

# Introduction

## Application-driven terminology engineering

Maria-Teresa Cabré, Anne Condamines and  
Fidelia Ibekwe-SanJuan

### 1. Introduction

Following a two day workshop, held in Lyon (France) in 2004, on “Terminology, Ontology and Knowledge Representation”, this special issue aims at examining how current research in terminology engineering is influenced by particular target applications. The Lyon workshop showed that investigators from many disciplines (linguistics, information retrieval, knowledge management, philosophy, computational linguistics) use the notion of “term” to refer to different text units with different morphological and syntactic properties. For the majority of authors in the information retrieval field, terms are synonymous to single word units. Indeed, document vectors are often single word units which are used to compute index terms for query expansion. At the other extreme, for certain applications of domain knowledge modelling, the units of interest can be whole clauses like the “units of understanding” referred to in Kerremans and Temmerman (2004).

The majority of researchers concerned with automatic term extraction consider terms as short phrases, mostly noun phrases (Bourigault 2002; Daille 1996; David and Plante 1990; Jacquemin 2001; Justeson and Katz 1995; Lauriston 1994); some authors extract other syntactic constructs, e.g., verb phrases (Bourigault 2002). Although terms can be single word units, empirical studies have shown that, in a specialised field, most terms are multi-word units.

This diversity of approaches to the notion of “term” is determined by the needs of specific applications. But needs must not lead to confusion about the nature of terms. Terminological units constitute the central object of the

knowledge field of terminology and they must therefore be clearly delimited. In terminology, a terminological unit is a lexical unit with a morphological or a syntactic structure which corresponds to a minimal autonomous conceptual unit in a given field (Cabré 2003). This conception does not exclude the use of larger or smaller units in a given application.

In order to solve this terminological problem, some authors have introduced other terms, e.g., *unit of understanding* (Temmerman 2000) or *specialised knowledge unit* (Cabré 2003). Beside morphological, phraseological and sentence units, terminological units correspond to a particular type of specialised knowledge unit, all of them are always defined inside a precise knowledge field. Consequently, every application needs different specialised knowledge units for its purposes and selects them according to this purpose. This justifies the diversity of uses that *term* or *terminological unit* have in applied terminology, a diversity clearly shown by the contributions at this workshop.

This special issue examines how an application (the end use to which the processed terms are destined) influences the choice of the text units and their processing. In such an application-oriented framework, *term* designates the meaningful text unit in specialised discourse considered useful for an application. This approach can be expressed more simply as, “the application justifies the notion of ‘term’ and their choice”.

### *What is an application?*

It seems necessary to define the notion of “application” in the title “Application-driven Terminology Engineering” chosen for this special issue, since it lends itself to multiple interpretations. In its broadest sense, an application is conceived as the end use to which the results of a study are put. Such a study may use different methodologies, including automation, in its different stages.

Where analysis is restricted only to fields closely related to terminology (linguistics, computational linguistics and artificial intelligence) and to studies involving the use of computer programs, “application” is usually understood to be the end target involving real users and real needs. Taken in this strict sense, possible applications of terminology engineering could be information retrieval, question-answering, information extraction, text mining, machine translation, science and technology watch, all are external fields that can benefit from the incorporation of terminological knowledge. However, a looser acceptance of the term is equally widespread. Researchers involved in building language resources (dictionaries, lexicons, parallel or aligned corpora) or in modelling

domain knowledge (semantic relation identification, thesaurus, taxonomy, ontology building) also consider these objects as the “application” to which their studies are destined. In this sense, the application is defined according to the use to which the extracted relations, the dictionary, lexicon, thesaurus or ontology are put. Thus the second type of “applications” can be perceived as “intermediary” and as serving the first.

While the automated or semi-automated building of these resources requires the use of complex methodologies (symbolic, statistical, machine learning), the studies may not actually test the resources in the framework of an end application with real users. The diversity of papers presented in this special issue clearly illustrates the polysemous nature of the concept “application”. Most of the seven papers in this issue are concerned with “intermediary applications”, mainly building domain knowledge structures such as ontologies. For this, a prior stage of semantic relations acquisition from corpora is unavoidable. After recalling the state-of-the-art of methods on this topic, we will briefly discuss the specific contributions of the seven papers.

## 2. Acquiring semantic relations from corpora: The state-of-the-art

Current research in terminology engineering is witnessing an unprecedented effort directed towards the automatic detection of semantically-related terms from corpora. With the need for richly-structured language resources in order to accomplish such tasks as translation, text summarisation, question-answering, information extraction/retrieval or text mining, increased over the last few years, the terminology community has shifted its focus from term extraction to term structuring.

Although in term extraction, a few problems remain unsolved, e.g., adequate term filtering methods in order to accurately eliminate bad candidate terms, with the availability of terms extractors in various languages, the technology has reached a certain degree of maturity (*Acabit*: Daille 1996; *Fastr*: Jacquemin 2001; *Lexter*, *Syntex* and *Upéry*: Bourigault 2002; *Nomino*: David and Plante 1990; *TERMS*: Justeson and Katz 1995; *YATE*: Vivaldi 2001). In addition, the natural language processing (NLP) community has developed parsers which can also be tuned to perform this task. On the other hand, detecting semantic relations between domain terms and organising a domain terminology based on these semantic relations in order to build an ontology or a thesaurus, while having also been studied, remains a challenge.

