A two-tier corpus-based approach to robust syntactic annotation of unrestricted corpora

Núria Gala

A two-tier corpus-based approach to robust syntactic annotation of unrestricted corpora

Núria Gala Pavia

XRCE, 6 chemin de Maupertuis F-38240 Meylan
LIMSI-CNRS, Bt 508 Université de Paris-Sud, F-91403 Orsay Cedex
Nuria.Gala@xrce.xerox.com

RÉSUMÉ. *Cet article présente un état de l'art des analyseurs robustes existants et propose un système automatique d'annotation syntaxique de corpus plus efficace fondé sur un diagnostic préalable à l'application de grammaires spécialisées. Après avoir décrit quelques analyseurs et avoir montré leurs limites en ce qui concerne le traitement de certains corpus, une approche d'analyse en deux étapes est proposée. Les différents modules grammaticaux formalisent tout d'abord des phrases considérées comme noyau puis certains phénomènes syntaxiques particuliers comprenant de la ponctuation ou entraînant des ambiguïtés structurelles. L'avantage de cette approche est, pour tout type de corpus, l'application d'une même grammaire stable optimisée puis l'adaptation du parseur en fonction de la présence de certains phénomènes qui sont traités spécifiquement. Cette stratégie garantit des taux de précision et rappel élevés quelle que soit la typologie du corpus.*

ABSTRACT. *This article gives a state of the art of robust parsers and proposes a more efficient automatic way of syntactically annotating corpora based on a diagnosis of a sentence before the application of specialized grammars. After describing some available systems and showing their limits in terms of parsing certain type of raw corpora, a two-tier approach is proposed for the architecture of a robust parser. The splitting of the grammar rules into several modules permits to formalize first core sentences and in a second time some syntactic phenomena containing punctuation or implying structural ambiguities. The advantage of this approach is, for any kind of corpora, the application of a single optimized grammar followed by the parser's adaptation to the presence of certain phenomena which are specifically processed. This strategy guarantees high precision and recall rates for any kind of unrestricted corpora.*

MOTS-CLÉS : *Analyseurs robustes, analyseurs de surface, grammaires de constituants vs grammaires de dépendances, annotation syntaxique de corpus tout-venant.*

KEYWORDS: *Robust parsers, "shallow" parsers, phrase-structure grammars vs dependency grammars, syntactic annotation of unrestricted corpora.*

"Real-world natural language processing must deal with huge amounts of data, which involve many, and messy, details."
[JEN 88]

INTRODUCTION

With the tremendous growth of the Web and the increasing number of electronic documents, automatic syntactic analysis (*parsing*) has been focusing over the last ten years on the capability of extracting valuable information from raw corpora, rather than on analyzing in depth complex linguistic phenomena. Thus, the goal of robust systems is more a partial representation of the syntactic information for all the sentences of a given text than a deep analysis which, due to the complexity of natural language, may be ambiguous most of the times. The analysis is generally deterministic instead of global and recursive.

One of the highlights of this approach is that the rules are empirically built after the observation of real corpora. This guarantees a formalization of linguistic phenomena a posteriori, taking into account heterogeneity and variability.

Robust parsing has been further strengthened as a core component in Natural Language Processing applications because it provides mechanisms for identifying major syntactic structures and/or dependency relations between words on very large collections of documents in an efficient way. A number of systems have been developed, be them rule-based ([JOS 96], [JAR 97]) or statistical ([CHU 88]).

Another proof of the increasing interest given by the scientific community to Robust Parsing is the sequence of events dedicated to this issue: ESSLLI Robust Parsing Workshop in 1996, the meeting on Constraint Satisfaction for Robust Parsing of spoken Language in 1998 and the recent COLING 2000 tutorial "Trends in Robust Parsing" (Jacques Vergne) as well as the Workshop ROMAND (Robust Methods in Analysis of Natural Language Data) 2000 (EPFL, Lausanne).

Two of the domains of application for robust parsing are document processing and semantic navigation/disambiguation/representation. Robust parsing plays a major role in document processing: from Information Retrieval to Data Mining [GAU 00], including Information Extraction, Indexing, Terminology Extraction [JAC 99] and Normalization [GRE 99]. As for semantic applications the results of a parser can be used to extract information from dictionaries as in [DIN 99] to create disambiguation rules and calculate the most probable meaning of a word in context.

