodologique pour la production des documents de spécification des logiciels scientifiques*[7].

§ + def_det + CT1 + Vtobe (present) + undef_det + CT2 + {relative clause, past participle, present participle, adjective}

For example: § *La documentation de spécification est un des documents essentiels d'un projet*[8].

By applying the hyperonymic structures in our corpus, we found 199 pairs of concepts. It must be stressed that the hyperonymic relationship is also established using morphological criteria. Consider the following noun phrases: *dossier de conception, dossier de test, dossier de modification*. It is possible to say that *dossier* is the hyperonym of these noun phrases.

It will be repeated that certain linguistic patterns may denote hyponymy or meronymy as well. However, in the case of meronymy, the heads of the related

---

[6]Insofar as the typographical and dispositional markings are concerned, we simply note here that three constraints appear to be sufficient for an exclusively hyponymy pattern: the fact that the structure is located at the start of a paragraph (shown by the symbol §), that the CT2 is typographically marked (by a bold or italic font, for example). The CT1 is a definite noun phrase (notated def_det). The CT2 is always an indefinite noun phrase (notated undef_det).

[7] *The handbook of specification documentation development is a methodological handbook for specification documents of scientific software production.*

[8] *The specification documentation is one of the essential documents of a project.*

Candidate Terms must be the same. In this study, we have identified only five pairs using to this pattern. It is exemplified by the relationship between *dossier de conception* (*conception document*) and *dossier de conception générale* (*general conception document*), *dossier de conception détaillée* (*specific conception document*) where *dossier de conception* (*conception document*) is the hyperonym of *dossier de conception générale* (*general conception document*) et *dossier de conception détaillée* (*specific conception document*).

The semantic analysis of the 204 pairs identified with relevant linguistic patterns enables us to grade these concepts according to their semantic features. We have distinguished four classes:

— a class of activity noun (classA) as *specific conception*

— a class of document noun (classD) as *specific conception document*

— a class of human noun (classH) as *project leader*

— a class of time section noun (classT) as *conception stage*

3.4. *Searching for syntagmatic relationships*

As seen in 2.2.1, non-taxonomic relationships have been identified by combining taxonomies. 16 combinations have been examined, as classA {context} classD, classA {context} classH, etc. and their reverse.
In the following paragraphs, we firstly present the syntagmatic relationships (3.4.1), and secondly we take an example of the different steps of the method, (3.4.2).

3.4.1. *Presentation of syntagmatic relationships*

Our linguistic analysis of combinations has shown that half of them (eight combinations) are particularly interesting.

$\text{classD}_{(\text{document})}$ {context} $\text{classD}_{(\text{document})}$

The relationship identified is "**is composed of**", a well known type of meronymic relationship. In order to analyze results and identify linguistic patterns, we have used our knowledge of French, on the one hand, and the results of work carried out on meronymic patterns (such as that of Jackiewicz, 1996) on the other hand. We have described a set of lexico-syntactic patterns for this relationship, including:

CT2 + Vtobe (present) + undef_det {group, family, etc.} + of + CT1

Here is an example of this pattern: *Un Etat de Configuration est une famille de fichiers sources*[9].

CT2 + {contains, encompasses, includes, etc.} + CT1

For example: *Les unités documentaires sont constituées des documents de spécification et de conception*[10].

This relationship connects 133 pairs of concepts in our corpus.

$classA_{(activity)}$ {context} $classA_{(activity)}$

The relationship identified is "**precedes**". We have called this relationship "precedes" because it stresses firstly how activities are linked together, and secondly how time sections are related (see below).

$classT_{(time)}$ {context} $classT_{(time)}$

The "precedes" relationship also holds between time sections. Another relationship identified by this combination is "**occurs during**". The following sentence provides an example: *Le cycle de développement Composant se déroule pendant la phase de réalisation Produit*[11].

$classA_{(activity)}$ {context} $classT_{(time)}$

This combination is the most frequent one. The principal differences between the five relationships identified stem from their semantic significance.

The first two relationships identified are: "**starts during**" and "**ends during**". In this type of combination, the verbs which hold between the terms are *to conclude, to terminate, to finish, to start, to begin*, for example.

We have added the "**occurs during**" relationship which can also stress how activities are organized within a time frame. Here is an example: *Les vérifications qualité doivent être réalisées au cours de la phase de conception générale composant*[12].

Two other relationships have been identified: "**conditions the start of**" and "**conditions the end of**". They have been described thoroughly in another paper (Condamines and Rebeyrolle, 2000).

