Graph based system purpose – built for automatic retrieval and extraction of the electronics data.

Dziczkowski Grzegorz
Ecole des Mines de Paris
35, rue Saint-Honore
77305 FONTAINEBLEAU
ESIGETEL
grzegorz.dziczkowski@esigetel.fr,

Wegrzyn-Wolska Katarzyna
Ecole Superieur d'Ingenieurs en Informatique et Genie
des Telecommunicatiom (ESIGETEL)
1,Rue de Port de Valvins
77-215 Avon-Fontainebleau Cedex
katarzyna.wegrzyn@esigetel.fr

ABSTRACT
The work presented in this paper concerns automatic information retrieval and extraction in precise field of electronics based on linguistic knowledge. The extraction and the filtering of data is carried out automatically by methods based on the construction of local grammar. To carry out an automatic search for the events in the corpus, a linguistic base, for example the base of the graphs, was created. A comparison of methods is given.

KEY WORDS
natural language processing, pattern matching, lexicon-grammar, syntactic parsing, syntactic grammars, recursive transition networks (RTN).

1. Introduction

The field of study presented in this paper is the filtering and extraction of information. The extraction of information consists in the identification of quite precise information in a text in natural language and its representation in a structured form [1]. It consists of information retrieval, which aims at finding a group of relevant documents to the query in a corpus [2].
The filtering and extraction requires lexicons and specialized grammar. The development of such resources is a long and tiresome task, which generally requires an expertise on the field approached and a knowledge in data-processing linguistics like techniques of filtering, categorization of documents and extraction of information.
Systems of text comprehension, for the majority, have been designed as generic comprehension systems, but they appear unlikely to be used in real applications. Comprehension is seen as a transduction which transforms a linear structure, i.e. text (the linear structure) is transformed into an intermediate logico-conceptual representation, which is then used to make conclusions. In our case the conclusions are the question answers.
To understand the meaning of the text, it is necessary to carry out syntactic and the semantic analysis. Syntactic analysis is the largest task due to ambiguities. The semantic analysis aims to produce a structure representing as accurately as possible, an unit of the sentence, with its meanings and its complexity; then it has to integrate all structures into a single textual structure. At the end, we obtain a logico-conceptual representation of the text. The semantic representation varies from one system to another. We end up, in the system "Core Language Engine", with forms known as logically inspired partly by the grammar of Montague [3]. In the system "Kalipsos", the semantic representation is carried out by the conceptual graphs [4] whereas in the system "Acord", we end up with structures of discursive representation [5]. The semantico-conceptual structures can be more or less broad, rich and complex and more or less ambiguous. The adaptation of these systems poses the traditional problem of the re-use of the systems and the bases of knowledge which they integrate. The adaptation of a new task to a new field requires the rebuilding of a great part of the knowledge bases, particularly in the semantic lexicon.
The objective of our work is the automatic extraction of the information concerning electronic field, from any text according to user query. Our method uses linguistic rules for automatic data analysis and extraction. It is evident, that this linguistic rule depends of the selected language. Our work is done now only for the Polish language. We use Unitex as a corpus analyzer.

2. Previous Work

The relative failure of the generic systems comprehension is well-known today. It should however be recalled that these systems resulting from work of automatic treatment of the languages of years 1980 really made it possible to explore this generic approach of the comprehension of text. Researchers are trying to develop relatively complete syntax and semantics dictionaries.
This pushed large numbers of researchers to describe the natural languages in the same way as formal languages. Maurice Gross [6] undertook with his team of the LADL[1] the exhaustive examination of simple sentences of French,

