The Th(IC)2 Initiative: Corpus-Based Thesaurus Construction for Indexing WWW Documents

Nathalie Aussenac-Gilles* and Didier Bourigault**

* Université Toulouse 3, Institut de Recherche en Informatique de Toulouse (IRIT)
118, route de Narbonne, 31062 TOULOUSE Cedex 4 (F) - aussenac@irit.fr

** Université Toulouse Le Mirail, Etudes et Recherches en Syntaxe et Sémantique (ERSS)
Maison de la recherche, 5, allées Antonio Machado, 31048 TOULOUSE Cedex (F)
didier.bourigault@univ-tlse2.fr

Abstract. This working paper reports on the early stages of our contribution to the Th(IC)² project, in which, together with other French research teams, we want to test and demonstrate the interest of corpus analysis methods to design domain knowledge models. The project should lead to produce a thesaurus in French about KE research. The main stages of the method that we apply to thisexeprimentare (a) setting up a corpus, (b) selecting, adapting and combining the use of relevant NLP tools, (c) interpreting and validating their results, from which terms, lexical relations or classes are extracted, and finally (d) structuring them into a semantic network. We present the LEXTER system used to automatically extract from a corpus a list of term candidates that could later be considered as descriptors. We also comments upon the validation protocol that we set up : it relies on an interface via the Internet and on the involvement of the French KE community.

1 The Th(IC)2 Initiative

1.1 A contribution to the (KA)2 initiative

The Th(IC)2 project is an initiative from of the French TIA[1] group of interest. With this project, some French researchers in Knowledge Engineering (KE) intend to contribute to the (KA)2[2] project [4]. Initiated in 1998, the (KA)2 initiative aims at building an ontology that would be used by researchers in the domain of KE in order to index their own web pages with "semantic tags " corresponding to concepts in this ontology. In its current state, the (KA)2 ontology contains the knowledge necessary to describe the administrative organisation of the research in the field, but few items related to the content of the research itself. The target of the Th(IC)2 contribution is to enrich the part of (KA)2 ontology dedicated to the description of research topics in the KE community.

---

[1] TheTIA special interest group (http://www.biomath.jussieu.fr/TIA/) is a research group in Linguistics, NLP and AI concerned with text-based acquisition of ontological and terminological resources. The authors, as well as the members of the TIA group, thank the " Direction Générale à la Langue Française" (DGLF) for supporting the Th(IC)2 project.

[2] http://www.aifb.uni-karlsruhe.de/WBS/broker/KA2.html

With a larger scope, our methodological proposals can prove relevant in the broader context of designing community web portals.

The first purpose of the Th(IC)2 project is to build a thesaurus in French which will describe how KE research develops in the French speaking area, with its specificity and strengths. Indeed, we will first draw a state-of-the-art on research topics that are currently addressed in the French KE community. This must be done before it can be included into a broader description. This thesaurus will have a conventional structure: a set of descriptors referring to research topics will be organised in taxonomy and connected via synonymy and "see also" links. The correspondence between this thesaurus and the (KA)2 formal ontology will be established in a second stage.

1.2 Using corpus-based methods to build a thesaurus

The overall process proposed by the promoters of the (KA)2 project is to use tools, methods and languages developed by the Knowledge Acquisition (KA) community in order to build the ontology. This recursive prerequisite explains the square in (KA)2. In the same spirit, the TIA group wants to test and demonstrate the interest of some KE results, and particularly those resorting to corpus analysis methods. A new trend appeared recently, derived from a major evolution of Terminology [3]. It resorts to both acquisition tools based on linguistics and browsing and modelling tools with links between models and texts. This evolution is due to new corpus-oriented Natural Language Processing (NLP) tools whose efficiency has increased thanks to valuable collaborations between linguists, terminologists and knowledge engineers. This trend is clearly less ambitious than the automatic transfer approach: NLP tools are viewed as aids for knowledge engineers who select and combine their results to develop a model.

The tools and methods developed by the TIA group members should be useful for ontology design. We assume that a thesaurus is a kind of lexico-conceptual resource similar to ontologies, at least enough to resort to the same corpus-based techniques to design them. Comparing thesaurus and ontologies is one of the issues that could be made clearer thanks to this project.