$classD_{(document)}$ {context} $classT_{(time)}$

---

[9] *A configuration state is a family of source files.*
[10] *The documentary units include specification documents and design documents.*
[11] *The Component development cycle takes place during the product realisation phase.*
[12] *The quality checks must be done during the general component design stage.*

The relationship identified is "**updated during**" which indicates at which point of a process a document is modified, as shown in the following example: *Le plan de validation produit est mis à jour au cours de la phase d'intégration produit*[13].

$\text{classD}_{(\text{document})}$ {context} $\text{classA}_{(\text{activity})}$

This combination provides the relationship labeled "**is the result of**". In fact, closer examination of contexts showing a precedence relationship as opposed to one of succession has allowed us to identify pairs which are related by this relationship. It indicates which document is the product of an activity. The characteristic patterns of this relationship are *is followed by, results in, is the consequence of*, for example. As shown in Table 1, only 8 pairs of concepts are linked by this relationship which is illustrated in the following example: *Le Dossier de Spécification Logiciel est le résultat de la phase de spécification composant*[14].

$\text{classH}_{(\text{human})}$ {context} $\text{classH}_{(\text{human})}$

The two relationships concerned with this combination are "**is responsible for**" and "**plays the role of**".

This combination is established between a few pairs of concepts (cf. Table 1). The first one is exemplified by the following sentence: *Le chef de projet est le seul responsable de l'ensemble de l'équipe de projet*[15]. The second one in fact connects two human nouns when one human plays a different hierarchical function than the usual one, as in this example: l'ingénieur qualité joue le rôle du responsable qualité[16].

$\text{classH}_{(\text{human})}$ {context} $\text{classA}_{(\text{activity})}$

This is the relationship labeled "**has responsibility for**" as it occurs in this example: *Pour un projet donné, le chef de projet a la responsabilité des activités de gestion de projet*[17]. We give a specific description of this relationship below.

Notes:

a) We have added another relationship which is not expressed by linguistic patterns in the corpus but occurs implicitly in schemata: "**occurs in**". It holds between activities such as *test validation activity, test qualification activity,* etc. and location concepts such as *space of delivery, space of receipt*, etc.

---

[13] *The product validation plan is updated during the product integration stage.*
[14] *The Software Specification File is the result of the component specification phase.*
[15] *Only the project leader is accountable for the project group.*
[16] *The quality engineer plays the role of the quality leader.*
[17] *For a given project, the project leader is in charge of the projet management.*

b) We had to deal with several case of ellipsis. In some cases, using ellipses entails a wrong interpretation, as in this sentence: *Le chef de projet a la responsabilité du plan de développement*[18], where *plan de développement* (*development plan*) is an ellipsis of *rédaction du plan de développement* (*drafting of development plan*). So the relationship does not hold between a human and a document, as it seems, but between a human and an activity. This example is to be classified together with the "has responsibility for" relationship.

3.4.2. *Example of the process*

This section focuses on the "has responsibility for" relationship. The five steps described in 2.2 are applied as follows for this relationship.

1- We start the analysis by labeling (using SATO) all the CT belonging to the same class (either human, activity, time or document) on the one hand, and on the other hand, all the parts of the text containing a hyperonymic linguistic pattern. The point is that when we combine the members of our taxonomies, we do not want to reexamine the contexts found during the first steps - that is to say hyponymy pairs.

2- We examine more closely the distribution of the most frequent term of the human class — here it is *chef de projet* (*project leader*) —with all the terms of the activity noun class.

3- It shows some stable verb phrases. The linguistic analysis of the occurrences enables us to bring out a sub-class of verbs phrases as *has the responsibility for, is entrusted with, is responsible for*, as shown in the following example: *L'ingénieur qualité a la charge du contrôle qualité* [19]. Then if we look for reverse combinations where an activity noun occurs with a human noun, we obtain verb phrases, such as *to be in charge of, to be within the competence of*, which correspond to the same relationship. All these verb phrases are linguistic patterns of the "has responsibility for" relationship. They are labeled as such in the corpus.

4- The next step of the method consists in finding places in the corpus where these patterns occur with a CT. We have to note here that we have not specified the noun class (all the CT are used) which can occur with the pattern because we hope to discover nouns which have not been identified in the first step of the method. That is to say that they did not occur in a hierarchical relationship. In the case of the "has responsibility for" relationship, we found no more term by this way.

