Terminological knowledge bases From Texts to Terms, from Terms to Texts.

Anne Condamines

Terminological Knowledge Bases: From Texts to Terms, from Terms to Texts.

Anne Condamines

Cognition, Langues, Langage, Ergonomie (CLLE)

University of Toulouse, CNRS, University of Toulouse Jean Jaurès, France

1. Introduction

This chapter addresses issues concerning terminological knowledge bases (TKBs). A TKB was initially defined at the crossroads of terminology and knowledge engineering as a knowledge base taking into account the twofold nature of a term: conceptual and linguistic. In fact, at the beginning, issues concerning TKBs mainly concerned the way knowledge engineering could help terminologists to structure specialized knowledge. The main aim was to build knowledge representation models that were usable by tools. On the one hand, this step was of great importance for the improvement of terminology studies, particularly because terminologists had to analyze the link between terms and specialized discourses. But, on the other hand, the necessity of representing knowledge under a network model sometimes appeared to be too restrictive and too far remote from discourse and this point generated dissatisfaction. Nevertheless, an unforeseen result of studies developed with the aim of building TKBs from texts was that tool-assisted methods were designed for exploring specialized texts systematically. First, this chapter presents the origins and the aim of terminological knowledge bases. Second, it details the tool-assisted linguistic methods for building TKBs from texts. Finally, it shows how similar methods may be applied to meet other needs in which terms and relations are not the result of linguistic analyses but rather the point of departure for the study of specialized texts. Note that the viewpoint adopted in this chapter is mainly that of a linguist.

2. Historical Perspectives

In the case of terminological knowledge bases, a historical perspective is very important because this concept symbolizes the confluence between at least two disciplines: Terminology and Knowledge Engineering. The name itself reflects this dual heritage. The first occurrence of this name was in articles produced by I. Meyer and her team in the Cogniterm project (Meyer et al., 1992), but other teams, at almost the same time, were developing similar projects and methods to build them (in Surrey, Ahmad, 1993; in Toulouse, Condamines and Amsili, 1993). In all cases, the teams involved in the project were composed of linguists/terminologists and computer scientists, a clear sign that the concept presented an interest for both communities. In the following two sections, the respective interests of the two communities in this concept are presented.

2-1 Knowledge Engineering point of view

In the 1990s and even before, some authors noted the possible convergence between terminology and artificial intelligence (Parent, 1989). For example Wijnands wrote: “Terminology and artificial intelligence face the same problem” (Wijnands, 1993: 168). For knowledge engineering, three elements motivated the encounter with terminology.

a- Taking into account the linguistic nature of computerized representations

When knowledge has to be represented, knowledge engineering models always use linguistic forms to designate the represented objects. This way of labelling representations is very important because these models are built to give access to textual data; thus, they have to use the same character strings as the ones used in texts. The problem is that these labels are not only character strings, they are also words, which implies meaningful elements. They allow the switch between the formal representations and the texts. For most knowledge engineers, it appeared crucial to take this twofold aspect of linguistic labels into account.

b- Integrating terminological data into existing models of representation

The most widely used model of knowledge representation in the 1990s, and even now, is the one using conceptual networks in which both nodes and links are labelled by linguistic forms. In such models, the node labels may easily correspond to terms and the link labels to conceptual relations. This mode of representation influenced terminologists and led them to develop methods to build terminological networks from texts, but also to focus only on contexts assumed to be useful for building a relational representation (see below).

c- Exploring textual data rather than just soliciting experts

For several years, knowledge acquisition was done by interviewing experts about their own knowledge. This method was rather unsatisfactory, however, for two main reasons. Experts were not always prepared to contribute to this process either for lack of time or because they did not understand how to collaborate with knowledge engineers. In addition, it appeared that domain experts had some difficulty describing their own knowledge and sometimes, their description did not correspond to their practices. The use of expert texts was considered as a good solution to tackle these difficulties.

2-2 Terminological point of view

From a terminological viewpoint, four elements contributed to the junction with knowledge engineering.

a) Need for the management of technical documentation (Condamines, 1995)

Since the 1980s, the e-management of massive amounts of documentation had become a great challenge within firms. For example, in the Airbus Company, it was estimated that the documentation concerning a plane would fill the plane. So, a large number of projects were developed in order to assist in processing the documentation. In most of these projects, terminology played a crucial role.

b) Inadequacy of existing terminological resources

With these new needs emerging within firms, most of the terminologies developed by official bodies appeared to be inadequate because they were too far from real uses. Moreover, most engineers did not even know that there were terminological standards built by official bodies. For terminologies to be usable in documentation management within firms, it became necessary to build them based on real usage.

c) Development of corpus linguistics methods

At the end of the 1980s and the beginning of the 1990s, several projects to build general corpora were carried out, especially for English. One of their main aims was to design methods to help dictionary definitions. One of the most important projects was the design of the Cobuild dictionary, using a large corpus. Sinclair anchored the process and inspired generations of lexicographers (Sinclair, 1991). At the same time, studies were carried out on definitions in discourse (Flowerdew, 1992). The main bases of e-lexicology were then defined and usable for building terminologies from texts. See this book chapter 8.

d) Development of Natural Language Processing tools

During the 1990s, Natural Language Processing perspectives evolved. The main point became not only to implement formalisms assumed to be relevant for describing the functioning of language, but rather to help a user to use the texts. Concordancers were designed and made available to linguists, but also tools dedicated to the exploration of specialized corpora such as candidate-terms extractors or candidate-relations extractors (see below and this book, chapter 12).

