SURFACE GRAMMATICAL ANALYSIS FOR THE EXTRACTION OF TERMINOLOGICAL NOUN PHRASES

Didier BOURIGAULT

Ecole des Hautes Etudes en Sciences Sociales
et
Electricité de France
Direction des Etudes et Recherches
1, avenue du Général de Gaulle
92141 Clamart Cedex
France

Tel : +33 1 47 65 50 64

ABSTRACT

LEXTER is a software package for extracting terminology. A corpus of French language texts on any subject field is fed in, and LEXTER produces a list of *likely terminological units* to be submitted to an expert to be validated. To identify the terminological units, LEXTER takes their *form* into account and proceeds in two main stages : *analysis, parsing*. In the first stage, LEXTER uses a base of rules designed to indentify *frontier markers* in view to analysing the texts and extracting maximal-length noun phrases. In the second stage, LEXTER parses these maximal-length noun phrases to extract subgroups which by virtue of their grammatical structure and their place in the maximal-length noun phrases are likely to be terminological units. In this article, the type of analysis used (*surface grammatical analysis*) is highlighted, as the methodological approach adopted to adapt the rules (*experimental approach*).

1) Constituting a terminology

Constituting a terminology of a subject field, that is to say establishing a list of the *terminological units* that represent the concepts of this field, is an oft-encountered problem. For the Research Development Division of Electricité de France (French Electricity Board), this problem arose in the *information documentation* sector. An automatic indexing system, using different thesauri according to the application, has been operational for three years or more [Monteil 1990]. The terminologists and information scientists need a terminology extraction tool in order to keep these thesauri up to date in constantly changing fields and to create "ex nihilo" thesauri for new fields.

This is the reason why the terminological extracting software, LEXTER, was developed, forming the first link in the chain that goes to make up the thesaurus. *A corpus of french-language texts is fed into LEXTER, which gives out a list of likely terminological units, which are then passed on to an expert for validation.*

2) What is a terminological unit ?

The main aim here is not to provide a rigorous definition of what a terminological unit is, but rather to outline its essential features, and thus to justify the hypotheses (concerning the *form* of terminological units) on which LEXTER is based.

Semantic function : the representation of the concept

The first characteristic of the terminological unit is its *function as the representation of a concept.* The terminological unit plays this role of representation in the framework of a terminology, which is the linguistic evidence of the organisation of a field of knowledge in the form of a network of concepts; the terminological unit represents a concept, uniquely and completely, taken out of any textual context. The existence of this one-to-one relationship between a linguistic expression and an extra-linguistic object is, as we shall see, a situation which particulary concerns the terminological units.

The appearance of a new terminological unit is most often a parallel process to that of the birth of the concept which it represents. This "birth" is marked by the consensus of a certain scientific community. This consensus is attested only when the *occurrences* of this linguistic expression, or term-to-be, shows a stable correlation to the same object in the subject field, uniquely and completely, in the writings of the agents of this scientific community. When this is the case, the object in question takes its place in the network describing the subject field, and the expression takes on the status of a terminological unit. This referential function is, for E. Benveniste, the "*synaptic*" mark of a syntagm [Benveniste 1966].

It is thus because occurrences in text of a terminological unit systematically refer to a concept, that a relationship of representation is established, *out of any textual context,* between the terminological unit and the concept. This underpins the specific status of the terminological unit as opposed to that of the word in language, a status close to that of a descriptor in Information Science ([Le Guern 1984]).

Syntactic form : synaptic composition

We put forward the hypothesis that this function of representing the concept *out of context* puts a certain number of constraints on the form that terminological units may take on. It has been seen that the construction of terminological units obey well-known rules of syntactic formation, called synaptic composition ([Benveniste 1966]). For example : terminological units are noun phrases, generally made up of nouns and adjectives, and pratically never containing conjugated verbs; the prepositions used most often are "de" and "à", rarely followed by a determiner.

