Corpus analysis for conceptual modelling

Nathalie Aussenac-Gilles *, Brigitte Biébow **, and Sylvie Szulman**

* Université Toulouse 3, Institut de Recherche en Informatique de Toulouse(IRIT)
118 route de Narbonne, 31062 TOULOUSE Cedex 4
Nathalie.Aussenac@irit.fr
**Université de Paris-Nord, Laboratoire d'Informatique de Paris-Nord(LIPN)
Av. J.B. Clément, 93430 VILLETANEUSE (France)
{Brigitte.Biebow,Sylvie.Szulman}@lipn.univ-paris13.fr

Abstract We promote a new approach for knowledge modelling based on knowledge elicitation from technical documents. It benefits of the increasing amount of available electronic texts and of the maturity of natural language processing tools. The approach defines a framework where the knowledge engineer selects the appropriate tools, combines their use and interprets their results to build up a domain model. The paper presents the method.

1 Introduction

Within the modelling cycle, texts have long been hardly used. We claim that, when available, documents or any natural language support (messages, text files, paper books, technical manuals, notes, protocol transcripts, etc.) are one of the possible forms the knowledge may take. Each of these supports will provide different types of knowledge to be elicited and exploited with specific techniques. The general approach described in this paper is keeping the principles of the French TIA (Terminology and Artificial Intelligence) group. It involves researchers from from KE, Terminology and Linguistics communities. Its major statements are the following:
(1) to start from texts to acquire knowledge: texts are a tangible support, collecting stabilized knowledge which may be referred to in the model; unlike individual expertise, texts hardly contain very specific and practical know-how acquired through experience. Indeed, they reflect a consensual view on the domain. This might be an advantage, or a useful starter, especially for applications that address a large variety of users. However, it does not mean that texts will be the single knowledge source.
(2) to connect source texts to conceptual models: relevant connections from concepts to the texts where they are defined or used in improve the model interpretation. Labels play a larger role that is hardly acknowledged. They help the reader to understand concept meanings in the domain (referential interpretation) and their representation in the model. Such connections also guarantee the model understanding and maintenance by keeping tracks of modelling choices.
(3) to explore texts by applying natural language processing tools and

techniques based on results in linguistics: these tools help systematic text analysis and make the modelling process easier. We do not promote here fully automated text interpretation. Current investigations tend to organize the application of such tools into efficient methods dedicated to specific application types.

We present in this article our method for domain knowledge elicitation and modelling from texts by analysing corpora with NLP tools.

2 Methodological framework

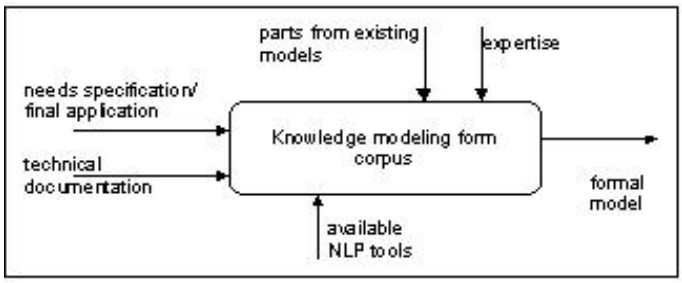

Figure 1. Global view of the method

Our method is a general one, independent of the language used in texts. It defines a framework where some technical and methodological choices are left to the knowledge engineer, depending on different factors:

- the application requirements for which the conceptual model is developed;
- the technical documents that are put at disposal;
- the elements of existing models that could be reused (ontologies or terminological resources as thesaurus, glossaries etc.);
- the expertise that can be given;
- the natural language processing tools that are available, which are not the same depending on the processed language in the texts.

The designer who uses the method is more or less qualified in linguistics, in knowledge modelling and in formalization. At every step, he/she must decide which techniques must be used depending on the previous factors and his/her own ability. To use the method, he/she needs a specific software to manage a great amount of information (terms, concepts and relations), to describe it, to organize it and to formally represent it. Such an environment must allow an easy access to the terms and the lexical relations, the texts from which they are extracted and the model in which they will be inserted.

We have already developed Terminae [5] and Géditerm [2] for this purpose. Terminae offers to consult a corpus and to integrate the results of the Lexter extractor of term-candidates [1]. The designer extracts terms from the list of term-

[1] A term-candidate is a syntagm extracted from texts that may become a term if validated by an expert.

candidates and defines notions from term meanings. In Terminae, a notion refers to a concept under modelling, whereas a concept is formal. These notions are then structured and differentiated, to be finally formalized as concepts (keeping the link to the corpus). In Terminae, the formal language is close to terminological logics. The link is kept between each formal concept, the notion, its associated terms and their occurrences in the corpus. Géditerm assists the first steps to select terms. The term/concept link is justified by the occurrences of the terms in the corpus. Terminae and Géditerm both may take as input the list of term-candidates given by Lexter. Géditerm does not allow the resulting conceptual network to be formalized but it provides tools for a better management and visualization of the conceptual network before its formalization.
In what follows the methodological framework is presented, focusing firstly on the nature of the data used and produced during the process, from the corpus to a domain model. Then the main stages of the process itself are described.