The 1990s and 2000s were also marked by the fact that a large number of researchers criticized the General Theory of Terminology, developed by an Austrian engineer, Eugen Wüster, during the 1930s. In order to limit the difficulties inherent to language functioning, he proposed to standardize terminology, considering that specialized languages were very different from general language because they were used in highly constrained situations. He thought that it was easy to standardize the terminology in such controlled situations (Wüster, 1974). With such a point of view, it was necessary to keep terminology away from linguistics and to consider terms as just labels of concepts which became the main elements to be studied (Wüster, 1979).

Unfortunately, even if they have some specific characteristics, specialized texts are not so different from general ones and it is impossible to control all the possible variations of terms.

New theories such as socioterminology (Gaudin, 1990), sociocognitive terminology (Temmerman, 2000) or the communicative theory of terminology (Cabré, 1999) were proposed to describe terms according to their functioning in real situations. All these approaches can be subsumed under the term of “textual terminology” (Pearson, 1998).

So, both from an applicative and a theoretical point of view - knowledge engineering and terminology -, all the parameters were converging towards the definition of a new concept,

developed to take into account the methods and needs common to both knowledge engineering and textual terminology, namely the terminological knowledge base.

3. Core issues and Topics

Even if many projects have developed different models of TKBs, some characteristics are always present in the definition of a terminological knowledge base.

a) Distinction between terms and concepts

Whatever the model, in a TKB, there are always two levels for representing knowledge. The first one may be considered as the conceptual level, that supports the network representation; the second one is the linguistic level, in charge of the lexicalization of the conceptual level in each language or language community considered. From a linguistic point of view, this representation can be criticized because it assumes that there is a unique knowledge representation, common to all languages. In the wake of Sapir and Whorf's work, it is difficult to accept such a point of view because, as they claimed, language shapes our behavior and not the opposite. Nevertheless, this linguistic relativism should be qualified, for two reasons. First, TKBs are concerned by a limited domain and, in most cases, by a limited application. Within this domain and this application, one may consider that there is no crucial variation in knowledge representation among languages, even if it is a strong hypothesis. Second, this representation is very useful in order to deal with phenomena such as synonymy (two terms/one concept) or polysemy (one term/two linked concepts) and even inter-linguistic equivalence (one concept/terms in different languages). Finally, the representation is controlled both by the uses within the corpus and by the aim of the TKB design.

b) Knowledge representation in a network form

Probably the most important characteristic of a TKB, at least for terminologists, is the fact that knowledge is represented in the form of conceptual relations. In previous terminological data bases, definitions appeared in a discursive form. The challenge with TKBs was to replace most of these discursive forms by a reticular one, that is to say, first, to de-contextualize the terms and, second, to retain only the contexts that can be used to code knowledge in a network form. This form is not always adequate and may lead to a loss of knowledge. However, as for the separation between concepts and terms, this representational choice may have advantages. The most important one is that terminological data become manageable by tools and then tools can be used to verify the terminological data: consistency, completeness and so on. For example if a father (hypernym) appears as having only one son (hyponym), which seems impossible from a linguistic point of view, the user may be alerted and then try to identify at least one co-hyponym (a brother) within the corpus or by consulting an expert.

c) Corpus as a source of knowledge

TKBs are always built by using a corpus as a source of knowledge. As a result, the first step in the study consists in building a corpus that is well adapted not only to the domain but also, most of the time, to the application concerned by the resource. Then, as for general corpora,

the specialized corpus becomes the object to be studied and must be as representative as possible.

From a linguistic point of view, the emergence of the TKB concept led to two kinds of studies, the first concerning the characterization of the specificities of terms and how to spot them within a corpus, the second concerning the description of relationships and how they are expressed within corpora.

Of course, studies on these two issues already existed. The definition of terms had been addressed especially in lexicology or in translation. In the theory of sublanguages, the functioning of terms was very often described in comparison with the general lexicon by using the notion of "deviant mode" (Lehrberger, 1986).

Conceptual (or semantic) relationships and the way they are expressed in languages had also been studied by semanticists, lexicologists and even philosophers (for example, Green et al., 2002; Winston et al., 1987; Cruse, 2002).

However, the need to propose systematic descriptions for specialized languages, usable by tools, strengthened the research and opened up new perspectives.