To illustrate this, take the concept of a "screen belonging to a portable computer". Without going in to the linguistic phenomena behind this, it can be said that, in context, both the syntagms "l'écran d'un ordinateur portable" ("the screen of a portable computer") and "un écran d'ordinateur portable" ("a portable computer screen") can refer generically to the concept. However, if one wished to represent this concept out of any textual context, the chances are that one would reject the expression "écran d'un ordinateur portable", for wich the interpretation of the article "un" may be ambiguous, and accept the expression "écran d'ordinateur portable", more naturally used in isolation and thus more suitable to go into a terminology.

From these considerations on the form and the function of terminological units, two mains ideas are relevant to developing a computer based system of terminology constitution :

1) It is possible to devise an extraction program solely based on syntactic data, since the grammatical form of terminological units is relatively predictable;

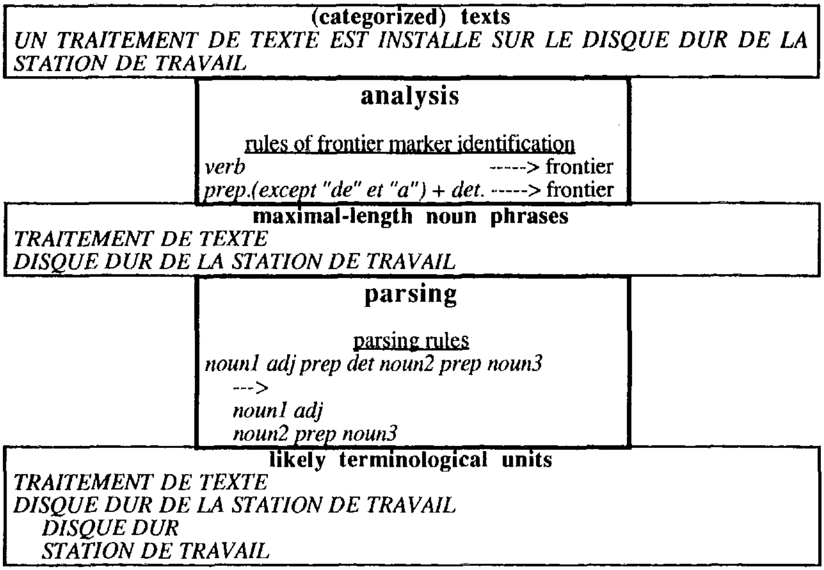
2) It is not possible to expect this program to extract terminological units *and nothing else,* given the basically referential semantic function of occurrences of terminological units : this means that the results obtained can only be considered, a priori, as *likely* terminological units.

3) How LEXTER works : analysis and parsing

To detect terminological units, LEXTER takes the form of these units into consideration, and works in two phases : *analysis* and *parsing*.

LEXTER treats *categorized* texts, which have been submitted to a morphological analysis : each word is tagged with its grammatical category (noun, verb, adjective, etc.).

Figure 1 : Simplified example of how LEXTER works



(categorized) texts

UN TRAITEMENT DE TEXTE EST INSTALLE SUR LE DISQUE DUR DE LA STATION DE TRAVAIL

analysis

rules of frontier marker identification

verb -----> frontier

prep.(except "de" et "a") + det. -----> frontier

maximal-length noun phrases

TRAITEMENT DE TEXTE

DISQUE DUR DE LA STATION DE TRAVAIL

parsing

parsing rules

noun1 adj prep det noun2 prep noun3

--->

noun1 adj

noun2 prep noun3

likely terminological units

TRAITEMENT DE TEXTE

DISQUE DUR DE LA STATION DE TRAVAIL

DISQUE DUR

STATION DE TRAVAIL

First step : analysis by identification of frontiers

At this stage, LEXTER takes advantages of "negative" knowledge about the form of terminological units, by identifying those grammatical patterns which never go to make up these units and which can thus be considered potential terminological limits. Such patterns are made up by, say, conjugated verbs, pronouns, conjonctions, certain strings of preposition + determiner, etc.