2.1 From texts to a formal model

The method applies on technical documentation and ends to a formal domain model. Terms and lexical relations are syntagms existing in the corpus and regarded as important in the domain. Lexical clustering puts together syntagms which occur in some similar contexts. The syntagms may be interpreted in a local context (sentence or paragraph) then in a global one (text or whole corpus). If they are considered as terms, they give rise to concepts and semantic relations that they label. The set of concepts and relations make up a semantic network, informal but understandable by the designer. In the formal model, concepts and relations are formalized into Terminae terminological language: concepts and roles are structured into an inheritance hierarchy. The concepts are characterized following two dimensions, a linguistic one to express how close to a syntagm in the corpus a concept is, and a pragmatic one, reflecting the reasons why the concept has been integrated into the formal model. This information makes both the model and the knowledge base easier to understand and to maintain [4].

2.2 Used natural language processing tools

A full study of the linguistic tools that could be used for knowledge elicitation and structuring is still to do, and the following list is not exhaustive. We have selected the most frequently used tools in the French TIA community. We differentiate tools dedicated to terminological knowledge acquisition (TKA) from texts (most of them work on French only), from other TKA software that specialized in conceptual modelling, and from classic linguistic tools for NLP.

Terminological extractors Term-candidate extractors extract from a corpus a list of terms that must be validated. They return a great amount of data often with some noise; so a long and boring selection must be made that requires both a good domain expertise and a good anticipation of the way terms will be used.

These tools can be based on syntactic principles, as Lexter [6] and Nomino [9], or on statistic principles as Ana [10] and Startex [11]. Their use does not imply a great competence in linguistics.

Relation extractors are usually based on linguistic patterns such as Prométhée [15] or Caméléon [17]. Some of these tools need first to be provided with general linguistic relation patterns like "X IS INDEFINITE_ARTICLE Y" for the hyperonymy relation (kindOf). Patterns are applied on the corpus in order to visualize the pieces of texts where the lexical relation appears. Other tools require couples of related forms as input, from which specific patterns are identified. Starting from some predefined patterns their application onto the corpus rises up terms from which domain specific patterns may be created for new lexical relations. The use of these tools requires some linguistic skills but gives significant information for structuring the domain.

Term and relation extractors may be used in a separate or complementary way. If a term extractor is firstly used, then relations between terms may be searched for by exploring their contexts. If a relation extractor is firstly used, then projecting the relations onto the corpus may rise up related terms. These tools usually offer an environment to browse their results.

Other terminological tools Some TKA tools are more oriented towards concept discovery. Conceptual clustering tools like Zellig [13] or Lexiclass [1] put together noun phrases that share syntactic dependency relations. The resulting clusters must be manually analyzed to define semantic classes. Results interpretation is difficult but term structuring and concept definition is made easier. Asium [12] uses learning techniques to propose term clusters. Each cluster must be manually validated before defining concepts. Synoterm [18] offers potentially synonym clusters, that can be also considered as concept-candidates. Lexis [16] finds names in a corpus, which may be useful to find some class instances.

An example of sophisticated acquisition tool working in English is KAWB (Knowledge Acquisition WorkBench) [14]. It acquires some semantic classes of a domain from large text corpora. It uses various methods from computational linguistics, information retrieval and KE. A data extraction module includes word class identification based on linguistic annotation of texts, statistical word clustering, with access to external linguistic and semantic sources. A pattern finder collects word collocations, searches for regularities and proposes lexico-semantic patterns for a conceptual characterization to the user. An analysis and refinement module helps the user to test patterns which represent his/her hypotheses, groups together the cases and generalizes them to ask the user for a final decision.

Classic linguistic tools Some simple and very easy to use linguistic tools have been available for many years now, like concordancers and KWIC tools. KWIC (KeyWord In Context) tools bring into vertical alignment along a given word or phrase all the sentences of a corpus in which this word occurs. This is very practical to study all its contexts, its linguistic behaviour and, first of all, to get

an idea of its meaning from the way it is used. Concordancers offer a similar assistance: they look into the corpus for every occurrence of any user given syntagm. They are more powerful than KWIC tools because these syntagms may be characterized by syntactic or semantic properties, not only by giving explicit nouns phrases or verbs. So concordancers result very practical to apply and test some patterns on a corpus, study their occurring contexts and compare them.

Generic tools for text analysis, such as Sato [8] offer a variety of options, which range from research of occurrences and text alignment to syntactic analysis and corpus tagging, including statistics on word frequency, disambiguation at a syntactic or sometimes semantic level. They may be useful for extracting and structuring knowledge when looking for very specific information.