3-1 Tools and methods

Since the 1990s, many tools have been developed in order to assist the extraction of terms and relations from texts. In most cases, researchers speak about "candidate-terms extraction" rather than "terms extraction" to highlight the fact that the results must be validated by terminologists and/or domain experts. From a linguistic point of view, three assumptions about how terms function underlie the design of these tools.

The first assumption is that terms are mainly nominal groups. Several studies have shown that, in existing terminologies, around 70% of terms are nominal groups. Tools applying this characteristic use a tagged corpus and seek all the strings of characters corresponding to nominal structures, for example: adjective noun (*incandescent lava*), noun preposition noun (*eruption of magma*), noun preposition adjective noun (*mantle of molten rock*), noun preposition determiner noun preposition determiner noun (*collapse of the summit of a volcano*) etc.

The second assumption is that, in a specialized corpus, the most recurrent character strings are likely to correspond to terms. When implemented, this statistical method makes it possible to spot not only nominal compounds but all the recurrent forms: verbs (*to extrude*), adjectives (*eruptive*), etc.

The third assumption is that terms can be extracted by comparing the number of occurrences of a character string in a specialized corpus with the number of occurrences of the same character string in a general corpus. The point is that if a character string (or a sequence of character strings) appears significantly more often in a specialized corpus than in a general one, this string of characters may correspond to a term. With such a method, terms that also correspond to general words may be proposed as results. Such cases are not rare since many

terms are used in general corpora (see for example, *telecommunication satellite, hemoglobin, oceanographer…).*

Most of the time, tools implement two of these assumptions. For example, Termostat (Drouin, 2003) combines statistical and comparative approaches, while TerMine (Frantzi et al., 2000) combines statistical and linguistic approaches.

3-2 The notion of "Knowledge rich contexts"

The term "knowledge rich context" was proposed by Meyer in the context of building TKBs. It was defined as "a context indicating at least one item of domain knowledge that could be useful for conceptual analysis" (Meyer 2001: 281).

From a corpus linguistics perspective, knowledge rich contexts can be described as all the linguistic elements that can be used in order to identify a conceptual relationship. Depending on the authors, these linguistic elements have different names, e.g., *formulae* (Lyons, 1977), *diagnostic frames* (Cruse, 1986), *hinges* (Pearson, 1996). From a linguistic point of view, these patterns correspond mainly to local grammars (Gross, 1997) that may be designed (and then used in a tool), in order to describe relationships (Barnbrook and Sinclair, 2001). More precisely, local grammars aim to spot triplets such as [T1-relationship-T2], for example, [branch is-a-part-of tree]. Indeed, such triplets are searched in order to build a network by combining them. For example, the two triplets [leaf is-a-part-of branch] [branch is-a-part-of tree] make it possible to start the design of a network.

Here are some examples of relational patterns and the triplets spotted.

[NP1 especially NP2] (hypernym between NP1 and NP2)

1) Furthermore, compared to the minorities especially blacks, the majority population generally has a higher likelihood of living in a nuclear family within a stable community.

[NP1, the most adjective NP2] (hypernym between NP2 and NP1

2) *Breakfast is the most important meal of the day.*

[NP1 be composed of NP2] (meronymy between NP1 and NP2)

3) The vasculature of human skin is composed of the nutritional capillaries and the thermoregulatory blood vessels.

Several studies have described relational patterns (Auger and Barrière, 2008).

Concerning the productivity of such relational patterns, variations have been identified among corpora, in particular according to the textual genre, which has been described as crucial in the functioning of patterns of conceptual relations patterns (Condamines, 2002; Marshman et al. 2008). The role of textual genre is important both for the presence vs absence of patterns but also for the way polysemic patterns may be interpreted. For example, *to provoke* may have the meaning of *to cause*. But this meaning mainly occurs in scientific or technical texts where the verb appears mostly with inanimate arguments as in (4):

4) *Higher lengths and lower diameters can provoke excessive voltage fall.*

In general texts, where animate arguments are frequent, it is probably the case that the main use is equivalent *to excite* as in (5):

5) *The child provokes the animal.*

Several studies have described step by step how to use these patterns to build a conceptual network (Condamines and Rebeyrolle, 2001; L'homme and Marshman, 2006).

However, the design of triplets [term-relationship-term] using texts is far from easy. In (Aussenac-Gilles and Condamines, 2012), several difficulties of this process are presented.

- One of the terms-concepts is missing

This is the case for example when an anaphoric pronoun is used instead of the term itself. It may be very difficult to identify the correct antecedent.