[1] French Laboratory for Linguistics and Information Retrieval

in order to have reliable and quantified data on which it would be possible to make rigorous scientific experiments. For that, each verb was studied in a way as to test if it responded or not to syntactic properties like the fact of admitting a noun clause in a position subject. 6000 verbs were examined using approximately 300 properties. The result is that for 6000 verbs, we have approximately 15000 different uses and syntactic behaviour. It is obviously realized that French can't be described with general rules: the same situation applies to all the other languages include Polish. The results of this study were coded in matrix called lexicon-grammar tables. These tables show a precise description of the syntactic behaviour of each verb of French. The objective is to use all the resources of the lexicon-grammar tables to obtain a system able to analyze any structure of simple sentences.
The minimal unit of direction, according to Maurice Gross, is the sentence, and not the word. The principle is thus to study the transformations of the simple sentences. The simple sentences were indexed by their verbs. For a verb we can have several different uses. Syntactic properties allow to distinguish the uses of a verb. Each verb has a unique characteristic so there are not two verbs having exactly the same syntactic behaviour. Thus we can't formulate general rules to explain any language.
To exploit the linguistic knowledge an application Unitex[2] was created at LADL by Mr. Sebastien Paumier under the direction of Mr. Maurice Gross. The application is based on linguistic tools like AGLAE [7] and INTEX [8]. Unitex [9], [10] is an environment of enhancement used to build formalized descriptions to broad coverage of the natural languages and apply them as texts of important size in real time. Descriptions of the natural languages are formalized as numeric dictionaries, grammars represented by graphs and lexicon-grammars. Unitex makes it possible to treat in real time the texts of several mega-bytes for the indexing of morpho-syntactic reasons, the search for set phrases or semi-fixed phrases, the production of agreements and the statistical study of the results.

2.1 Linguistic resource.

The linguistic resource to achieve the information retrieval and extraction are as follows:

- Dictionaries
- Networks of the recursive transitions (local grammar)
- Tables of lexicon-grammar

2.1.1 Dictionaries

The digital dictionaries employed by Unitex use formalism of DELA[3]. Numeric dictionaries describe both the simple words and the complex words of a language. Dictionaries associate the word with a lemma and a series of grammatical, semantical and inflexional codes.

2.1.2 Networks of recursive transitions (local grammar)

Grammar is a representation of linguistic phenomena by recursive transitions (RTN), a formalism close to that of the finite state automaton. Many studies have highlighted the adequacy of automats on linguistic problems. A transducer with a finite number of states is a graph which represents a whole of entry sequences, and associates sequences produced as an output. Generally a grammar represents sequences of words and produces linguistic information like the information on the syntactic structure.
A local grammar [6] is an automaton representation of the linguistic structures witch is difficult to formalize in lexicon-grammar tables or numeric dictionaries. The local grammars, represented in the forms of graphs, describe elements which concern the same syntactic or semantic field. The linguistic descriptions grouped together in the form of local grammars are used for a large variety of automatic processes applied to the text. Thus various methods of lexical clarification were developed to implement grammatical constraints described before using this type of graph.
The corpora of text are represented by automats, in which each state corresponds to a lexical analysis. The linguistic phenomena are represented by local grammar, and are then translated into finite state automaton in order to be easily confronted with the corpora of text [11].

2.1.3 Tables of lexicon-grammar

Tables of lexicon-grammar are matrixes that outline the properties of all the simple verbs which are described by syntactic properties. Each word having an almost unique behavior, the tables give the grammar of each element of the lexicon, which is why they are called lexicon-grammar tables. With Unitex we can build grammar from such tables. The lexicon-grammar is a systematic description of the syntactic and semantic properties of the syntactic functors of French that is predicative verbs, nouns and adjectives. It is organized in groups of tables, which are associated with the syntactic category like full verbs, verbs supports, names, etc... A table corresponds to a particular syntactic construction and gathers all the words entering this construction. Currently lexicon-grammar is especially developed for the verbs and the predicative phrases. This lexicon currently contains 15.000 entries of simple verbs. Moreover, 25.000 predicative phrases were described, and 20.000 phrases built with *etre* (en: to be) or *avoir* (en: to have) [12][13].