2.3 Detailed description of the method steps

The modelling process is detailed below from setting up the corpus along to designing the formal model.

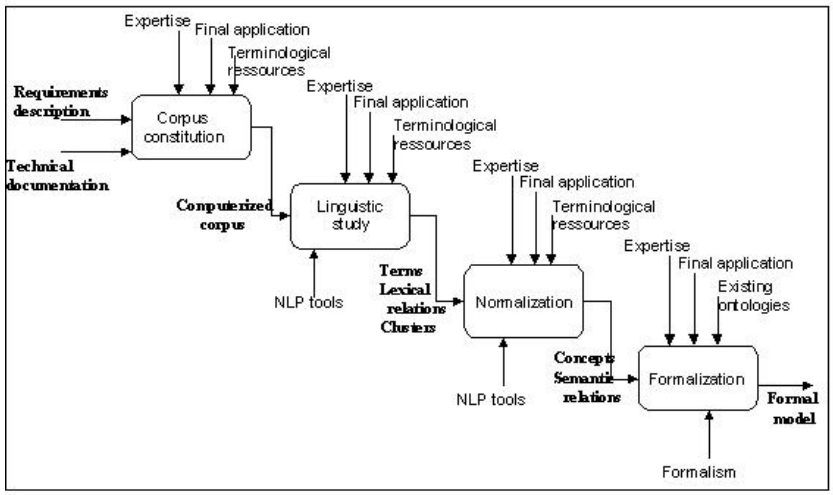

Figure 2. Steps of the modelling process from text according to our approach

Setting up the corpus From the requirements that explain the objectives underlying the model development, the designer selects texts among the available technical documentation. He must be an expert about texts in this domain to characterize their type and their content. The corpus has to cover the entire domain specified by the application. A glossary, if it exists, is useful to determine sub-domains and to verify that they are well covered. The corpus is then

digitalized if it was not. Beginning the modelling may lead to reconsider the corpus.

Linguistic analysis This step consists in selecting adequate linguistic tools and techniques and in applying them to the text. Their results are sifted and a first linguistic based elicitation is made. The objective is to allow the selection of the terms and lexical relations that will be modelled. The results of this stage are quite raw and will be further refined.

Normalization Normalization is a particular conceptualization process based on corpus analysis, in line with [7] in contrast with expert introspection. The expertise and the target system influence concept definitions in a second time. Indeed, the restricted meaning of concepts is mainly derived from the study of term occurrences in texts. These terms become concept labels. Thus concepts are described thanks to the use of their label together with the other terms in the corpus. So, the corpus plays an important role during normalization. Linguistic study and normalization are closely intertwined and cyclic activities. At any time of the normalization process, we use linguistic tools and principles to explore the text and to decide whether a concept, an attribute or a relation should be defined or not.

Normalization includes two parts: the first one is still linguistic, it refines the previous lexical results; the second one concerns the semantic interpretation to structure concepts and semantic relations. The modelling goes from terminological analysis to conceptual analysis, that means from term to concepts and from lexical relations to semantic ones. During normalization, the amount of data to be studied is gradually restricted.

During the linguistic step, among the set of terms and lexical relations, the designer has to choose those that will be modelled. This choice is mainly subjective, the terms and relations are kept when they seem important both in the domain and for the application. Because of this subjectivity the selection is rather large. Then, from the study of each syntagm occurrences, the designer writes in natural language a definition that remains close to the texts. In the same time, he determines for each term and relation if it has one or several meanings in the domain. In case of polysemy, he decides which meanings attested by the corpus have to be kept because they are relevant.

The second step is conceptual modelling. Concepts and semantic relations are defined in a normalized form using the labels of the concepts and relations already defined. These definitions may be less close to the text as long as they must be relevant for the task for which the model is built. These descriptions are structured into a semantic network, with a strong emphasis on the hierarchical relations (kindOf, partOf). Only the rigor of the work and perhaps the modelling environment may guarantee the coherence of this semi-formal ontology.