As more domain-specific texts are produced, the semantic resources of domains have to be continually updated. Manually-built resources are known to be expensive, labour intensive and no longer feasible in the context of instantaneous electronic publications, which are inevitably accompanied by the appearance of new terms and the evolution of the semantic relations between them and already existing terms. Below, we summarize the main approaches to the acquisition of semantic relations from corpora.

### *Approaches to corpus-based semantic relation mining*

Research into the acquisition of semantic relations between terms stems from two main approaches: exogenous (top-down) or endogenous (bottom-up). The first approach consists in consulting an external semantic resource such as a domain dictionary, a taxonomy or an ontology as a gold standard, and using this resource to determine if terms extracted from a corpus share a particular semantic relation. The second approach is subdivided into a statistical and a linguistic approach.

Following the exogenous approach, to cite but a few recent studies on synonymy identification, Blondel and Senellart (2002) have measured the similarity of two words based on words they share in their dictionary definitions. Building on this work, Wu and Zhou (2003) add to this dictionary approach a measure of the similarity of contexts in which the two words appear in monolingual and multilingual corpora. In Hamon and Nazarenko (2001), synonymy identification between terms involved the use of external resources.

The endogenous, or “corpus-based” approach, relying on statistical evidence, views the problem of semantic relations as a term similarity identification problem where similarity is understood as a distributional problem. This idea, that can be traced back to Harris (1968), assumes that the more similar the distributional behavior of the units, the more likely that these units can be defined as synonyms. Following this tradition, Church and Hanks (1990) used a clustering technique to produce classes of words found in each other’s vicinity. Lin (1998) proposed a word similarity measure to automatically construct a thesaurus. However, this measure can only work on very large corpora (tens of millions of words) as it requires very high frequency thresholds.<sup>1</sup>

Although statistical approaches are very robust, since they do not require a linguistic analysis or a semantic resource, they could lead to grouping heterogeneous concepts or even antonymous ones in the same class. In Lin (1998), the most frequent words associated with the noun *brief* were *affidavit*, *petition*,

*memorandum, motion, lawsuit, deposition, slight, prospectus, document, paper* which all hold different relations with the initial word, including collocational ones. Distributional similarity has been explored for more complex language tasks like automatic thesaurus expansion (Grefenstette 1994).

The alternative endogenous approach relies on linguistic evidence and considers the discourse as the main information source. This approach gave rise to two types of methods: semantic similarity based on internal evidence called morpho-syntactic variations or lexical association (Daille 2003; Ibekwe-SanJuan 1998; Jacquemin 2001) and semantic relations signaled by lexico-syntactic patterns (Hearst 1992), also called *relational markers* (Condamines 2002). For the internal evidence approach, while it has been shown that some types of morpho-syntactic variations lead to semantic relations as in *British library* and *British national library*, where the second term is a hyponym of the first by pure lexical inclusion, this is not always the case, especially when the variation affects the head words as in *British library* and *British library reconstruction*. Here, the hypernym/hyponym relation is no longer obtained, although there is obviously an association of ideas which may be useful for other applications like topic mapping (Ibekwe-SanJuan and SanJuan, 2004).

While internal evidence undeniably enables the acquisition of a certain number of semantic links such as hypernym/hyponymy (*blood cell/white blood cell*), the approach is inherently limited in that it cannot capture conceptually related terms which do not share any lexical element. For instance, *AltaVista* and *search engine* will not be related, nor will *car* and *vehicle* be, whereas both pairs obviously share a hypernym/hyponym relation important to capture for many applications. In other words, the system will not allow the detection of semantic variants when the relation is materialized by other linguistic devices.

Since internal evidence is not sufficient to identify all the semantically related terms and it cannot be guaranteed that all the identified variants are semantically close, there was a need to find a complementary approach, independent of internal evidence. Thus current research on semantic relation mining between terms constitutes the natural complement of studies on internal evidence as a means of structuring terms.

External evidence searches for recurrent lexico-syntactic patterns which are supposed to signal specific semantic relations. The underlying hypothesis is that semantic relations can be expressed via a variety of surface lexical and syntactic patterns. According to Condamines (2002: 144–145), this idea can be traced back to Lyons (1978) who used the term *formulae* to refer to terms linked by a hypernymic relation, and also to Cruse (1986) who spoke of

*diagnostic frames*. Morin (1998) found traces of such a hypothesis even earlier. In the seventies, Robison (1970) tried to extract semantic information using lexico-syntactic patterns such as “*transformation of S from S into S*”.

This hypothesis, namely that semantic relations can be expressed via a variety of surface lexical and syntactic patterns, has combined research from different fields. In computer science, specifically for knowledge acquisition from texts, Ahmad (1992) and then Bowden et al. (1996) use the terms *knowledge probes* or *explicit relation markers* respectively. In computational terminology, pioneering work on terminological knowledge bases uses the expressions *knowledge rich contexts* (Meyer et al. 1992) and *conceptual relation patterns* (Condamines and Reyberolle 2000). This hypothesis has triggered many empirical studies (Aussenac-Gilles and Séguéla 2000; Hearst 1992; Morin and Jacquemin 2004; Nenadic et al. 2004; Suarez and Cabré 2002).