The LEXTER analysis module is thus set up with a base of rules for identifying *frontier markers*, which it uses to analyse the texts. This analysis phase produces a series of text sequences, most often noun phrases. The way the rules are worked out is presented in section 5.

These noun phrases may well be likely terminological units themselves (as is the case with *TRAITEMENT DE TEXTE*, in the example in figure 1), but more often still, they contain subgroups which are also likely units (such as *DISQUE DUR DE LA STATION DE TRAVAIL*, which contains *DISQUE DUR* and *STATION DE TRAVAIL*). That is why it is preferable at the analysis stage to refer to the noun phrases identified as "maximal-length noun phrases".

Second stage : parsing the maximal-length noun phrases

It is thus necessary, in the second stage, to parse these maximal-length noun phrases in order to obtain subgroups which are likely terminological units by virtue of their *grammatical structure* and their *position* in the maximal-length noun phrase. The LEXTER parsing module is made up of parsing rules which indicate which subgroups to extract on the basis of grammatical structure. An example of a rule is given in figure 1. In its present form, the parsing module can recognize up to 800 different structures, enabling it to treat around 95% of the maximal-length noun phrases obtained from our test corpus on completion of the analysis stage, that is around 43,500 groups out of 46,000. This module, the core of LEXTER, is described more fully in [Bourigault 1992b].

4) Surface grammatical analysis versus complete syntactic analysis

At the beginning of the conceptual phase in the development of LEXTER, it was hypothesized that a complete syntactic analysis of the sentences of the corpus could be foregone, given the limited aim of extracting terminological noun phrases with their characteristic grammatical structure.

In LEXTER, the basic linguistic data is the *grammatical categoty* of the lexical units which make up the sentences, and the analysis and parsing which make use of this data take into account the surface form of utterance considered as sequences of categorized units : only the "place" of the units in the surface sequence is taken into account and not their position (cf [Milner 1989]) in a syntactic structure. This is why it is more accurate to speak of a *surface grammatical analysis* than of a complete syntactic analysis.

The quality of the results obtained by the present prototype justify the non-necessity of a complete syntactic analysis. The advantages of restricting the analysis to surface structure are obvious : the program can deal with texts written in styles that are not necessarily academic; it is sturdy and quick, not negligible virtues when it come to the development and extension stages.

Although it is not necessary to go into a complete syntactic analysis of the sentences to extract the terminology from a corpus, it would seem highly likely that a syntactic analyser (parser) would be much more efficient if it could use a glossary of the terminological units of the subject area. The syntactic structures of a natural language text, and the syntactic structures of the terminological units, representing out of context, in a terminology, the concepts of a subject field, are to be placed on two different organisational levels. It is thus advisable to dissociate these two analysis, though using the results of one for the other. Since the terminological unit, as its name suggests, is always a *semantic unit*, which is at the basis of its status (cf §1), it should be treated as such on the syntactic level as well.

This makes it possible to envisage a text analysis in two phases, the first identifying terminological units, the second using these results to analyse the sentences syntactically, with the view to constructing a semantic representation. This is the principle which we intend to adopt to make LEXTER a text analysis tool to aid knowledge acquisition (cf [Bourigault 1992a]).

5) An experimental approach to work out rules of analysis

To analyse texts, LEXTER uses an analysis rule base which detects frontier markers. Some of these rules are simple : one of them detects all punctuation marks; another all the words belonging to certain grammatical categories : verbs, conjonctions, pronouns, etc.

As well as theses simple rules, it is necessary to add more complex rules which examine sequences of lexical units to find frontiers, in particular to spot the boundaries between noun phrases that are complements of the same verb or the same noun.