- T1 and T2 do not belong to the same grammatical category

This is the case for example with nouns and verbs. It is not possible to link them in a terminology. Nevertheless, this situation can be found in discourse as in:

6) *The numbering of cables consists in identifying and numbering each cable for an electrical cabinet.*

- Pattern and T2 are present in the same word

This phenomenon is seldom described. Some words composed of a base and an affix may contain both a term and a relational pattern (the affix). For example, in:

7) *The willow has been uprooted.*

One can deduce that, generally, *roots* are parts of a *willow* and, consequently, that a *willow* is probably a tree. The *up* prefix can be considered as a part-of-all pattern. Then *uprooted* is a term and it also contains a relational pattern (*up*) and another term (*root*).

- Polysemy of patterns

It is well known that some patterns can be polysemic. See for example *to provoke* in (4) and (5).

- Implicit relationship

Sometimes, the pattern cannot be interpreted directly but must be deduced. This can be the case with nominal anaphora. It is well known that, in some cases, there is a hypernymic relation between a head noun within a nominal anaphor and its antecedent.

For example in

8) *A cat entered, the animal had shiny fur.*

There is a hypernymic relation between *animal* and *cat*. This relation is not posed by the sentence but is assumed to be known. Nevertheless, some textual genres foster this type of relation (Condamines, 2005).

- Indirect interpretation (Condamines, 2000)

In some cases, the relation must be deduced because the utterance does not directly express it. In (9),

9) *Chez les colobinés, le nez fait saillie sur la lèvre supérieure.*

[In colobines, the nose juts out over the upper lip.]

Here, there is a meronymic relation between *nose* and *colobines*. The knowledge about this element is presupposed and not posed by the sentence. In other cases, the same pattern may correspond to another relation or not correspond to any relation at all:

10) *Chez les colobinés, la nourriture…*
[Among colobines, food…]

- Multiple binary relations

From some utterances, it is impossible to build triplets corresponding to the meaning because depending on the discourse, the relationships are interdependent as in:

11) *Each subdivision transmits to CIGT a form related to a complete site*

This sentence corresponds to the syntaxico-semantic structure:

NP1 (*person) communicates* NP2 *(information) to* NP3 *(person)*

In which all the arguments are linked. It is impossible to represent this sentence by a binary-relation or even by several binary-relations.

3-3 Distributional contexts

In fact, few results are obtained using relational patterns, even in didactic texts. An alternative is to use a distributional approach. This approach originated within two schools of linguistics. The first one, based on a behaviorist and mathematical approach, was developed by Bloomfield, then Harris. The second one, based on a sociolinguistic point of view, was developed by Firth. Both are based on the same idea: if you do not know the meaning of a word, you can guess it by examining the contexts in which it appears. Hence, if you identify several contexts (or rather, categories of contexts), you may decide that the word in question has more than one meaning.

“You shall know a word by the company it keeps” (Firth 1957: 11).
“Difference in meaning correlates with differences of distribution” (Harris 1954: 156).

With distributional approaches, it is often necessary to use different contexts, containing what can be called "cues", rather than transparent patterns, in order to build a conceptual relation or a lexical relation such as synonymy or polysemy.
While the relational pattern approach can be considered as top-down because the patterns are described for a language and then projected into the discourse, the distributional approach can be considered as bottom-up because the interpretation is mainly built directly from the textual contexts. Examples (12) and (13) are extracted from a corpus on volcanology. For French readers, the meaning of *are extruded* is not clear in (12). However, as (13) contains some elements close to (12), i.e. *from the vents* as origin argument and *magma*, which can be considered as pertaining to the same paradigm as *lavas* in (12), this orients the interpretation towards a synonymy (or at least a semantic proximity) between *extruded* and *ejected.*

12) Lavas and ash of lavas and ash of granitic composition are extruded from vents
13) Gas-rich magma is ejected from the vent to produce a bomb

From a linguistic point of view, what we can retain from this brief description of TKBs is that this new concept boosted research in textual terminology and, especially, in the systematic exploration of specialized texts.

4. Looking into the Future

At the moment, three observations concerning TKBs can be made.

First of all, TKBs are still built, even if less frequently, and always with the same main characteristics: distinction between terms and concepts, knowledge representation in a network form and role of texts as knowledge source. This is the case for example with Ecolexicon, built in Granada by Faber's team (Faber and Buendia-Castro, 2014).

Second, the original "symbiotic relationship" between terminology and knowledge engineering, evoked by Skuce and Meyer (Skuce and Meyer, 1991), no longer applies. Now, the most widely used term within knowledge engineering to refer to a network representation is *ontology* (see this book, chapter 20). One difference with TKBs is that concepts and linguistic forms are not identified separately. However, the main difference likely resides in the methods used in knowledge engineering to build networks, that are now rarely symbolic (linguistic). Most often, machine learning methods are applied on very large corpora (very often, the entire web) in order to spot new patterns and new triplets. The initial machine learning method, proposed by Hearst (Hearst, 1992), used triplets linked by a known relation (hypernym) and learned new patterns from texts, using the recurrent linguistic forms appearing between the couple of terms. The hypothesis was that these recurrent forms could correspond to patterns of conceptual relations. Then, new triplets were detected by these patterns and, in their turn, projected on the corpus. This recursive method was used and improved in different projects (Agichtein and Gravano, 2000; Buitelaar et Ciminao., 2008). In more fine-grained analyses, belonging to what is named "distributional semantics", the syntactic links between the terms are used, which leads to more precise results from a linguistic point of view (Lenci, 2008). But with machine-learning methods, interpretation of