---

[18] *The project leader is responsible for the development plan.*
[19] *The quality engineer is entrusted with the quality check.*

In this section, the main steps of the method have been described and applied to a corpus: searching for conceptual relationships, identifying linguistic patterns and finally giving the term status to some Candidate Terms (that are the ones) linked by linguistic patterns.

Up to now the discussion is centred on the applicability of the method.

4. Evaluation

This section deals with two types of evaluation of the CTKB, one viewing the CTKB as a corpus model and the other evaluating the possibilities of using a CTKB for various applications.

4.1. *CTKB as corpus model*

As previously stated, the presence of a conceptual relationship is systematically justified by a reference to those parts of the corpus which support it. This allows us both to evaluate manually what proportion of text has been used to identify conceptual relationships, as well as to examine the passages which have not been used. To this end, we have kept in our study the division into Textual Units (TU), which roughly correspond to sentences, made by LEXTER. Out of the 4832 initial TUs, 1216 contain a concept pair linked by a linguistic pattern of one of the 14 conceptual relationships of the corpus. Note that certain "term-relationship-term" triples can be extended over several TUs, mostly due to anaphora.

A close examination of the remaining TUs has not led us to retain any relationship other than the 14 identified. We hesitated over one or two relationships (for example, "has the function of") but did not retain them in the end as there is no real linguistic pattern but only several lexical elements distributed in a paragraph : this kind of limit is also mentioned by (Davidson et al,1998).

We can thus conclude that about a quarter of the corpus contains defining elements. Recall that this text provides recommendations, that is, the more or less normative defining elements of the concepts belong to the genre of the text.

These results seem to confirm the relevance of our hypothesis : in corpora which can be characterized as belonging to the expository-genre, it is possible to use linguistic patterns in order to identify relationships.

4.2. *CTKB as a prerequisite to applications*

Our hypothesis is that a CTKB is a first model for different kinds of application, so that it can be consulted, added to or even modified by any user who needs to use a corpus to access the knowledge it contains.

In the GIS project, we have been able to evaluate the possibilities of using the CTKB that we built for two different tasks aiming to improve access to the text:

construction of a task model for writing software engineering specifications;

construction of an index for consulting the handbook.

In the first case, the contents of the CTKB were shown to be very useful. In fact, several sections of the conceptual network could be used directly to model the user's task, that is, an engineer having to write software engineering specifications. This relies on the genre of the corpus which describes the normative linking between the stages of software engineering point by point, and which are modelled in the CTKB. The description of this linking is the basis of the task modelling which has been carried out by the project's ergonomists.

In the second case, the experiment has been a little bit less conclusive. This is not surprising insofar as an index contains descriptors which are not necessarily (in fact, rarely) terms from the corpus, but rather meta-terms, which encompass several terms. Many meta-terms have been added while several C-terms have been removed.

Another experiment is currently being carried out. It sets out to formalize the results of the CTKB, so that it can be used in a knowledge base system.

Finally, an experiment should soon be started at the EDF. It involves rewriting the text so as to remove linguistic and conceptual inconsistencies (both of which have been revealed by our study).

These experiments will provide us with a better evaluation of the relevance of the CTKB, that is, the relevance of a corpus-based modeling which contains only a few formalized elements.

Conclusion

The main aim of this study was to evaluate a new form of TKB which we have called CTKB. It is application-independent and is defined as a corpus model. At the same time, we have tried to outline the stages in the method for constructing

the CTKB, focussing on identifying conceptual relationships. We have thus been faced with the problem of identifying patterns which signal these relationships in the texts.

At the present time, it would be useful to evaluate, using new corpora, the relevance of the patterns modeled during the experiment, so as to test the hypotheses of stability versus variability of linked structures in the domain and/or text genre.

The real usability of a CTKB to create a ATKB is currently being evaluated. This experiment will also allow to integrate a more advanced formalism in the CTKB in the form of proposals concerning conceptual relations made to the linguist/terminologist on the basis of calculations of completeness and coherence, for example.

The initial evaluation of the usability of the CTKB that we have made seems to suggest the relevance of this concept and the interest of the corpus-based approach that we have presented.

Acknowledgments

We wish to thank Nathalie Aussenac for her helpful comments on an earlier version of this paper.

References

Aussenac N. 1999. “GEDITERM : Un logiciel pour gérer des bases de connaissances terminologiques”. *Terminologies Nouvelles* n°19, proceedings of TIA'99, Terminologie et Intelligence Artificielle : 111-123.