The constraint imposed by the choice of a surface grammatical analysis make it difficult to base the detection of frontiers between noun phrases on reliable theorical morphosyntactic hypothesis (even though the works of F. Debili showed that this choice, for french language, is pertinent for computer processing [Debili 1982]). This is particulary so for the semantic-syntactic type of lexical information for the subcategorization of verbs (nouns and adjectives as well) which must be foregone, making even more difficult the tricky task of identifying to what prepositional noun phrases are attached.

The alternative is then to rely on intuitive ideas and to compensate for the absence of theorical justification by adopting an empirical approach based on large-scale corpus experimentation. Before any rule is put into one of LEXTER's modules (rules of analysis, rules of parsing), it must pass the test of the results it produces every time it is applied to the test corpus.

This is why it is necessary to work on a test corpus of sufficient volume to be representative of possible cases of analysis and parsing, and to produce a sufficiently fast working software to make this experimental approach worthwhile. For this, a test corpus of 2 5000 pages (arround 1 200 000 words) was used, gathered from 1 700 texts, in which the scientific officers of the Research Development Division of Electricité de France describe in a short paper (1 or 2 pages) each of their medium term research projects. The analysis and parsing module of LEXTER were programmed in *C*, using *lex* and *yacc* tools in a *Unix* environment. Each of the stages of analysis and parsing is *less than 2 minutes* on a Sun work station, making very frequent tests easy and thus enabling far reaching updating and ajustment.

It is through an experimental approach that the analysing (and parsing) rules were worked out. By way of illustration, the analysis rules treating the sequences *preposition + determiner* are presented in figure 2.

Figure 2 : Analysis rules for the sequences : *preposition + determiner*

	ø	*LE*, *LA* or *LES*	*UN*, *UNE* or others
DE or *A*	_	_	frontier
others	_	frontier	frontier

It is true too that these rules, as all the rules in LEXTER, have their limits and that there are cases where they apply (or do not apply) "wrongly". These limitations come from the strong hypotheses and the methodological choices which have already been outlined. But in the field of Linguistic Engineering, exceptions do not have the same status as in Linguistic Science; it is here a question of *compromise.*

In the experimental approach adopted, this risk of error is taken into account and kept under control, as each rule is tested separately against the corpus, and it is the test of the number of cases to which it applies which decides whether it gets into LEXTER or not. The principle is not to include rules of analysis which are too strict; it is preferable to drop a rule which is productive in many cases (as for the rule A + *LE, LA* or *LES* = frontier) if the number of residual cases of erroneous analysis is too high. This principle, called "of relative strictness", is justified in that it will be easier for the terminologist to eliminate certain likely units than to find real terminological units that escaped detection by LEXTER

6) References

[Benveniste 1966] **Benveniste Emile** (1966) "Formes nouvelles de la composition nominale", in *Problèmes de linguistique générale*, Tome 2, PP 163-176, Gallimard, Paris

[Bourigault 1992a] **Bourigault Didier** (1992) "LEXTER, vers un outil liguistique d'aide à l'acquisition des connaissances", *Actes des 3èmes Journées d'Acquisition des Connaissances*, Avril 1992, Dourdan

[Bourigault 1992b] **Bourigault Didier** (1992) "LEXTER, un logiciel d'extraction de terminologie", *Actes du symposium TAMA 92*, Juin 1992, Avignon

[Debili 1982] **Debili Fathi** (1982), "Analyse syntaxico-sémantique fondée sur une acquisition automatique de relations lexicales sémantiques", Thèse d'état, Orsay

[Le Guern 1984] **Le Guern Michel** (1984), "Les descripteurs d'un système documentaire. Essai de définition", *Actes du colloque "Traitement automatique des langues naturelles et systèmes documentaires"*, Clermont-Ferrand

[Milner 1989] **Milner Jean-Claude** (1989), "Introduction à une science du langage", Seuil, Paris

[Monteil 1990] **Monteil Marie Gaelle, Pénot Nadine** (1990), "Indexation Automatique, fonctionnement - Principes généraux", *Note interne HN46464*, EDF, Direction des Etudes et Recherches, Service IPN, Clamart