the relationships is not as fundamental as in TKBs. This is due to the fact that the main aim of the learning is not to build a precise representation of the knowledge, but, rather, to detect enough regularities to assume that some couples of terms have a constant and relevant relationship. In these cases, the most important application is to improve information retrieval. So, the objectives of linguists and those of knowledge engineers moved apart. Nevertheless, some meeting-points still exist that strengthen the collaboration between the two communities. This is the case of the two conferences, "Terminology and Artificial Intelligence" and "Terminology and Knowledge Engineering" that are held every two years.

But, from a purely linguistic point of view and as a result of the studies carried out in order to build TKBs, two main perspectives emerge. The first one concerns the improvement of the methods for studying terms in discourse systematically and the second deals with the use of the terms and their contexts for other objectives than building conceptual networks.

4-1 Terms in discourse

The aim of building conceptual networks from texts, as systematically as possible, has enabled the improvement of terminological analysis methods using the results of natural language processing tools.

This issue has brought lexicology and terminology closer, since, with a fine-grained textual analysis, it is impossible to consider terms as just concept labels (as in the Wüsterian perspective); rather, they have to be considered as words (or word groups) used in specialized corpora.

With such a point of view, many of the analyses conducted for the general lexicon can be adapted for terminology. One of the main consequences of considering terms as words is to re-contextualize them and study them within discourses. Hence, there is no reason to take only nouns into account. On the contrary, other parts of speech and especially verbs, which are generally considered as sentence pivots, can be seen to play an important role. In the building of a TKB, verbs are mainly used in relational patterns (as in (4) and (11)), that is, only for their capacity to link terms. However, verbs may also be terms and, moreover, they may be used to describe specialized nouns in discourse (L'homme, 2002). This idea underlies the Framenet project, which is adapted to terminology description. The Framenet project derives from frame semantic theory (Fillmore et al., 2003) in which verbs play the role of pivot around which arguments are organized. Some projects try to adapt this approach in specialized domains in order to take into account not only paradigmatic relations (as in traditional TKBs) but also syntagmatic ones (Faber, 2015). This is an obvious way to bring terms and the discourses that use them closer together.

But more than anything, what appears most specific to terms in use is, on the one hand, the situation in which the texts (in the broad sense including speaking situations) are produced and, on the other, the aim of the study (either theoretical or applied). One of the unforeseen consequences of the development of methods for building TKBs was that, depending on the needs, TKBs may be different in a given domain. The role of the application then becomes crucial in constructing the TKB. Moreover, methods aimed at building conceptual networks

may be adapted for other types of studies. So, what appeared is that computer-assisted methods defined for building conceptual networks may be adapted to other aims, as shown in the following section. Two main methods have been described in section 3: one based on patterns (a top-down method) and the other based on distributional contexts (a bottom-up method). The same two approaches can be adapted for other aims, as discussed in 4-2.

4-2 Terms as a key for entering the texts.

There are many needs for which the above two approaches may be used. However, while in TKBs the aim is to build a relational network, with other needs, the network (or even just the terms) becomes the starting point of the study. In most cases, needs are linked to the variation of terms (presence or not, frequency variation, variation in meaning) and the study requires detecting the variations and explaining them by situational elements.

To a certain extent, terms may be used as pivots of contexts whose study may lead to results that are relevant for new needs. The adaptation of the methods concerns different aspects:

- The corpus to be analyzed. The corpus must be built in accordance with the aim of the study. In most cases, the corpus is organized in sub-corpora according to the situational element of variation to be compared (time, place, communities, etc.). The sub-corpora are then compared and the linguistic variations are linked with situational variations. So the structuration of the corpus plays a crucial role in the study.
- The distributional contexts. The choice and the interpretation of the relevant contexts of the terms are linked to the aim of the study.
- The top-down contexts. They are also defined (pre-defined) depending on the aim.