Within this framework, in this article we first present a state of the art of robust parsers and then describe an approach based on the splitting of the analysis into two tiers. We consider the hypothesis that any given corpora contains a certain number of sentences which do not present difficulties for their analysis (we call them *fi rst-tier* sentences). The annotation of this subset is done by a core grammar with a very high level of linguistic precision. The coverage of the core grammar (the percentage of ana-

lyzed sentences) remains stable for most kinds of corpora (i.e. newspapers, technical manuals, scientific reports, legal documents, etc.).

The sentences containing phenomena requiring specific analysis (*second-tier* sentences) are analyzed by different specialized grammars. In this second-tier analysis, the parser can adapt its strategy depending on the presence of particular phenomena.

The results obtained show that the splitting of the grammar as well as the specialization of some empirically built rules improve the performance of the system and thus the syntactic annotation of the whole corpora.

The article is organized as follows. The first section presents a characterization of robust systems, giving special attention to the linguistic foundations. This section also provides a typology of several parsers and describes them in detail. Section two evaluates some of the available robust parsers and leads to the third section in which the problem of variety in raw texts is discussed. In section three, a classification of sentences based on their "syntactic complexity" is also proposed.

Section four illustrates an approach for robust parsing that considers the variety of linguistic phenomena as a key point for developing the overall architecture of the system in a modular way. Punctuation and structural ambiguity, which until recently have received little attention (except prepositional attachment) in NLP and Computational Linguistics [WHI 95], are considered essential within this framework. Finally, the conclusions look at the advantages of the approach and give an outlook for further research.

1. ROBUST PARSERS FOR UNRESTRICTED TEXT: AN OVERVIEW

The notion of "robust parsing" gained favor in the nineties following a general effervescence of computer techniques for the analysis of corpora [CHA 94]. Giving priority to the robustness of the system, i.e. the capability of giving a result even for unknown or ungrammatical input, this approach is completely focussed on the processing of large collections of texts, rather than on the depth of the linguistic analysis. Robustness is essential to efficiently treat large amounts of heterogeneous corpora.

Some of these parsers are often called *shallow* or *partial* because they do not attempt to analyze in depth but rather to produce a "minimal" analysis -annotation- for a whole text (i.e. bracketing of nominal phrases). However, it is important to make clear that not all robust parsers are "shallow": some systems calculate complex relations [1] between words and the output produced cannot really be considered a "partial" syntactic representation of the sentence.

1. Some syntactic relations (e.g. embeddings, control) require deduction rules that take into account diverse parameters.

For Abney ([ABN 94]), determinism, a local evaluation of pieces of a sentence and the existence of a cascade of specialized grammars (NPs, PPs, etc.) are three of the main characteristics of *partial parsing*. Taking them into account, we specify and expand these features to define *robust parsing*:

– a single syntactic analysis is produced for *chunking* (determinism);

– the analysis of linguistic phenomena is multi-stage (incremental);

– the grammars are empirically founded (corpus-based);

– an output is produced for any kind of unrestricted text.

Robust parsers are generally deterministic (i.e. [HIN 83]): one single analysis is produced for the phrase-structure analysis of each given sentence. This analysis is performed in different deterministic stages instead of globally and recursively. The aim is not to explore all the potential structural ambiguities nor to provide all the possible analyses, but rather to apply heuristics in order to extract the most plausible analysis for a given sentence. In certain systems, only when the syntactic information is not enough to apply a decision, all the possible results are given. This is the case of prepositional attachment, often needing semantic information to establish the appropriate link between a prepositional phrase and a noun or a verb.

The analysis of a robust parser is very cautious, decisions that are difficult to take at a given stage are left for later stages. Following [EJE 93] two main stages can be distinguished during the parsing process:

– a first stage of pre-analysis aiming at recognizing minimal syntactic structures for producing a preliminary representation that will be used as input for the following stage;

– a second stage having as purpose to calculate more complex structures and/or syntactic relations between the words of the sentence.