3. Our Approach

In the preceding chapter, we presented the possibility of carrying out an effective search in any text using

[2] Unitex – a corpus processing system, based on automata – oriented technology. (www-img.univ-mlv.fr/~unitex/)
[3] Dictionaries Electronic of the LADL

linguistic knowledge. In the Unitex application, we can seek the correspondences by using graphs, the dictionaries and grammars. To find the correspondence with the request that user wishes to apply, he is obliged to build the graphs by himself. To build graphs, a knowledge of the linguistic bases is necessary. To facility the search task we improved the application so that research is automatically created.
In an automatic research, the user has to choose only the corpus he wants to parse as well as the request with the required words. Automatic research is carried out for a quite precise field. The field considered is electronics for which we generated a base of graphs.

3.1 Graph based system for automatic retrieval and extraction

To use the analyzer Unitex with a new language, a procedure for processing is obligatory. It is necessary to specify the alphabet, the dictionary, the graphs, and the corpus,. It might be possible that we observe the grammatical rules to have the inflected terms of canonical forms of the dictionary. One of our task was to add the Polish language to Unitex application. To perform the automatic search and to add Polish language the following steps are necessary:

- integration of the Polish language in the Unitex software
- adaptation of the dictionary to cover the maximum number of the words of the field considered
- creation of the graphs able to seek the most interesting events

To be able to seek in the text the occurrences of the words requested by the user, we proceed in the following way:

- choice of the various parameters to carry out a research via the user interface. These parameters are: the input of the text, the choice of the method of research and the request
- to find the corresponding graph starting from the request of the user. Two different methods enable us to establish a link between graphs and user requests
- creation of the graphs corresponding to the request if, for the request, no graph of the base is appropriate
- generation of results.

For the words of the dictionary belonging to the field considered, the correspondences with the graphs are carried out by the semantic classes. For that, a division of the field was obligatory. The semantic classes allow the possibility of binding the graphs with the user requests.
We divided the electronic field into several parts. All the fields are represented in the form of a tree. This construction can be visualized on figure 1. For each branch, we associate a semantic class. One word can belong to several semantic classes, as for example a photodiode is an optoelectronic component, but has however the characteristics of a diode which is located in the category of the semiconductors. We created the base of the graphs using the tree structure.

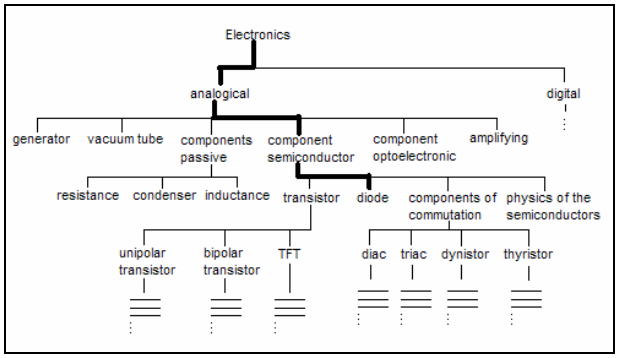

Figure 1 - Parts of electronic field

When the categorization of the field is done, we can associate the semantic classes with the graph. It is possible that one word has several semantic classes. Figure 2 presents the example of a simple graph for a diode category:

Electronics → analogical → semiconductor components → diode.

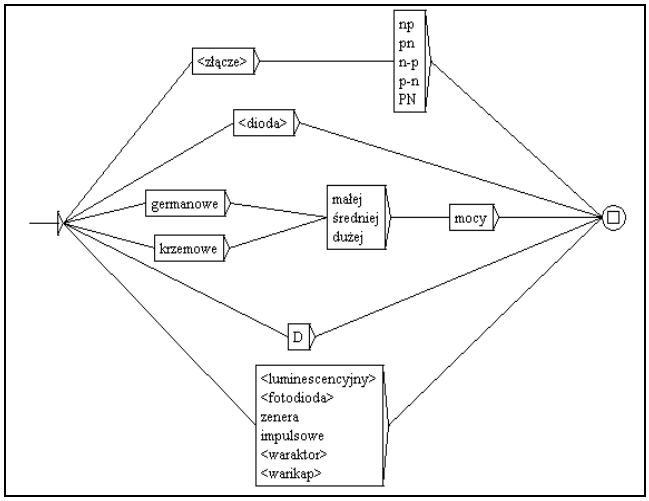

Figure 2 - Example of graph for the "diode" term.