Here are two examples of studies using the two different contexts for two very different needs (Condamines et al., 2012).

a- Detecting variations in terms: contexts in order to control knowledge evolution

Generally, one considers that the meaning of terms evolves slowly and that a dictionary remains valid over a number of years. But this is not always the case. In some domains, knowledge evolution may be very rapid, but not only because experts make discoveries in a short time but also because external elements (social pressure) impact the evolution of knowledge. What can be very problematic is that speakers may be unaware of this evolution. We encountered this situation in some CNES (Centre National d'Etudes Spatiales) projects. In order to propose a method to detect knowledge evolution via the study of terms, A. Picton, in her PhD thesis, used the two kinds of contexts occurring with candidate-terms (Picton, 2009). With the top-down approach, she used linguistic patterns such as: *nouveau* (new), *autrefois* (formerly), *est apparu* (appeared), which are directly linked to the idea of novelty and occur in utterances containing terms (or candidate-terms). In such cases, we can consider that speakers are more or less aware of the phenomenon. With bottom-up contexts (distributional ones), speakers are probably much less aware of the variation among corpora. For example, it may concern the nature of verbs of which a term is an argument (see for example the case of

11). Once identified, these variations are submitted to the domain experts who decide if they could be problematic or not.

b- Detecting variations in terms' contexts in order to help to stabilize a neo-discipline

We encountered this case with a project concerning exobiology, the study of life beyond the earth's atmosphere. Four disciplines are involved in the characterization this neo-discipline: astronomy, biology, chemistry and geology. Within the context of a project funded by the CNRS (Centre National de la Recherche Scientifique) we built a corpus organized in four sub-corpora and we detected the terms that were present in at least two of them. The items that came top of the list were *atmosphère, eau, temperature, planète, acide, vie*. Then we analyzed their contexts. With top-down patterns, we detected terms for which speakers are aware of the variation in meaning. We used patterns such as: "as said in...", "it is not the same meaning in…". As for evolution over time, we also examined distributional contexts and we spotted variations according to disciplines. Note that, in this study, distributional variation was not a problem, unlike in the previous study. If experts are made aware of it, it can be very fruitful and help them to better define new concepts in the neo-discipline (Condamines, 2014).

In these two examples, the terms constitute the base of the study and the different contexts in which they appear are interpreted from a specific perspective, contributing to help experts in the situations they face. In both cases, the study used NLP tools (candidate-terms extractors, statistical tools, concordancers or Perl programs). The method was similar but the results, for domain experts and for linguists, were different. In the first case, 17 types of linguistic phenomena (variations linked to the evolution over time) were identified (Picton, 2009). In the second case, 12 types of phenomena were identified.

5. Conclusion

In the 1990s, TKBs represented the convergence between terminology and knowledge engineering. Three elements characterized this new concept: terminological knowledge representation under a conceptual network, the use of corpora as knowledge sources, and the development of tools and methods to systematically spot terms and conceptual relations in corpora. Twenty-five years later, the representational model is still used in both terminology and knowledge engineering and the corpus is still the departure point of study for both disciplines. But the methods and the aims of the two disciplines have evolved. Now, the aim of knowledge engineering is mainly to build ontologies from texts using large corpora. From a linguistic point of view, while TKBs continue to be built (and, especially, conceptual networks), new needs emerged which the systematic study of terms and their contexts can satisfactorily meet. Two kinds of contexts may then be used. The first belong to a top-down approach: linguistic patterns linked to the need are defined a priori and searched in co-occurrence with a (candidate-)term in the corpus. The second ones, belonging to a bottom-up approach, consist in choosing and interpreting the contexts of (candidate-)terms depending on the final aim. Whereas the main initial aim of TKBs was to analyze texts in order to define terms, new aims lead mainly to using terms in order to approach texts. In both cases, the tool-assisted methods are very similar and in both cases, the main issue, from a linguistic point of view, concerns the study of the semantic link between terms and texts.

6. Further Reading

Khurshid A., Rogers, M. (2007) *Evidence-based LSP: translation, text and terminology*, Bern, Berlin, Bruxelles, Frankfurt am Main, New York, Oxford, Wien: Peter Lang
This book presents different studies in LSP, observing language in use through the use of corpora. Part four focusses especially on "Terminology and Knowledge Management". Many domains are taken into account and several applications are described in order to show the different aspects of knowledge management.

Proceedings of Computerm (Computational Terminology) workshop associated to Coling (Computational Linguistics). The workshop has existed since 1998. The proceedings are on line. The workshop brings together Natural language processing researchers and linguists/terminologists around issues such as term extraction, conceptual relations extraction, text mining, or summarization.

Ibekwe-San Juan, F., Condamines, A., Cabré Castellví T. (2007) *Application-Driven Terminology Engineering*. Amsterdam/Philadephia: John Benjamins.
The fundamental role of the application in the design of a terminology is the core issue of this book. It presents various types of applications, seen either as the end use of the terminology (information retrieval, information extraction, competitive intelligence) or as an intermediate step serving the previous ones (dictionaries, the lexicon, taxonomies - namely terminological ressources).

Bowker L., and Pearson, J. (2003) *Working with Specialized Language — A practical guide to using corpora*, London: Routledge. This book is a good introduction for beginners to the study of specialized corpora. It looks at issues such as Corpus design, compilation and processing or Corpus-based applications in LSP.

Temmerman R., Van Campenhoudt M. (2014) *Dynamics and Terminology: An interdisciplinary perspective on monolingual and multilingual culture-bound communication*
Amsterdam/Philadephia: John Benjamins.
This book presents different views on the dynamicity within terminology. Different languages, different disciplines (linguistics, sociology, psychology, ethnology and even language philosophy), different domains and different types of needs are taken into account in this rich panorama.