This paper describes one of the experiments carried out within the TIA group to build this thesaurus. The group set up a general method [3] that defines a framework within which knowledge engineers select and adapt the relevant tools for the application at hand, according to the documents and expertise available, the corpus language and the kind of resources to build. The main stages of this method are (a) setting up a corpus, (b) selecting, adapting and combining the use of the relevant NLP tools, (c) interpreting and validating their results, from which terms, lexical relations or classes are extracted, and finally (d) structuring them into a semantic network. This working paper reports on a particular experiment that illustrates most of these stages:

1. A first corpus as representative as possible of research activities within the French speaking KE community is set up (section 2).
2. The LEXTER system is used to automatically extract from this corpus a list of term candidates that could later be considered as descriptors (section 3).
3. A validation protocol is defined: the single term list is automatically subdivided into sub-lists according to the number of texts comprised in the original corpus; these sub-lists are validated through an interface via the Internet (section 4).

Further stages (section 5) include the selection of terms and their organisation into a thesaurus that is then structured with the help of additional tools. Finally, the French KE community will be asked to validate the whole.

2 Building a reference corpus

The TIA group used all available criteria to set up an exhaustive and representative corpus. To this end, the corpus gathers many documents produced in the domain and distributed as follows:

- 32 descriptions of laboratories or teams working in the field of KE *("AFIA sub-corpus")* published in a special report on KE in the 34th issue of the *Bulletin de l'Association Française d'Intelligence Artificielle.* Each description (of an average size of 975 words) shortly outlines the main directions of investigation of a team or laboratory, its main results, collaborations and publications.
- 35 papers of a recently edited book on KE *("LIVRIC sub-corpus")* [8]. This book collects a selection of papers from the proceedings of the French conference in KE (IC) that were organised between 1995 and 1998. The average size of the papers is 5095 words. Most of the topics addressed by research in KE at this time are quite well represented.

	AFIA sub-corpus.	LIVRIC sub-corpus
Document type	Laboratory descriptions	Scientific papers
Number of documents	32	35
Average number of words per document	975	5 095
Total number of words	31 212	178 336

Table 1: Some figures about the reference corpus of the Th(IC)2 project

3 Extracting term candidates with LEXTER

A preliminary selection of terms is performed using LEXTER, which is a term extractor [6] [7]. The input of LEXTER is an unambiguously tagged corpus. The output is a network of term candidates, that is words or sequences of words which are likely to be chosen as entries in a thesaurus or concept labels in an ontology. The extraction process is composed of two main steps.

1. Shallow parsing techniques implemented in the Splitting module detect morpho-syntactical patterns that cannot be parts of terminological noun phrases, and that are therefore likely to indicate noun phrase boundaries. In order to process correctly some problematic splitting, such as coordinations, attributive past participles and ambiguous preposition + determiner sequences, the system acquires and uses corpus-based selection restrictions of adjectives and nouns.
2. Ultimately, the Splitting module produces a set of text sequences, mostly noun phrases, which we refer to as Maximal-Length Noun Phrases (henceforth MLNP). The Parsing module recursively decomposes the maximal-length noun phrases MLNP into two syntactic constituents: a constituent in head-position (e.g. 'model' in the noun phrase 'conceptual model'), and a constituent in expansion position (e.g. 'conceptual' in the noun phrase 'conceptual model'). The

Parsing module exploits rules in order to extract two subgroups from each MLNP, one in head-position and the other one in expansion position. Most of MLNP sequences are ambiguous. Two (or more) binary decompositions may compete, corresponding to several possibilities of prepositional phrase or adjective attachment. Disambiguation is performed by a corpus-based method that relies on endogenous learning procedures.