As these studies gather impetus, the focus is shifting to the ontology population using the terms and the semantic relations mined from specialised corpora (Biébow and Szulman 1999; Navigli and Velardi 2004).

### 3. The papers in this issue

This issue includes seven papers; four are dedicated to domain knowledge modelling tasks such as ontology building (Gillam et al.; Kerremans et al.; Malaisé et al., 3.1) or terminology structuring (Weeds et al., 3.2) although all have real end applications as their main focus. The paper by Wanner et al. (3.3) is devoted to automatic extraction of collocations from corpora using machine learning techniques. Daille (3.4) offers a review of the work on term variation (lexical association or term similarity by internal evidence) and demonstrates how different definitions found in the literature are dependent on the target applications. Finally, Nazarenko and Aït El Mekki (3.5) present a novel application of the exploitation of terminology processing for building back-of-the-book indexes.

#### *Corpus-based terminology: An anchor for ontology building*

Three papers in this issue deal with the relation between terminologies and ontologies. First used in philosophy, over the last 15 years the notion of “ontology” has taken a particular sense within the information processing community. Nowadays, an ontology is a formal knowledge representation that may be used

in order to perform automatic reasoning. More concretely, this representation appears as a network composed of nodes joined by arcs, both labelled with linguistic forms. Since this representation is more often linked with a domain and since it must be interpretable both by machines and by humans, these labels can be considered terms of the domain. This kind of relational representation is used by several disciplines whose aim is to describe knowledge in a structured form: natural language understanding, knowledge engineering but also terminology and information science (thesauri, glossaries, indexes, etc.). All these disciplines are now concerned with the general issue of establishing links between language and knowledge formalisation.

This interaction between terminology and knowledge representation began 15 years ago with the work on terminological knowledge bases by Meyer et al. (1992). However, similarities between terminology and NLP had been observed some years before (see, for example, Parent 1989).

Concerning the relation between language and knowledge formalisation, two approaches — that are sometimes complementary — exist. The first one argues in favour of general ontologies (Gruber et al. 1993), and the second one, in favour of local ontologies (Aussenac et al. 2000). In the first case, language is considered as a communicative tool, shared by all the speakers and concerned with a common knowledge (Guarino 1995). The main interest of such general ontologies is that they are supposed to be useful for all domains and/or all applications. It means they are supposed to be reusable. Wordnet and Eurowordnet (Felbaum 1999; Vossen 1998) but also Cyc (Lenat et al. 1990) are considered such general ontologies. In the second case (local ontologies), knowledge is considered as domain-dependent and even application-dependent. Local ontologies are quite suitable for specific applications (translation, indexing, knowledge representation) but their reusability is poor. These local ontologies are increasingly built from texts rather than by interviewing experts. This method of building ontology (from texts) is currently the most frequent and gives rise to numerous studies (Jacquemin 2001).

Such an approach raises important issues: how do we build the corpus? Is there a link between the corpus and the final application? Is it possible to determine a method from texts to the formalisation without human interpretation? And mainly, what kind of pre-existing resources are relevant for building an ontology from texts?

The three papers concerned deal with a method for building ontologies from texts. In two papers (Malaisé et al. and Gillam et al.), linguistic patterns are used in order to identify relations between terms. But they are very different

regarding other points, especially in their use of pre-existing resources, namely in their way of considering links between general knowledge (supposed to be shared by all readers) and specific knowledge (present within the corpus). While Malaisé et al. think that it is preferable to use mainly only domain corpus data, Gillam et al. also use also a general corpus: the British National Corpus (BNC).

Malaisé et al. support their opinion by an original experiment. They have compared dictionary definitions from the *Trésor de la Langue Française Informatisé* (TLFi) with corpus data (childhood domain from the point of view of anthropologists). An important number of definitions (236/354) are inadequate in comparison with corpus defining contexts. So, they prefer to use only knowledge rich contexts, that is contexts containing linguistic patterns allowing to spot definitional elements. One of their main studies concerns definition modelling and they try to situate definitional elements identified within the corpus as a part of the described model. Their aim is to build a tool for assisting the construction of what is called a *differential ontology*. A differential ontology explains how semantic features allow to situate parents and sibling(s) each *vis-à-vis* the other. This purpose of structuring terms is to assist the definition process. As in the other paper, the study concerns mainly hypernym relations but it shows also how co-hyponyms may be identified on the basis of shared words in their definitional contexts.

Gilliam et al. also use lexical patterns for spotting relations, but their use becomes operative only at the third step. The first step consists in the use of the BNC in order to compare the specific corpus with this “general” corpus. The aim is to select “interesting” (i.e. domain specific) words. The second step consists in the construction of a tree of collocating terms (based on a statistical method developed by Smadja (1993)). After the use of linguistic patterns (third step), the results of step 2 and 3 are unified and a *candidate conceptual* structure is presented. All these four steps are performed automatically. The results are integrated within a terminology metamodel (ISO 16642). The method is exemplified by the domain of nanotechnology.