7. Related Topics
8. Chapter Endnotes
9. References

Agichtein, E., Gravano, L. (2000) 'Snowball: Extracting relations from large plain text collections', *Proceedings, 5th ACM Conference on Digital Libraries*, San Antonio, Texas: 85–94.

Ahmad, K. (1993) 'Terminology and Knowledge Acquisition: A Text Based Approach, in K.D Schmitz (ed.), *Proceeding, Terminology and Knowledge Engineering (TKE 2013)*, Frankfurt: Indeks Verlag: 56-70.

Auger, A., Barrière, C. (eds) (2008) *Terminology*, 14(1): *Pattern based approaches to semantic relation extraction: a state-of-the-art*: 1-19.

Aussenac-Gilles, N., Condamines, A. (2012) 'Variation and semantic relation interpretation: Linguistic and processing issues', in Aguado de Cea et al. (eds.) *Proceedings, Terminology and Knowledge Engineering Conference (TKE 2012)*, Madrid, Spain: 106-122.

Barnbrook, G., Sinclair, J. (2001) 'Specialised Corpus, Local and Functional Grammars', in M. Ghadessy, A. Henry & R.L. Roseberry, (eds), *Studies in Corpus Linguistics*. Amsterdam/Philadephia: John Benjamins, 237-75.

Buitelaar, P., Cimiano P. (2008) *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, Amsterdam: IOS Press.

Cabré, M.-T., (1999) *Terminology. Theory, Methods and Application*, Amsterdam/Philadelphia: John Benjamins.

Condamines, A., (1995) 'Terminology : New needs, new perspectives', *Terminology*, 2(2) : 219-238.

Condamines, A. (2000) '*Chez* dans un corpus de sciences naturelles: un marqueur de méronymie?'. *Cahiers de Lexicologie* 77(2): 165-187.

Condamines A. (2002) 'Corpus Analysis and Conceptual Relation Patterns', *Terminology*, 8 (1): 141-162.

Condamines, A. (2005) 'Anaphore nominale infidèle et hyperonymie: le rôle du genre textuel', *Revue de Sémantique et Pragmatique* 18 : 23-42.

Condamines, A. (2014) 'How Can Linguistics Help To Structure A Multidisciplinary Neo-Domain Such As Exobiology', in *Bioweb of conferences*. http://dx.doi.org/10.1051/bioconf/20140206001

Condamines A., Amsili P. (1993) 'Terminology between Language and Knowledge: an example of Terminological Knowledge Base', in K.-D. Schmitz (ed.), *Terminology and Knowledge Engineering* (*TKE 199),* Frankfurt: Indeks Verlag: 316-323.

Condamines, A., Rebeyrolle, J. (2001) 'Searching for and Identifying Conceptual Relationships via a corpus-based approach to a Terminological Knowledge Base (CTKB) : method and results', in D. Bourigault, M.C. L'homme, C. Jacquemin (eds), *Recent Advances in Computational Terminology,* Amsterdam/Philadelphia: John Benjamins: 127-148.

Condamines, A., Dehaut, N., Picton, A. (2012) 'Rôle du temps et de la pluridisciplinarité dans la néologie sémantique en contexte scientifique. Études outillées en corpus', C. Gérard and J. Kabatek (eds.), *Les Cahiers de Lexicologie* n°101, *Néologie sémantique et analyse de corpus*: 161-184.

Cruse D.A. (1986) *Lexical Semantics.* Cambridge: Cambridge University Press.

Cruse, D.A. (2002) 'Hyponymy and its Varieties', in R. Green, C.A. Bean & S.H. Myaeng (eds) *The semantics of relationships: an interdisciplinary perspective*, Dordrecht/Boston/London: Kluwer Academic Publishers: 3-21.

Drouin, P. (2003) 'Term extraction using non-technical corpora as a point of leverage', *Terminology*, 9(1): 99-117.

Faber, P. (2015) 'Frames as a framework for terminology', in Kockaert, H.J. & Steurs, F. (eds), *Handbook of Terminology*. Amsterdam/Philadelphia: John Benjamins: 14-33.

Faber, P., Buendía Castro, M. (2014) 'EcoLexicon', in Andrea Abel, C.V. & Ralli, N. (eds) *Proceedings of the XVI EURALEX International Congress*, Bolzano: 601-607.

Fillmore, C.J., Johnson C., R., Petruck, M. R. (2003) 'Background to FrameNet' *International Journal of Lexicography* 16(3): 235-250.

Flowerdew, J. (1992) 'Definitions in Science Lectures', *Applied Linguistics*, 13(2), Oxford University Press: 203-221.

Frantzi, K., Ananiadou, S., Mima, H. (2000) 'Automatic recognition of multi-word terms', *International Journal of Digital Libraries*, 3(2): 117-132.