Term candidate	freq.	Term candidate	freq.	Term candidate	freq
modèle conceptuel *conceptual model*	135	type de connaissance *knowledge type*	38	espace de connaissances *knowledge space*	25
résolution de problème *problem solving*	121	méthode de résolution de problème *problem solving method*	37	domaine d'application *application domain*	25
ingénierie de la connaissance *Knowledge engineering*	120	travail coopératif *co-operative work*	37	système expert *expert system*	25
acquisition des connaissances *knowledge acquisition*	106	représentation de la connaissance *knowledge representation*	36	Base de Connaissance *Knowledge base*	24
système d'information *information system*	106	gestion de la connaissance *knowledge management*	33	système informatique *compute supported system*	24
connaissance du domaine *domain knowledge*	92	fouille de donnée *data mining*	33	langage de représentation *representation language*	23
candidat terme *term candidate*	63	niveau d'abstraction *abstraction level*	33	unité linguistique *linguistic unit*	23
système à base de connaissances *knowledge based system*	56	contexte partagé *shared context*	32	relation sémantique *semantic relation*	23
génie logiciel *software engineering*	55	langage de modélisation *modelling language*	32	premier temps *first stage*	23
modélisation de la connaissance *knowledge modelling*	50	méthode de résolution *problem solving method*	32	haut niveau *high level*	22
base de données *data base*	47	ontologie de l'expertise *expertise ontology*	32	base de cas *case base*	22
logique de description *description logic*	46	acquisition de connaissances *knowledge acquisition*	31	modèle de connaissances *knowledge model*	22
aide à la Décision *computer supported decision making*	46	appel d'offre *call for proposal*	29	système coopératif *co-operative system*	22
modèle d'expertise *expertise model*	45	processus de conception *design process*	29	processus d'acquisition *acquisition process*	22
structure prédicative *predicative structure*	44	mémoire d'entreprise *corporate memory*	28	primitive de modélisation *modelling primitive*	21
points de vue *point of view*	43	mot clé *key word*	28	dossier médical *medical file*	20
ingénieur de la connaissance *knowledge engineer*	41	fonction test *test function*	27	relation causale *causal relation*	20
mesure de similarité *similarity mesure*	39	Management par projet *Project management*	27	primitive conceptuelle *conceptual primitive*	20
modèle générique *generic model*	39	modèle de raisonnement *reasoning model*	27	niveau connaissance *knowledge level*	20
graphe conceptuel *conceptuel graphs*	38	cycle de vie *life cycle*	26	type de document *document type*	20

Table 2: The most frequent term candidates in the Th(IC)2 corpus

The sub-groups generated by the Parsing module, together with the MLNP extracted by the Splitting module, are the term candidates produced by LEXTER. This set of term candidates is represented as a network: each multi-word term candidate is connected to its head constituent and to its expansion constituent by syntactic decomposition links. Building the network is especially important for the purpose of term acquisition.

LEXTER was used in many applications aiming at gathering lexical and/or conceptual resources, such as terminological knowledge bases, ontologies, thesaurus, etc. [6], [1].

In this experiment, the number of term candidates extracted by LEXTER from the Th(IC)2 corpus is given in table 3 and the most frequent term candidates are listed in table 2.

	freq = 1	freq > 1	Total
Number of term candidates	17189	3879	21068

Table 3 : Number of term candidates extracted by LEXTER from the Th(IC)2 corpus

4 Evaluation protocol

4.1 Generating sub-lists of term candidates for individual validation

The most frequent term candidates appear to be relevant descriptors, and thus must be considered as valid entries in the thesaurus. However, this simple numeric criterion is not powerful enough to select without any error or omission a set of descriptors that will cover the whole range of research activities in KE in a precise and exhaustive manner. Some term candidates with a low frequency should be considered. So the validation process should bear on the entire list of extracted term candidates.

Given the very large size of this list, it is hard to imagine that a small number of persons would undertake the validation of the entire list. It is doubtful that such a group has the competence and time required to check the whole domain and corpus. Moreover this thesaurus will not be used to massively index large document bases, but rather as a precise map of the KE domain intended as a reference documents for researchers. This is why we have set a collective and manual validation process: we ask every researcher to validate the term candidates extracted from his/her own texts.

In order to make this individual validation possible, we have decomposed the list of term candidates into as many sub-lists as documents in the corpus.

- For each document in the LIVRIC sub-corpus, we have selected those candidate terms occurring at least twice in the document, or only once in the document and at least once in an other document from the LIVRIC sub-corpus. The average number of term candidates of the sub-lists is 81.
- For each document in the AFIA sub-corpus, we have selected those candidate terms occurring at least twice in the document. The average number of term candidates of the sub-lists is 48.

This validation protocol requires involving all the researchers concerned as authors. We consider this participation as very beneficial. Firstly, it is a very enriching experiment for an author: he has a picture of his document in a form both unusual for him and familiar

enough to be interpreted. Secondly, we assume that, in line with the (KA)2 project promoters, the success of an experiment like the Th(IC)2 project strongly depends on the important involvement of the community members. They should not be only users of the thesaurus, but they should take part in the early stages of its design *("Do not ask what the community can do for you. Ask what you can do for the community!").*

4.2 A validation interface on the web

To implement this collaborative validation process, we designed a web interface through which the authors can access and validate the sub-list of term candidates built up from their text. A snapshot of the validation interface is given on figure 1.