The aims and methods described in the two papers are different: in addition to the fact that Malaisé et al. use only corpus data, the corpus itself is not very large and seems homogeneous according to genre, while Gillam et al. use a very large corpus within a domain, regardless of genre homogeneity. Malaisé et al.’s method involves only linguistic knowledge while Gillam et al.’s method combines linguistic and statistical knowledge.

However, the two approaches share a common feature: both use linguistic patterns for spotting and interpreting “knowledge rich contexts”, i.e. contexts expressing relationships or even Aristotelian formulae (Genus, i.e. hypernym, and differentiae). This method was first proposed by Hearst (1992), probably inspired by such authors as Cruse (1986) or Lyons (1978). It supposes that linguistic patterns may express some predefined relationships, specifically paradigmatic ones (hypernymy and meronymy). This approach is now well known for different languages such as English, French or Spanish.

New ideas are now being developed in order to take into account variations between corpora concerning their genre (Condamines 2002). This supposes that the use of linguistic patterns does not yield the same results for each corpus genre. The difficulty lies then in identifying relevant genres; this may be very difficult since genre classification may vary not only with speakers’ situation but also with the situation of readers or interpreters.

What is more problematic are cases of silence which correspond to parts of text that may be interpreted as expressing relations and that are not spotted by classical patterns. Some linguistic patterns are indeed specific to a corpus and are not predictable. These cases are probably not very frequent but it would be regrettable not to detect them, specifically in the case of small corpora (corpora from firms are often small). It would be necessary to devise methods for detecting such patterns. The perspectives regarding linguistic patterns will probably consist in taking into account variation between genres, which are predictable, and the non-predictable specificities of corpora.

The paper of Kerreman et al. presents an ongoing application-oriented terminography project for financial forensic purposes. This research, part of the European project FF POIROT, deals with the prevention of value-added tax carousel fraud in the European Union and the detection of securities fraud using multilingual knowledge repositories. The paper shows clearly how users and applications determine what textual information can be included in a multilingual terminology base and how to structure it at macro and micro-levels. This application is based on a multidisciplinary approach in terminography called *Termtography* (Kerremans et al. 2003; Temmerman and Kerremans 2003). Termtography integrates a knowledge specification phase that assists the corpus selection process and the specific criteria for selecting relevant knowledge units or “terms” (representation in a natural language of a unit of understanding considered relevant to given purposes, applications or groups of users).

*Combining statistical and symbolic methods for term similarity mining*

Weeds et al. deal with the acquisition and structuring of semantically-related terms through a combined use of linguistic (grammatical roles) and statistical similarities in order to mine term similarity. The authors applied four similarity measures (L1 Norm, Jensen-Shannon, Lin's measure, co-occurrence retrieval) to identify classes of semantically related multi-word terms in the biomedical domain. The effectiveness of the measures in correctly predicting the semantic type of a class of terms is evaluated against a "gold standard", the hand-built GENIA ontology. The authors showed that distributional similarity can be used to predict semantic types with a high degree of accuracy, reaching an optimal value of 63.1%. The major interesting features of this paper reside in the following :

- the focus is on multi-word term similarity in contrast to the majority of statistical approaches to word similarity almost always focused on single word units;
- the statistical measures used in distributional similarity generally need huge corpora (several millions of words) where the frequency thresholds need to be very high (=100) in order for the measures to be effective. Here the authors applied the statistical measures to sparse data where the frequency of each term rarely exceeds 5. They applied the measures to classes of lexically related terms (sharing the same head word but different modifiers) instead of to atomic terms;
- while studies on distributional similarity were applied to general language vocabulary, Weeds et al. apply them to a specialised corpus in the biomedical domain. Parsing a specialised corpus poses vocabulary coverage problems to most electronic resources;
- more importantly, the authors first perform a prior deep syntactic analysis on the corpus, using a parser (Pro3Gres) which enables them to identify grammatical dependency relations (subject, verb, object) between the terms and verb phrases and to resolve long distance dependency relations. Distributional similarity is thus not mined directly on the co-occurrence of terms but on the co-occurrence of grammatical relations such that two terms often appearing as "subject" or "object" of a given verb will tend to be more similar semantically.

Despite the sparseness of the data, their study shows a significant correlation between distributional and semantic similarity, where similarity is defined in terms of identical grammatical functions between classes of lexically-associated

terms. The mined similarities can be used in the initial stage of ontology construction as they enable the system to robustly acquire classes of semantically-related terms and to predict their semantic types, thus their position in the ontology. Another potential application is the expansion of user's query terms to its semantic class in order to enhance recall.

### *Using Machine Learning Techniques (MLT) for information extraction*

One of the strategies used nowadays for improving large scale information extraction is machine learning (ML). This strategy always follows the same process: some cases are identified as representatives of a given type of the information that it is desired to extract automatically; these examples are provided to the system which "learns" to detect them on the basis of the regularities that can be observed. The article of Wanner et al. is based on ML techniques to extract collocations automatically from corpora and to label them semantically by means of lexical functions (Mel'čuk 1996). The final purpose is the automatic compilation of a dictionary of collocations specifically for translators and technical editors. Collocations are conceived in the paper as lexico-semantic links between terms or lexemes. For that reason, Lexical Functions (LFs) are added in order to identify collocations.