Formalization The formalization step includes building and validating the ontology. Some existing ontologies may help to build the highest levels and to structure it into large sub-domains. Then semantic concepts and relations are translated into formal concepts and roles and inserted in the ontology. This may imply to restructure the ontology or to define additional concepts, so that the inheritance constraints on the subsumption links are correct. Inserting a new concept triggers a local verification to guarantee the syntactic validity of the added description. A global validation of the formal model is performed once the ontology reaches a quite stable state to verify its consistency. This step and its consequences on the validation of the ontology have been detailed in [4]

3 Conclusion

In this article we have presented a method to create a domain conceptual model from a corpus analysis, by using NLP tools. A report on the early stage of an experiment where we applied it to organize the concepts of the Knowledge Engineering domain in French may be found in the proceedings of the conference EKAW 2000 ([3]). This method raises several methodological issues that could be worth debated during the workshop :

- What are the good criteria to set up an appropriate corpus? Further investigations still need to be done to list explicit criteria for text selection, and to measure how much the nature of texts influences the kind of analysis process to be carried out.
- What are the available NLP tools? A listing of existing tools processing other languages than French could be useful. We have mentionned only a few of them.
- Who should be in charge of result interpretation? It requires specific abilities and training, both in linguistics and knowledge engineering. But the method should also bring general principles and guidelines. We have proposed some guidelines that should be extended when new tools will be taken into account.
- How much does the type of the final application influence the process and the kind of tools to be used ? An on going experiment will help us specify those relevant to design a thesaurus for indexing documents. Another one, based on the same corpus, aims at designing an ontology and will identify a proper way to design them from texts. Comparing the two experiments will underline the influence of the application type on the process.

References

1. H. Assadi. Construction of a regional ontology from text and its use within a documentary system. In N. Guarino, editor, *Proc. of the 1st International Conference on Formal Ontology and Information System (FOIS'98))*, pages 236–249, 1998.
2. N. Aussenac-Gilles. Gediterm, un logiciel de gestion de bases de connaissances terminologiques. *Terminologies Nouvelles*, 19:111–123, 1999.

3. N. Aussenac-Gilles, B. Biébow, and S. Szulman. Revisiting ontology design : a methodology based on corpus analysis. In *Proc. of the 12th International Conference on Knowledge Engineering and Knowledge Management (EKAW'2000)*, LNAI. Springer-Verlag, 2000.
4. B. Biébow and S. Szulman. Terminae: A linguistics-based tool for building of a domain ontology. In D. Fensel and R. Studer, editors, *Proc. of the 11th European Workshop (EKAW'99)*, LNAI 1621, pages 49–66. Springer-Verlag, 1999.
5. B. Biébow and S. Szulman. Terminae: une approche terminologique pour la construction d'ontologies du domaine à partir de textes. In *Proc. of Reconnaissance des Formes et Intelligence Artificielle (RFIA'2000)*, volume II, pages 81–90, 2000.
6. D. Bourigault. Lexter, a natural language processing tool for terminology extraction. In *Proc. of the 7th EURALEX International Congress*, Goteborg, 1996.
7. J. Charlet and B. Bachimont. De l'acquisition l'ingénierie des connaissances: Applications et perspectives. In *Actes des Assises Nationales 1998 du PRC-I3*, http://www.irit.fr/ACTIVITES/EQ_SMI/GRACQ/index-commf.html, 1998.
8. F. Daoust. *Système d'Analyse de Textes par Ordinateur.* Centre ATO, Université du Québec à Montréal, 1992.
9. S. David and P. Plante. *Termino version 1.0.* Centre d'Analyse de Textes par Ordinateur, Université du Québec à Montréal, 1990.
10. C. Enguehard and L. Pantéra. Automatic natural acquisition of terminology. *Journal of Quantitative Linguistics*, 2/1:27–32, 1995.
11. P. Frath F. Rousselot and R. Oueslati. Extracting concepts and relations from corpora. In *Proc. of the 12th European Conference on Artificial Intelligence (ECAI'96)*, 1996.
12. D. Faure and C. Nedellec. Knowledge acquisition of predicate argument structures from technical texts using machine learning: The system asium. In D. Fensel and R. Studer, editors, *Proc. of the 11th European Workshop (EKAW'99)*, LNAI 1621, pages 329–334. Springer-Verlag, 1999.
13. B. Habert, E. Naulleau, and A. Nazarenko. Symbolic word clustering formedium-size corpora. In *Proc. of the 16th International Conference on Computational Linguistics (COLING'96)*, pages 490–495, Copenhagen, 1996.
14. A. Mikheev and S. Finch. A workbench for acquisition of ontological knowledge from natural language. In *Proc. of the 9th Banff Knowledge Acquisition for Knowledge-Based Systems Workshop (KAW'95)*, 1995.
15. E. Morin. Acquisition de patrons lexico-syntaxiques caractéristiques d'une relation sémantique. *TAL (Traitement Automatique des Langues)*, 40/1:143–166, 1999.
16. T. Poibeau. Repérage des entités nommées: un enjeu pour les système de veille. *Terminologies Nouvelles*, 19:43–51, 1999.
17. P. Séguéla. Adaptation semi-automatique d'une base de marqueurs de relations sémantiques sur des corpus spécialisés. *Terminologies Nouvelles*, 19:52–60, 1999.
18. A. Nazareko T. Hamon and C. Gros. A step towards the detection of semantic variants of terms in technical documents. In *Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98))*, pages 498–504. Morgan Kaufmann, 1998.