Within these two main stages, the analysis is done gradually, in different specific levels. This is the reason why robust parsers are generally called *"incremental parsers"*: non ambiguous structures are first analyzed and the result of a first step is used to improve the decision of a following step. In general, as in *partial parsing*, each step is "specialized" in a particular structure or syntactic relation.

Unlike the so-called *deep parsers*, the formalization of linguistic phenomena is done *a posteriori*, corpus-based or corpus-driven instead of theory-driven. The construction of the grammars after the observation of large amounts of corpora aims at syntactically *represent* any given input without being explicative of the linguistic mechanisms encountered.

For any kind of text, no matter its domain or the structure of the items it contains, a robust parser produces a result which can later be used for other applications (terminology extraction, event recognition, etc.). The reusability of a robust parser's output (the syntactically annotated corpora) is a key feature, even if the syntactic analysis is not always totally accurate (see section *3*).

1.1. *Technical approaches*

According to [SRI 97] two major approaches to robust parsing can be considered.

The *symbolic approach* concerns grammar-based parsers, systems made of a set of rules describing syntactic phenomena. Some of the grammar-based parsers ([JOS 96], [APP 93], [ROC 93]) rely on finite-state technology; but others (Fidditch [HIN 83], [TAP 97]) do not.

For finite-state systems the grammar is represented by a cascade of transducers created from regular expressions. The analysis can be constructivist if structures are progressively added during the analysis ([ABN 91], [GRE 92]) and/or reductionist if restrictions are applied during the analysis in order to remove potential analyses that appear to be incorrect ([CHA 96], [KAR 95]).

Probabilistic parsing uses machine learning techniques (neural networks, unsupervised learning, etc.) in order to obtain a partial representation of a given text. The rules (weights) are automatically obtained from annotated corpora. Some representative statistical approaches to parsing are described in [CHU 88], [DAE 99], [MUN 99].

The main problem of this approach is the need for (manually or automatically) annotated corpora to create the heuristics.

1.2. *Linguistic approaches*

Depending on the linguistic approach underlining the parser, the syntactic information is represented differently. There are two different formal models to represent syntactic information: one that aims at giving evidence of “structures” and a second one showing “relations” between words.

1.2.1. *Structures*

A first approach is based on *Phrase Structure Grammar* (PSG). This theory appeared during the thirties in the US [BLO 33], in the same line as the structural school (trend that Chomsky followed later on with his transformational generative approach).

The main notion in this approach is the principle of immediate constituency between the elements of a sentence, i.e. the words of a sentence can be grouped into bigger structures (taking into account their meanings). Syntactic analysis is here reduced to mark categories and constituents, and to give evidence of how these constituents are linked within a sentence.

Phrase-Structure Grammar was the predominant trend in linguistics until the seventies. Some reasons of this predominant position [MEL 88] are the use English as a language for studying linguistics (the structure of this language makes constituents relatively easy to mark), the excessive trend to formalization and a rather negligent attitude towards semantics.

Within PSG, in the nineties [ABN 91] described a finer-grained decomposition of the structures of a sentence based on a criterion of prosody. For Abney, chunks are the non-recursive cores of "major" phrases (i.e. NPs, VPs, APs, etc.) motivated by psycholinguistic evidence.

1.2.2. *Dependencies*

In the sixties in Europe another approach appears showing more interest on semantics (i.e. [FIL 68]) and on the variety of languages (with more complex syntactic structures). *Dependency Grammars* [TES 59] refuse the strict representation imposed by the constituency model. Their focus is on the relations established between the words of a sentence and how these relations are produced ([KAH 00]).

This trend of syntactic analysis becomes fundamental during the eighties as different theories give evidence: *Relational Grammar* [PER 83], *Word Grammar* [HUD 94], *Functional Grammar* [HAL 94], *Constraint Grammar* [KAR 95].

1.3. *A typology of robust parsers*

Taking into account these linguistic approaches and the kind of analysis a robust parser produces, three kinds of systems are conceivable. We focus here on symbolic parsers, not characterizing probabilistic systems.