On the basis of the semantic description of items which occur together in syntagmatic relations, ML techniques can be an efficient strategy for capturing collocations. EuroWordNet (Vossen 1998) is used as a source of semantic descriptions of items. More specifically, in this paper, an experiment of automatic extraction of collocations corresponding to V-N and N-V structures in Spanish legal texts is carried out with two concrete ML-classification techniques: Nearest Neighbour (NN) and Tree Augmented Bayesian Network (TAN). Both techniques are applied to the same corpus in order to evaluate their results for automatic extraction of collocations. The NN algorithm matches the semantic description of a candidate as a whole against the descriptions of instances of lexical functions or against the descriptions of instances of freely combined bigrams in order to capture the level of interdependence between the components of the bigrams. The TAN-classifier uses the interdependence between individual meaning components of the bigram. The result is that NN performs better than TAN.

*Terminology structuring by internal evidence: A typology of term variations and target applications*

Daille presents a long-awaited synthesis of the concept of “term variation” and of the subsequent work thereon. Although this notion and the reality it covers have been studied by several authors amongst which Daille (1996), Ibekwe-SanJuan (1998) and Jacquemin (2001), no unifying theoretic framework exists under which the different definitions and types of variants can be situated. What these studies have established is that variations are morpho-syntactic operations which create relations between terms. Variations can be captured by a series of surface morphological, lexical and syntactic cues (thus by internal evidence). Far from being rare, variation phenomena are frequent in specialised corpora and concern between 15% and 35% of the terms.

However, a unified typology of variants remained highly desirable as different authors gave similar but not quite identical definitions of variations. In fact, a dichotomy seems to emerge from Daille’s survey: variations defined for terminology resource-building or language-related applications (translation) versus variations defined for end applications in other fields using terminological resources. In the first case, authors define variants as different forms of an existing reference term which are conceptually close to the reference term (Daille 2003; Jacquemin 2001; Jacquemin and Tzoukermann 1999). This includes variants such as synonyms, hypernyms, hyponyms or morphological variants of the same term. In essence, what is required by terminological resources (term base, lexicon, ontology, thesaurus) or for translation is that the variant of a term belong to the same semantic class as that term.

At the other end, researchers processing terms and their variants in applications such as information retrieval, question-answering, (Dowdall et al. 2003) or science and technology watch (Ibekwe-SanJuan 1998; Polanco et al. 1995) adopted looser definitions. A variant of a term is any term related to it by a series of variation operations and this could eventually lead to a semantic distance. For instance, for applications in science and technology watch, a variant does not necessarily need to belong to the same semantic class as the initial term. The expert user is more interested in the association of domain concepts which can include several semantic links (meronym, hypernym/hyponym, synonym, antonym, etc.) or simply a vague “see\_also” relation. This type of application is sensitive to topic shifts and evolution. Thus the interest here will be in capturing such links as that found between *frozen sweet dough* and *frozen sweet dough baking* involving a topical shift, thus a semantic distance. Daille’s survey concludes by a typology of core variations present in several

applications, albeit under different names and discusses the benefits/problems encountered in identifying them in several applications.

### *Terminology engineering for book index building*

Nazarenko and Ait El Mekki integrate terminological processing into the application of building book indexes. Book indexes are useful for rapidly accessing the key themes of any long document. It is usually onerous to read a book in a linear manner before distinguishing the key items relevant to our information need. While the fundamental problem raised by building book indexes is similar to that of automatic indexing, i.e. the selection of weighted index terms which can sufficiently distinguish the content of one document from another, this specific application has been rather neglected by the information retrieval (IR) community where research has been strongly focused on automatic indexing of large document collections for query expansion (Baeza-Yates and Ribiero-Neto 1999; Buckley et al. 1995; Salton 1986; Sparck Jones 1999). Although some professional index building tools exist, they poorly exploit the content of documents and largely ignore the term variation paradigm. The resulting indexes have poor readability and are also poorly structured (few hierarchical levels are allowed). Moreover, in these tools the bulk of the building work falls to the author or indexer who has to manually select the correct index entries. This is at best tedious in the case of very long documents (e.g., manuals, treaties, encyclopedias).

Nazarenko and Ait El Mekki exploit terminological tools for building back-of-book indexes and investigate whether this application imposes new requirements on terminological engineering. Their system, called *IndDoc*, consists of an original method for building back-of-the book indexes. It fruitfully combines experience and knowledge gathered by research on computational terminology and user needs issues in IR. *IndDoc* relies on a term extractor to extract the index entries. Semantic relations are identified by the combined use of internal evidence (lexical associations) and external evidence (relational markers). These relations help to gather the occurrences (variants) of various descriptors under the same entry and to make some semantic relationships explicit in the nomenclature. It also enables the system to offer different views of the same index: lists or semantic networks. Then, the index entries are weighted using various indices to rank the index entries: their discriminating power (a refined version of the *td.idf*<sup>2</sup> measure), and the informative strength of the

text segments in which they are found. The latter makes use of both the logical structure of the document and page formatting cues. Relevance of index entry is thus computed from several cues: formatting cues, document structure, distribution and semantic network density around a given entry. Index entry ranking is of great benefit at the validation step because the indexer can validate the most relevant descriptors first. Segment ranking is used to present the reference in the index by order of importance rather than by page number. The whole process is considered cooperative, since the user (the human indexer) has the final say.