In the graph, we have all the possible synonyms of each word, all the elements having similar parameters, the components which form the same family, etc... One graph can also call another graph.
The application associates the graphs with the request words. Unfortunately, it is not possible to cover the whole field of electronic with the graphs, even if we can note that this field is relatively "closed". In this case, we have the possibility to generate the graph, starting directly from the user request. The generation of this graph is very simple and far from being ideal. For this reason the use of a graph base is recommended. If any graph isn't found, we use the automatic generation of a graph. This method

consists in finding all the words requested by the user which are in the dictionaries covering the words of the field. After finding the words of a field, the algorithm creates the graph. We will take care only of the words which depend on the electronic field. We suggest that these words are more interesting for the user. Here is a simple example to clarify the idea:

The request of the user:

When electric and magnetic waves are at 90 degrees to each other, if I rotate the source of the waves by 90 degrees, would electric and magnetic waves interfere each other?

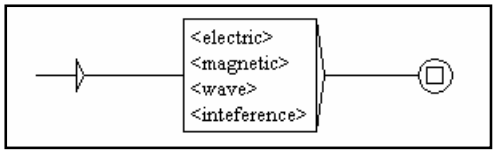

Figure 3 - Example of graph generated

If there is no graph for the magnetic waves the only words which will be taken into account are the words from electronic field. In this case, the graph will be as shown on figure 3.

3.2 Finding the corresponding graph.

We present three different methods to establish a link between the graph and user request.

3.2.1 Method of the semantic classes.

In the method of semantic classes we have correspondences between graphs and key words on the one hand, and entries of the dictionary and key words on the other hand; starting from the words of the request, we select the corresponding key words and then the graphs. Key words are called semantic classes. Searching process is presented of figure 4.

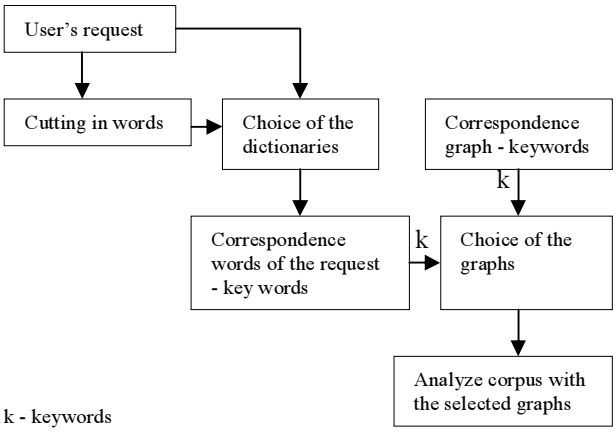

Figure 4 - Method of the semantic classes

We take all the dictionaries used for the current language. For each word from the request of the user, we parse all the dictionaries to find the dependent semantic classes attached to the required word. By the association of the graphs with the semantic classes we recover the graphs which were selected, to find information concerning the request. The correspondence enters the graphs and semantic classes are as follows:
< name of the semantic class 1> < names of the graphs> <name of the semantic class 2> < names of the graphs>.
We outline research in the corpus by using the graphs selected.
At the end of the procedure we obtain all the events found in the corpus chosen with the selected graphs of the base of graph. We hope to find good results given the construction of the graph base. To have satisfactory results, we must have the most complete and detailed graph base. The dictionary must contain the semantic classes detailed for all the words of the field.