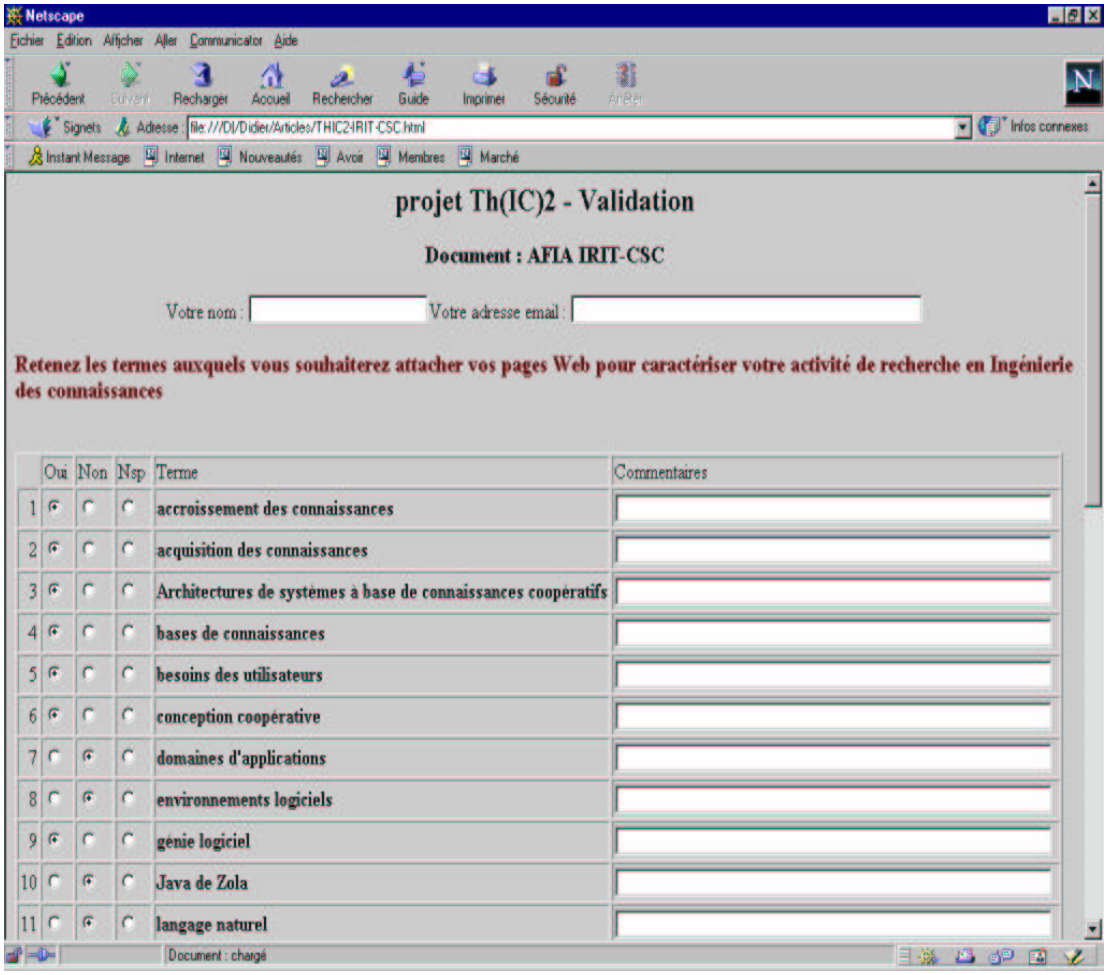

Figure 1 : A snapshot of the validation interface.

At this stage, the main difficulty was to formulate precise validation procedures so that any author would validate the list of term candidates “in the same spirit”. We led many

experiments in which specialists were asked to validate lists of term candidates. One of the main lessons learned from these experiments is that decision making is heavily dependent on the goal of the task, that is the type of lexical and/or conceptual resource that is under concern. Roughly speaking, with the same starting list of term candidates, the set of selected terms will not be the same whether the validated terms are to be integrated as descriptors in a thesaurus used by an automatic indexing system, or as concept labels in an ontology used by a knowledge-based system. For this reason, we will first explain the authors what the main goal of the Th(IC)2 project is (that is building a thesaurus for the KE community). We will then ask them not to index the document from which term candidates were extracted but to select term candidates according to their relevance and usefulness to characterise their own research within the field of KE.