The authors showed that incorporating terminological tools such as term and relation extractors improves the quality of indexes and that in the other sense, the application (index building) also influences the type of terminology processing. Building a back-of-the-book index calls for the integration of discursive markers (logical and formatting markers, semantic relations patterns) as well as frequency criteria. The authors also highlight the necessity for an index building interface to fully exploit the terminological network paradigm whose importance has already been established in the terminology community. In this paradigm, the place and meaning of a term can only be fully comprehended when the term is placed within its semantic network, in relation to other terms and not in a flat list. The end application considered here will go a long way in bringing closer researchers in the IR and terminology communities around common research issues.

## Conclusion

The number of submissions received in response to the call for papers for this issue was very high: 21 of which only seven could be accepted. This shows a real interest in this subject and proves that terminology engineering has now reached a new stage. It is now clear that processing, without considering the type of application envisaged, is misguided. This observation raises a number of points.

- First of all, since terminology processing has to take its application into account, it is necessary to identify the kinds of possible applications. Is it possible to draw up a list? And, more difficult: is it possible to anticipate new applications?
- Secondly, it is important to understand how these applications are linked. Is it possible to build categories of applications, i.e. is it possible to constitute types of processing corresponding to groups of applications?

- Finally, what seems important is to characterise precisely the knowledge from which terminologies are built (it means, more and more, texts) and the knowledge for which these terminologies are built (i.e. the knowledge necessary for a particular applications). This will allow the adaptation of processes to new applications, without incurring the danger of building *ad hoc* processes.

These issues have to be organised according to the two user dimensions already stated at the beginning of this introduction. First, the real users of terminologies: translators, indexers, documentalists, knowledge engineers, natural language engineers, etc. At the moment, their needs are well known, even if the compatibility between these needs is not completely established. But it is very difficult to anticipate new needs because they depend both on the availability of new tools and on new needs discovered as parts of projects within firms. The second dimension is intermediate between texts and final applications. It concerns general means of representation and specifically relational representations which are numerous (as the papers within this issue show). The ways of building these relational representations from texts are now one of the most important issues and many studies around this topic are in progress. Less frequent is research which evaluates to what extent the final application, which directly benefits end-users, influences the construction of relational representations. If a typology of final users were to become available, it would be possible to associate types of representation with different uses in order to construct adapted representations.

## Acknowledgements

The Guest Editors would like to thank the following people for acting as Advisory Board for this issue: Sophia Ananiadou, Nathalie Aussenac-Gilles, Caroline Barrière, Didier Bourigault, Béatrice Daille, Kyo Kageura, Geneviève Lallich, Widad Mustafa El-Hadi, Jean Royauté, Sylvie Sluzman, Rita Temmerman, Philippe Thoiron.

## Notes

1. Only words occurring more than a hundred times were considered.
2. It is the most popular measure in IR for selecting keywords to index a document or a document collection. This measure was proposed by Salton (1986) and has been refined since by several IR researchers.

## References

- Ahmad, K. and H. Fulford. 1992. *Knowledge Processing: Semantic Relations and their Use in Elaborating Terminology*. Computing Sciences report CS-92-Guildford: University of Surrey.
- Aussenac-Gilles, N., B. Biébow, S. Szulman. 2000. "Corpus analysis for conceptual modeling." In *Proceedings of the Workshop on Ontologies and Texts, EKAW'2000* (European Knowledge Acquisition Workshop). 13–20. Juan les Pins, France
- Aussenac-Gilles N. and P. Séguéla. 2000. "Les relations sémantiques : du linguistique au formel." *Cahiers de Grammaire : Numéro spécial Linguistique du corpus* 25, 175–198.
- Baeza-Yates, R. and B. Ribeiro-Neto. 1999. "Query operations." In Baeza-Yates, R. & B. Ribeiro-Neto (eds.). *Modern Information Retrieval*. 117–139. Boston: Addison Wesley.
- Biébow, B. and S. Szulamn. 1999. "TERMINAE : A linguistic-based tool for the building of a domain ontology." In *Proceedings of the 11th European Workshop, Knowledge Acquisition, Modelling and Management (EKAW '99)*. 49–66. Dagstuhl Castle, Germany.
- Blondel, V. and P. Senellart. 2002. "Automatic extraction of synonyms in a dictionary." In *Proceedings of SIAM workshop on Text Mining*. Arlington, USA
- Bourigault, D. 2002. "Upéry : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus." In *Actes de la 9<sup>e</sup> conférence annuelle sur le Traitement Automatique des Langues (TALN 2002)*. 75–84. Nancy, France.
- Bowden, P.R., P. Halstead and T.G. Rose. 1996. "Extracting conceptual knowledge from texts using explicit relation markers." In *Proceedings of the European Knowledge Engineering Workshop (EKAW'96), Lecture Notes in Artificial Intelligence*. 146–162. Berlin: Springer Verlag.
- Buckley, C., G. Salton, J. Allen and A. Singhal. 1995. Automatic query expansion using SMART: TREC-3. In Harman, D. (eds), *The Third Text Retrieval Conference (TREC-3)*. 69–80. NIST Special publication 500-225.
- Cabré, M. T. 2003. "Theories of terminology. Their description, prescription and explanation." *Terminology* 9(2), 163–200.
- Church, K.W and P. Hanks. 1990. "Word association norms, mutual information and lexicography." *Computational Linguistics* 16(1), 22–29.
- Condamines, A. 2002. "Corpus analysis and conceptual relation patterns." *Terminology* 8(1), 141–162.
- Condamines, A. and J. Reyberolle. 2000. "Construction d'une base de connaissances terminologiques à partir de textes : expérimentation et définition d'une méthode." In Charlet, J., M. Zacklad, G. Kasel and D. Bourigault (eds.). *Ingénierie des Connaissances, évolutions récentes et nouveaux défis*. 127–147. Paris: Eyrolles.
- Cruse, D. A. 1986. *Lexical Semantics*. Cambridge: Cambridge University Press.
- Daille, B. 1996. "Study and implementation of combined techniques for automatic extraction of terminology." In P. Resnik and J. Klavans (eds.). *The Balancing Act : Combining Symbolic and Statistical Approaches to Language*. 49–66. Cambridge: MIT Press.
- Daille, B. 2003. "Conceptual structuring through term variations." In *Proceedings of the ACL-2003, Workshop on MultiWord Expressions: Analysis, Acquisition and Treatment*. 9–16. Saporro, Japan.