3.2.2 Direct method.

In the direct method starting from the text of the request, we select the graphs by the intermediary of the lemmas or the inflected terms. If a word of the request has the same lemma or the same inflected term as a word in one of the graphs, we select this graph as the suitable graph for the analysis of the corpus. We don't take into account all the words of the request but only the words of the field. Searching is carried out in the following way [Fig 5].

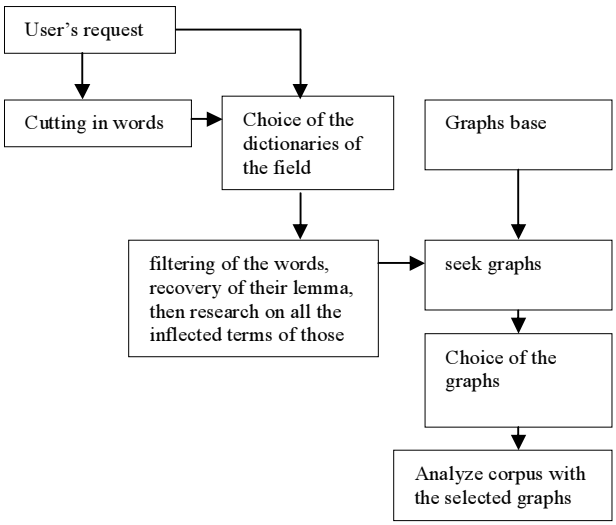

Figure 5 - Direct method

First we take the dictionary of the words corresponding to the electronic field. For each word of the request we parse this dictionary to find its lemma like all the inflected terms. We seek in all the graphs having the forms found previously - the lemma or the inflected terms. When the form is found, we take the graph as the adequate graph for research. We outline research in the corpus by using the graphs selected.

3.3 The automatic construction of the graphs

The third method consists in the recreation of graphs starting directly from the user's request. This method is recommended if no graph corresponds to the user request. We can't let as a result that no occurrence was found other it exists, due to a lack of graphs in our graph base. The creation of graph starting from the user's request takes into account only the words of the selected field. We supposed that they are the most significant words for research. The format of the user's request is not formalized. We must filter them to recover only the relevant words compared to the selected field. The construction of graphs consists in finding the lemma of all the words of the request which are in the dictionary of the field. The searching process is presented of figure 6.

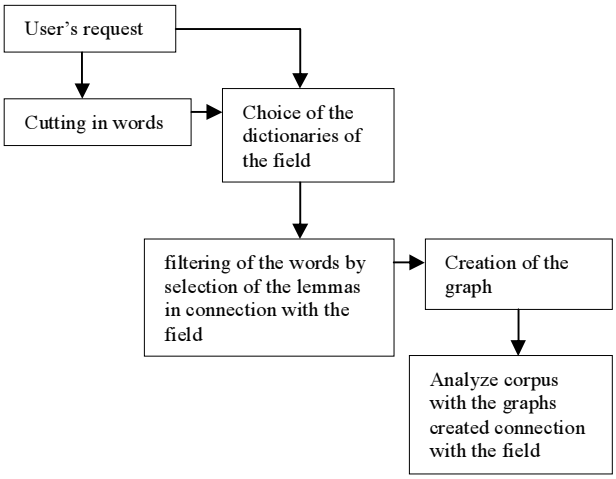

Figure 6 - Automatic constructions of the graphs

We take the dictionary of the words corresponding to the electronic field. For each word of the request we parse this dictionary to find the lemma. We added all lemmas found in the graph. We carried out searches in the corpus by using the graph created; search is thus carried out on all the lemmas, like all their inflected terms.

3.4 The system output

Figure 7 presents the results obtained by method of the semantic classes. An example of the text treatment with the user's request in the Polish language: *jakie sa rodzaje tranzystorow* (en: types of transistors exist).

4. Results

We carried out tests of each implemented method. It should be specified that the methods suggested are difficult to test, because there doesn't exist standard tests to be able to describe their exact performance. For that, we were interested in two parameters, which are:

- recall
- precision.