5 Further stages

5.1 Expertise based cross validation by the community

The next step planned in the project is to launch the validation process by soliciting members of the teams described in the AFIA sub-corpus and authors of papers of the LIVRIC sub-corpus. We will then synthesise all the results and build an initial list of descriptors by gathering all the term candidates that were selected by at least one author. During this stage, we will also have to gather synonym terms, to add simple terms that could help organise more complex ones, and to get to a consensual view by comparing the various lists. The resulting list will serve as a bootstrap for further work.

5.2 From term lists to a thesaurus

The main task will then be to structure this list with conventional thesaurus links [3]. This task will rely on two main approaches, carried out in parallel:

- A corpus based bottom-up analysis using results of natural language processing tools such as term extractors, relation extractors, clustering tools… . Links may indeed be revealed by term use in texts. Good means to identify these links may be either to browse term occurrences, which may be costly, or to look for co-occurrent terms, or to extract lexical relationships. We plan to use a lexical relation extractor, Caméléon [9], to check the links related to the selected terms, and to explore domain specific relations. By this means, additional information will be available to decide how to organise the thesaurus descriptors into a taxonomy. Lexical relations are also good inputs to precisely describe domain concepts. This analysis is likely to provide the lower level layers of the thesaurus.
- A top-down approach based on our expertise in the domain's global organisation as a research field. In short, this will lead to define the high level layers of the thesaurus which will organise the lower level layers previously mentioned. More precisely, expertise is very useful to directly get to the right interpretation of any textual item and to avoid further text investigations. It is also likely to shortcut some references to texts when trying to differentiate descriptors one with another. Moreover, most of the high level structuring descriptors are not in texts and must be acquired from domain experts. Although this process may seem very

pragmatic and intuitive, our goal is to make explicit the more modelling rules as possible.

A modelling tool, such as Terminae [5] or Géditerm [2], will help to store, to browse, to structure and to describe the terms, their relations and their definitions.

6 Discussion

Beyond the possible contribution to the (KA)2 project, this experiment raises two major issues for the KE community:

- What are the qualitative and quantitative benefits (in terms of design cost and time, domain coverage, quality of the final knowledge structure…) of a corpus-analysis-based approach?
- What are the right structuring and formalisation levels for an efficient indexing of researchers web pages? Is it worth undertaking the design of a formal ontology with very well-defined links or is a thesaurus enough?

References

1. Assadi H. (1998) Construction of a regional ontology from texts and its use within a documentary system. In *Formal ontology in information system* (FOIS'98), Guarino N. Ed. Frontiers in Artificial Intelligence and Applications, Vol. 46, pp 236-249. IOS Press, Amsterdam

2. Aussenac-Gilles N. (1999), GEDITERM, un logiciel de gestion de bases de connaissances terminologiques, in *Terminologies Nouvelles* n°19, pp 111-123. 1999

3. Aussenac-Gilles N., Biébow B. & Szulman S. (2000) Revisiting Ontology Design: a methodology based on corpus analysis, in *Proc. of the 12th Int. Conference on Knowledge Engineering and Knowledge Management* (EKAW'2000), Juan-les-Pins (F), Oct. 2000

4. Benjamins R., Fensel D., and Decker S. (1999) (KA)² : Building ontologies for the Internet: A Midterm Report. *International Journal of Human Computer Studies*, 51(3):687

5. Biebow B., Szulman S. (1999) TERMINAE : A linguistic-based tool for the building of a domain ontology, in *proc. of the 11th European Workshop on Knowledge Acquisition, Modelling and Management* (EKAW'99), Dagstuhl Castle, Germany, pp 49-66. 1999.

6. Bourigault D. (1995). LEXTER, a Terminology Extraction Software for Knowledge Acquisition from Texts, In *Proceedings of the 9th Knowledge Acquisition for Knowledge Based System Workshop* (KAW'95). Banff, Canada.

7. Bourigault D., Gonzalez-Mullier I. & Gros C. (1996) LEXTER, a Natural Language Tool for Terminology Extraction, In *Proceedings of the 7th EURALEX International Congres*s, Göteborg, Sweden. pp771-779

8. Charlet J., Zacklad M., Kassel & Bourigault D. (2000), *Ingénierie des Connaissances : évolutions récentes et nouveaux défis*, Paris: Eyrolles

9. Séguéla P., Aussenac-Gilles N. (1999), Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine, *Actes de IC'99 (Conférence Française d'Ingénierie des Connaissances*), pp 79-88, Paris, 1999.