- David, S. and P. Plante. 1990. "Le progiciel TERMINO : De la nécessité d' une analyse morphosyntaxique pour le dépouillement terminologique de textes." In *Actes du Colloque Les Industries de la Langue*. 140–155. Montréal, Canada.
- Dowdall, J., F. Rinaldi, F. Ibekwe-SanJuan and E. SanJuan. 2003. "Complex structuring of term variants for question answering." In Bond, F., A. Korhonen, D. MacCarthy and A. Villacencio (eds.). *Proceedings ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*. 1–8. Sapporo, Japan.
- Fellbaum, C. et al. 1999. *Wordnet. An Electronic Lexical Database*. Cambridge, London: The MIT Press.
- Grefenstette, G. 1994. *Exploration in Automatic Thesaurus Discovery*. Boston, MA: Kluwer Academic Publisher.
- Gruber, T. R. 1993. "A translation approach to portable ontology specifications." *Knowledge Acquisition* 5(2), 1993.
- Guarino, N. 1995. "Formal ontology, conceptual analysis and knowledge representation." *International Journal of Human-Computer Studies* 43, 625–640.
- Hamon, T. and A. Nazarenko. 2001. "Detection of synonymy links between terms: experiments and results." In Bourigault, D. C. Jacquemin and M. C. L'Homme (eds.). *Recent Advances in Computational Terminology*. 185–208. Amsterdam/Philadelphia: John Benjamins.
- Harris, Z. S. 1968. *Mathematical Structures of Language*. New York: Wiley.
- Hearst, M. A. 1992. "Automatic acquisition of hyponyms from large text corpora." In *Proceedings of the 14th International Conference on Computational Linguistics*. 539–545. Nantes, France.
- Ibekwe-SanJuan, F. and E. SanJuan. 2004. "Mining textual data through term variant clustering: the termwatch system." In *Proceedings "Recherche d'Information assistée par ordinateur" (RIA0)*. 487–503. Avignon, France.
- Ibekwe-SanJuan, F. 1998. "Terminological variation, a means of identifying research topics from texts." In *Proceedings of the Joint International Conference on Computational Linguistics (COLING-ACL'98)*. 564–570. Montréal, Canada.
- Jacquemin, C. 2001. *Spotting and Discovering Terms through Natural Language Processing*. Cambridge (MA): MIT Press.
- Jacquemin, C. and E. Tzoukermann. 1999. "NLP for term variant extraction: a synergy of morphology, lexicon and syntax." In Strzalkowski, T. (ed.), *Natural Language Information Retrieval*. 25–74. Boston, MA: Kluwer.
- Justeson, J. and S. Katz. 1995. "Technical terminology: some linguistic properties and an algorithm for identification in text." *Natural Language Engineering* 1(1), 9–27.
- Kerremans, K. and R. Temmerman. 2004. "Towards multilingual, termontological support in ontology engineering." In *Workshop Terminology, Ontology & Knowledge representation*. 80–86. Lyon, France.
- Lauriston, A. 1994. "Automatic recognition of complex terms: problems and the TERMINO solution." *Terminology* 1(1), 147–170.
- Lenat, D. B., R. V. Guha, K. Pittman, D. Pratt and M. Sheperd. 1990. "Cyc: towards programs with common sense." *Communications of the ACM* 33(8), 30–49.
- Lin, D. 1998. "Automatic retrieval and clustering of similar word." In *Proceedings of the Joint International Conference ACL-COLING'98*. 768–773. Montréal, Canada.