Firth J.R, (1957) *Papers in Linguistics 1934-1951*. Oxford University Press. (first edition: 1957).

Gaudin, F. (1990) 'Socioterminology and expert discourses', *Proceedings Terminology and knowledge engineering (TKE 1990):* 631-641.

Green, R., Bean, C.A., Myaeng, S.-H (eds.) (2002) *The semantics of relationships*, Dordrecht/Boston/London: Kluwer Academic Publishers.

Gross, M. (1997) 'The Construction of Local Grammars', in E. Roche & Y. Schabes (eds). *Finite-State Language Processing*, MIT Press: 329-354.

Harris, Z. (1954) 'Distributional Structure', *Word* 10 (23): 146-162.

Hearst, M.A. (1992) 'Automatic Acquisition of Hyponyms From Large Text Corpora', In *Proceedings 14th International Conference on Computational Linguistics*, Nantes, France: 539-545.

Lenci, A. (ed.) (2008) 'From context to meaning: distributional models of the lexicon in linguistics and cognitive science', *Italian Journal of Linguistics*, 20(1): 1-31.

Lehrberger, J. (1986) 'Sublanguage Analysis', in R. Grishman and R. Kittredge (eds), *Analyzing Language in Restricted Domains*, Hillsdale, New Jersey, London: Lawrence Erlbaum Associates Publishers: 19-38.

L'Homme, M.-Cl. (2002) 'What can Verbs and Adjectives Tell us about Terms?', *Proceedings, " Terminology and Knowledge Engineering"(TKE 2002)*: 65-70.

L'Homme, M.-C., Marshman, E. (2006) 'Terminological Relationships and Corpus-based Methods for Discovering them: An Assessment for Terminographers' In L.Bowker, *Lexicography, Terminology, and Translation. Text-based studies in honour of Ingrid Meye*r, Ottawa: University of Ottawa Press: 67-80.

Lyons, J. (1977) *Semantics: Volume 1*, Cambridge: Cambridge University Press.

Marshman E., L'Homme M.-C. & Surtees V. (2008) 'Portability of cause-effect relation markers across specialized domains and text genres: A comparative evaluation', *Corpora*, 3(2), pp.141-172.

Meyer I. (2001) 'Extracting Knowledge-rich Contexts for Terminography: A Conceptual and methodological Framework', in D. Bourigault, M.C. L'homme, C. Jacquemin (eds), *Recent Advances in Computational Terminology*, Amsterdam/Philadelphia: John Benjamins: 279-302.

Meyer I., Bowker L., Eck K., (1992) 'Cogniterm: An Experiment in Building a Terminological Knowledge Base', *Proceedings 5th EURALEX International Congress on Lexicography*, Tampere, Finland: 159-172.

Parent R., (1989) 'Recherche d'une synergie entre développement linguistique informatisé et systèmes experts : importance de la terminologie', *Meta*, 34-3 : 611-614.

Pearson, J. (1996) 'The expression of definitions in specialised texts: a corpus-based analysis', in J.J.M. Gellerstam, S.-G. Malmgren, K. Norén, L. Rogström, C. Röjder Papmehl (eds), *Proceedings, Seventh EURALEX International Congress on Lexicography*, Vol. Part II, Göteborg, Sweden: Göteborg University, Department of Swedish: 817-824.

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

Picton, A. (2009) *Diachronie en langue de spécialité. Définition d'une méthode linguistique outillée pour repérer l'évolution des connaissances en corpus*. *Un exemple appliqué au domaine spatial*, PhD Dissertation, Université Toulouse Le Mirail, France.

Sinclair, J. (1991) *Corpus, concordance, collocation*: *Describing English language*, Oxford: Oxford University Press.

Skuce D., Meyer I., (1991) 'Terminology and Knowledge Engineering: Exploring a Symbiotic Relationship', *Proceedings 6th International Workshop on Knowledge Acquisition for Knowledge-Based Systems* (Banff): 29/1-29/21.

Temmerman, R. (2000) *Towards New Ways of Terminology Description. The Sociocognitive Approach*, Amsterdam/Philadelphia: John Benjamins.

Wijnands P. (1993) 'Terminology versus artificial intelligence', in H.B. Sonneveld, K.L. Loening (eds), *Terminology, applications in interdisciplinary communication*. Amsterdam/Philadelphia: John Benjamins: 165-179.

Winston, M.E., Chaffin, R., Hermann, D. (1987) 'A taxonomy of part-whole relations', *Cognitive Science*, 11(4): 417-444.

Wüster, E. (1974) 'General Terminology Theory - Fine Line between Linguistics, Logic, Ontology, Information Science and Business Sciences.' *Linguistics* (119): 61-106.

Wüster, E. (1979) *Introduction to the General Theory of Terminology and Terminological Lexicography*. Springer, Wien.