Aussenac N., Bourigault D., Condamines A., Gros C. 1995. “How can Knowledge Acquisition benefit From Terminology”. *In Proceedings of 9th Knowledge Acquisition Workshop*, Banff (Canada).

Biébow B., Szulman S. 1997. “Méthodologie de création d’un noyau de base de connaissances en logique terminologique à partir de textes”. In *Actes des Deuxièmes rencontres de Terminologie et Intelligence Artificielle*, Toulouse (3-4 avril, 1997), 69-84.

Borillo A. 1997. “Exploration automatisée de textes de spécialité: repérage et identification automatique de la relation lexicale d'hyperonymie”. *LINX*, n°34-35: 113-121.

Bourigault D., Gonzalez-Mullier I, Gros C. 1996. “LEXTER, A Natural Language Processing Tool for Terminology Extraction”. *In Proceedings EURALEX'96*, Göteborg, 771-779.

Condamines A., Amsili P. 1993. “Terminology between Language and Knowledge: An example of Terminological Knowledge Base”. *In TKE’93: Terminology and Knowledge Engineering*, Frankfurt: Indeks Verlags, 316-323.

Condamines A., Rebeyrolle J. 1997. “Utilisation d'outils dans la constitution de bases de connaissances terminologiques: expérimentations, limites, définition d'une méthodologie”. In *Premières Journées Scientifiques et Techniques de l'AUPELF-UREF,* Avignon (15-16 avril 1997), 529-535.

Condamines A., Rebeyrolle J. 2000. “Construction d'une base de connaissances terminologiques à partir de textes: expérimentation et définition d'une méthode”. In *Ingénierie des connaissances*, J.Charlet, M.Zacklad, G.Kassel, D.Bourigault (eds), 225-241. Paris, Eyrolle.

Daille B., Gaussier E., Langé J. 1994. “Towards automatic extraction of monolingual and bilingual terminology”. In *Proceedings of COLING'94*, 515-521.

Enguehard C., Pantéra L. 1995. “Automatic natural acquisition of a terminology”, *Journal of Quantitative Linguistics 2 (1)*: 27-32.

Garcia D. 1997. “Structuration du lexique de la causalité et réalisation d'un outil d'aide au repérage de l'action dans les textes”. *In Actes des Deuxièmes rencontres de Terminologie et Intelligence Artificielle, Toulouse (3-4 avril, 1997),* 7-26.

Gillam L., Ahmad K. 1996. “Knowledge-engineered terminology (data)bases”. In *TKE’96: Terminology and Knowledge Engineering*, Frankfurt: Indeks-Verlag, 205-214.

Hearst M.A. 1992. “Automatic Acquisition of Hyponyms From Large Text Corpora”. In *Proccedings, 14th International Conference on Computational Linguistics*, Nantes, France, 539-545.

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

Jouis C. 1994. “Contextual approach: SEEK, a linguistic and computatioanl tool for use in knowledge acquisition”. In *University of Luxembourg, Proceedings of the first European Conference Cognitive Science in Industry*, Luxembourg (28th-30th september 1994), 259-274.

Meyer I., 2000 : “Extracting Knowledge –rich Contexts for Terminography : A Conceptual and Methodological Framework”, this volume.

Meyer I., Douglas S., Bowker L., Eck K. 1992. “Towards a new generation of terminological resources: An experiment in building a terminological knowledge base”. In *Proceedings 16th International Conference on Computational Linguistics*, Nantes, 956-957.

Morin E. 1999. “Acquisition de patrons lexico-syntaxiques caractéristiques d’une relation sémantique”, *TAL (Traitement Automatique des Langues)*, 40, (1), Paris: Université Paris VII, 143-166.

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

Pery-Woodley M.-P., Rebeyrolle J. 1998. “Domain and genre in sublanguage text: definitional microtexts in three corpora”. In *First International Conference on Language Resources and Evaluation,* Grenade (28-30 Mai 1998), 987-992.

Rousselot F., Frath P., Oueslati R. 1996. “Extracting Concepts and Relations from Corpora”. In Proceedings ECAI'96, 12th European Conference on Artificial Intelligence.

Séguéla P. 1999. “Adaptation semi-automatique d'une base de marqueurs de relations sémantiques sur des corpus spécialisés”. *In Terminologies Nouvelles, proceedings of TIA'99*, *Terminologie et Intelligence Artificielle, 19*: 52-60.