According to the tests on the two first methods applied (the methods which carry out the agreements between the request of user and the base of the graphs), we can notice that the research of the graphs directly from the user's request gives satisfactory results. I.e., that is, it is always possible to find the graph if the graph exists – for both methods - and if for the first method the key words (semantic classes) are present. Even if both research methods give satisfactory results, they are far from being complete and they lack precise details, because according to the results of the research, too many occurrences are found in the corpus, some of which are not interesting from the user's request view.

In our tests we used the corpus from the electronic field like for example the press release. For the requests from users, we used the questions asked in public groups of the domain. These tests have allowed us to improve our graph base. On the other hand, it is necessary to specify that with our methods, we often find nonpermanent graphs or many graphs not detailed enough. We also noticed that the field is much broader than we had imagined

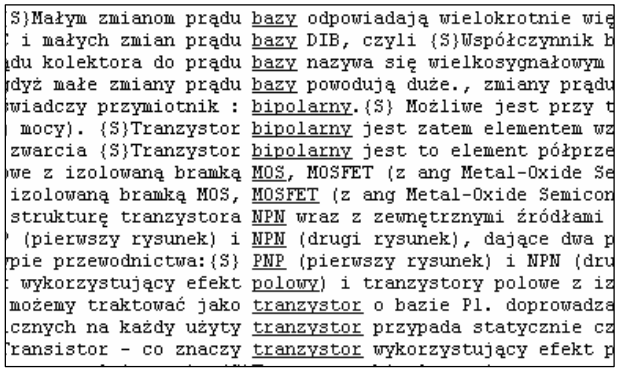
S}Małym zmianom prądu bazy odpowiadają wielokrotnie wię
i małych zmian prądu bazy DIB, czyli {S}Współczynnik h
du kolektora do prądu bazy nazywa się wielkosygnałowym
dyż małe zmiany prądu bazy powodują duże., zmiany prądu
wiadczy przymiotnik : bipolarny.{S} Możliwe jest przy t
mocy). {S}Tranzystor bipolarny jest zatem elementem wz
zwarcia {S}Tranzystor bipolarny jest to element półprze
we z izolowaną bramką MOS, MOSFET (z ang Metal-Oxide Se
izolowaną bramką MOS, MOSFET (z ang Metal-Oxide Semicon
strukturę tranzystora NPN wraz z zewnętrznymi źródłami
(pierwszy rysunek) i NPN (drugi rysunek), dające dwa p
pie przewodnictwa:{S} PNP (pierwszy rysunek) i NPN (dru
wykorzystujący efekt polowy) i tranzystory polowe z iz
możemy traktować jako tranzystor o bazie Pl. doprowadza
cznych na każdy użyty tranzystor przypada statycznie cz
ransistor - co znaczy tranzystor wykorzystujący efekt p

Figure 7 - Example of results

The various tests we carried out allowed us to:

- to increase the graph base while trying to have the most precise graphs
- to add semantic classes
- to increase the dictionary of the field

In spite of these improvements we made, we are still far from the ideal case. According to our tests results , and since it is necessary to start from the principle that more complex and complicated graphs are needed, we noticed that the method of the semantic classes gives better results that the direct method. Indeed, in the direct method, we can't benefit from a more enlarging complexity of the graphs, that is that by directly scanning for key words in the graphs we don't have the possibility to take into account the format of the graph, therefore often, the graph found is not permanent whereas in the first method this problem is regulated by the use of semantic classes.

The precision found in the corpus falls with the increase of the number of nonpermanent graphs used for search. However, we specified in the preceding chapter, that we based ourselves on the idea to find all the probable events and it is left to the user to carry out filtering.
Our research of the occurrences directly from the user's request gives satisfactory results despite the disadvantages mentioned above. We will extract information depending on the request, such as for example synonyms, electronics components of the same family or electronics components having the same parameters.
The results given by the method consisting in the creation of the graph directly from the request are easy to collect. We can describe this method as the method of direct search of key words by considering all the grammatical forms. This is why this method is used only in case where our graph base doesn't contain permanent graphs according to user's request. The results of our tests are presented in table 1.