- Lyons, J. 1978. *Éléments de sémantique*. Paris: Larousse Universités.
- Melčuk, I. 1996. "Lexical functions: a tool for the description of lexical relations in the lexicon." In Wanner, L. (ed.). *Lexical Functions in Lexicography and Natural Language Processing*. 37–102. Amsterdam/Philadelphia: John Benjamins.
- Meyer, I., L. Bowker and K. Eck. 1992. "Cogniterm: an experiment in building a terminological knowledge base." In *Proceedings of the 5th EURALEX International Congress on Lexicography*. 159–172. Tampere, Finland.
- Morin, E. and C. Jacquemin. 2004. "Automatic acquisition and expansion of hypernym links." *Computer and the humanities*, 38(4), 363–396.
- Morin, E. 1998. "Prométhée : un outil d'aide à l'acquisition de relations sémantiques entre termes." In *Proceedings Traitement automatique des langues naturelles*. 172–181. Paris, France.
- Navigli, R. and P. Velardi. 2004. "Learning domain ontologies from document warehouses and dedicated web sites." *Computational Linguistics* 30(2), 151–179.
- Nenadic, G., I. Spasic and S. Ananiadou. 2004. "Mining term similarities from corpora." *Terminology* 10(1), 55–81.
- Parent, R. 1989. "Recherche d'une synergie entre développement linguistique informatisé et systèmes experts : importance de la terminologie." *Meta* 34(3), 611–614.
- Polanco, X., L. Grivel and J. Royauté. 1995. "How to do things with terms in informetrics: terminological variation and stabilization as science watch indicators." In *Proceedings of the 5th International Conference of the International Society for Scientometrics and Informetrics*. 435–444. Illinois, USA.
- Robison, H.R. 1970. "Computer-detectable semantic structures." *Information storage and retrieval* 6, 273–288.
- Salton, G. 1986. "Another look at automatic text-retrieval systems." *Communications of the ACM* 29(7), 649–656.
- Smadja, F. 1993. "Retrieving collocations from text: Xtract." *Computational Linguistics* 19(1), 143–178.
- Sparck-Jones, K. 1999. "What is the Role of NLP in Text Retrieval?" In Strzalkowski, T. (eds.). *Natural Language Information Retrieval*. 1–25. Dordrecht: Kluwer Academic Publishers.
- Suárez, M. and M. T. Cabré. 2002. "La variación denominativa en los textos de especialidad: indicios lingüísticos para su recuperación automática." In *CD-Actas del VIII Simposio Iberoamericano de Terminología*. Cartagena de Indias, Spain.
- Strzalkowski, T. (ed.). 1999. *Natural Language Information Retrieval*. Dordrecht: Kluwer Academic Publishers.
- Vivaldi, J. 2001. *Extracción de Candidatos a Término mediante combinación de estrategias heterogéneas*. Tesis doctoral. Universitat Politècnica de Catalunya. Barcelona: Institute Universitari de Linguística Aplicada (IULA).
- Vossen, P. (ed.). 1998. *EuroWordNet: a Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers.
- Wu, H. and M. Zhou. 2003. "Optimizing synonymy extraction using monolingual and bilingual resources." In *Proceedings 2nd International Workshop on Paraphrasing: Paraphrase, acquisition and applications (IWP-2003), in ACL-2003*. 72–79. Sapporo, Japan.

*Authors' addresses*

M. Teresa Cabré  
IULA  
Pompeu Fabra University  
Rambla de Sta Mònica, 30–32  
Barcelona 08002-Spain  
Teresa.cabre@upf.edu

Anne Condamines  
ERSS  
CNRS and University of Toulouse Le Mirail  
Maison de la Recherche  
5 allées Antonio Machado  
31058, Toulouse — France  
anne.condamines@univ-tlse2.fr

Fidelia Ibekwe-SanJuan  
University of Lyon 3  
4, cours Albert Thomas  
69008, Lyon — France  
ibekwe@univ-lyon3.fr

*About the authors*

**M. Teresa Cabré** received her PhD in Romance Linguistics at Barcelona University in 1976. Since 1993 she has been a full professor at Pompeu Fabra University and a researcher at the Institute for Applied Linguistics (IULA) at the same University. She is leader of the research group IULATERM (Lexis, Terminology and Specialised discourse) formed by more than 20 full-time researchers. She was director of IULA from its creation in 1993 until 2004.

**Anne Condamines** completed her PhD in Linguistics at Toulouse II in 1990. Since 1993, she has been a full-time researcher at the CNRS within ERSS, Toulouse II, France. She is project leader for the group “Sémantique et Corpus” (Semantics and corpora). In 1993, she launched, together with Didier Bourigault, a research group funded by the French Department of Research. The group “Terminology and Artificial Intelligence” comprises terminologists, linguists and computer scientists. In 1994, she was awarded a national prize by the CNRS and its partners for her achievements in bridging the gap between theoretical research in linguistics and industrial applications.

**Fidelia Ibekwe-SanJuan** received her PhD from the University Stendhal, Grenoble 3 in 1997. She is currently lecturing at Jean Moulin University Lyon 3. Her main research interests revolve around clustering terminology units and relations for different applications related to strategic information awareness (business intelligence, science and technology watch, textmining) and to question-answering.