Method	Precision	Recall
Semantic classes	**70.7%**	**56.76%**
Direct method	**61.16%**	**55.58%**
Automatic construction of the graphs	**34.4%**	**49.04%**

Table 1 - Testes: precision and recall

5. Conclusion

We focused ourselves on the automatic search task of information in a corpus, more precisely on the linguistic adaptation of the tools for the extraction of information concerning a precise field. Our study was made on the application "Unitex" since it's the tool that makes it possible to carry out a major search by using grammars, tables of lexicon-grammar and dictionaries. Our objective was:

- to adapt the software to work with the Polish language
- to develop linguistic resources to be able to carry out our research
- to add several methods in order to allow an automatic research of the occurrences in the corpus starting from the user's request.

We succeeded in the creation and the integration of three new methods: method of semantic classes, direct method and automatic construction of the graph These methods make it possible to establish a link between the user's request and the graphs. The graphs latter are used for the permanent information retrieval via the question asked by the user. The adjustment of the linguistic resources like the creation of the graphs or the adaptation of the dictionaries was part of our work. We obtained satisfying results, but it is necessary to specify that they remain several points to improve. The solutions from the automatic information retrieval presented in this report give an image of the complexity of this field and highlight the need for making improvements and especially for opening several doors in the domain of research. The results obtained could be improved by carrying out a major research on comprehension of users' requests or by improving the linguistic resources, as for example the graph base. Unfortunately, the possibility of having ideal solutions to find that useful information seems to be a Utopia.

References

[1] M.T. PAZIENZA, Information extraction (a multidiscipilinary approach to an emerging information technology). *Springer Verlag (Lecture Notes in Computer Science)*, Heidelberg, 1997

[2] E.M VOORHEES, Natural language processing and information retrieval. In M.T. PAZIENZA, Information extraction, toward scalable, adaptable systems, *Springer Verlag (Lecture Notes in Computer Science)*, Heidelberg, 1999, pp32-48

[3] H. ALSHAWI, The core language Engine. *MIT Press (ACL-MIT Press Series in Natural language Processing),* Cambridge, 1992

[4] J. SOWA, Conceptual Structures. Information processing in Mind and Machine. *Addison Wesley Publishing CO.*, Reading, 1984.

[5] H. KAMP, Evenemts, representations discursives et reference temporelle . *Langages, nb 64*, 1981, pp.39-64

[6] M. GROSS, The construction of local grammars, *Finite-State Languauga Processing, E. Roche and Y. Schabes, Cambridge*, Mass./London, England : MIT Press, pp 329-354, 1997

[7] S. PAUMIER, Recherche d'expressions dans de grands corpus : le systeme AGLAE. *Memoire de DEA*. Universite de Marne-la-Valee, 2000

[8] M. SILBERZTEIN, Dictionnaires electronique et analyse automatique de texte, le sesteme INTEX. *Masson*, Paris, 1993

[9] S. PAUMIER, Unitex 1.2 Manuel d'utilisation. Université Marne la Vallée, 2004

[10] S. PAUMIER, De La reconnaissance de formes linquistique a l'analyse syntaxique. *These*, Marne-la-Valee, 2003

[11] A. Balvet, Grammaires locales et lexique-grammaire pour le filtrage d'information vers une (re)utilisabilite des ressources linquistique pour la recherche d'information. *Conference TIA*, Nancy, 2001

[12] C. GARDENT, B. GUILLAUME, I. Falk, G. PERRIER, Le lexique-grammaire de M.Gross et le traitement automatique des langues. *LORIA & ATILF,* Nancy, 2005

[13] C. MACLEOD, R.GRISHMAN, A. MEYERS Comlex syntax : Building a computational lexicon. *In Preceedings of COLING'94*, pp 2668-272, 1994