1.3.1. *Parsers based on constituency -chunks-*

CASS ([ABN 91] and [ABN 96]) is a cascade of cheap analyzers for tagging, NP-recognition, general chunking, simple clause and light (prepositional) attachment marking. Each one of these analyzers is deterministic so that one single best analysis is produced:

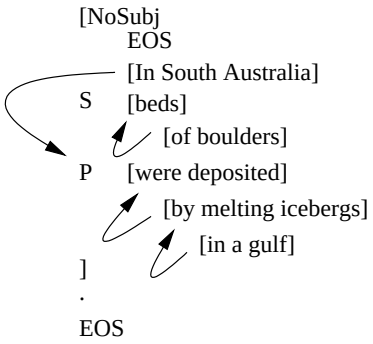

FIG. 1 Detail of the output for CASS.

The output is a bracketed text (with arrows marking the linked elements, as given below *"In South-Australia"*with *"were deposited"*, *"of boulders"*with *"beds"*,

etc. in the sentence *"In South-Australia beds of boulders were deposited by melting icebergs in a gulf."*) [2]:

1.3.2. *Parsers based on dependency relations*

Another type of robust parsers are based on extracting dependency relations between the different elements of a sentence.

Functional Dependency Grammar Parser [TAP 97], [TAP 99] (FDGP) is developed at the University of Helsinki and belongs now to Conexor Society. Some parts of this system have been adapted from a previously described version, ENGCG-2 [VOU 96].

The description of the surface syntax here is based on Constraint Grammar [KAR 95]. The main element is a *nucleus* (Tesnière model is followed very closely). Each *node* is a basic syntactic element and between these nodes different syntactic relations are established: there is a variety of relations and dependencies (*link names*): from morpho-syntactic relations such as *subject*, *object*, *complement*, etc. to low-semantic relations (*time*, *purpose*, etc.).

There is an on-line version for English [3] (Conexor tools version 3.4, Conexor Functional Dependency Grammar) in which the morpho-syntactic information is displayed in columns:

```
0
1 The           the           det:>2    @DN> DET SG/PL
2 decrease      decrease      subj:>3   @SUBJ N NOM SG
3 reflects      reflect       main:>0   @+FMAINV V PRES SG3
4 the           the           det:>5    @DN> DET SG/PL
5 repeal        repeal        obj:>3    @OBJ N NOM SG
6 of            of            mod:>5    @<NOM-OF PREP
7 catastrophic  catastrophic  attr:>8   @A> A ABS
8 health-care   health-care   attr:>9   @A> N NOM SG
9 benefits      benefit       pcomp:>6  @<P N NOM PL
.
```

LGP [SLE 93] (*Link Grammar Parser*) [4] is another dependency relation parser for English, based on Link Grammar, an original (dependency-based) theory of English syntax. Given a sentence, the system assigns a syntactic structure, which consists of a set of labeled links connecting pairs of words.

A new version has been released in 2000, which makes the parser hybrid deriving a "constituent" representation (showing noun phrases, verb phrases, etc.) from a *linkage*

2. The 'EOS' tags indicate the begin and the end of the sentence; "S" stands for subject and "P" for predicate.
3. http://www.ling.helsinki.fr/~tapanain/dg/eng/demo.html
4. http://bobo.link.cs.cmu.edu/link/

(the usual link grammar representation of a sentence, showing links connecting pairs of words) .

1.3.3. *Hybrid systems*

Some parsers merge the notion of constituents (*chunks*) and the notion of dependencies (*functions*) and produce as a result a marked sentence with a set of dependency relations between some words (usually the heads of the chunks).

CHAOS [BAS 92] is a non deterministic parser made of a finite-state grammar with lexicalized rules [BAS 99]. The syntactic information is represented by graphs (nodes are words and branches are dependencies).

This parser is made of different components (multi-agent) according to the action they perform (constituency oriented processors, dependency processors). The modularity of the system is based on the kind of processing performed for each task. Some modules of the parser are monotonic (chunker) and some non-monotonic (disambiguator). It works for Italian and English.

SEXTANT [GRE 92] is a low-level parser based on finite-state technology. Chunking and dependency extraction are deterministic and focused on nominal groups (including prepositional or verbal relations containing nouns). The main purpose of this system is to extract syntactic patterns by marking and filtering them like in [DEB 82]. The notion of chunk is not strictly faithful with Abney's definition: nominal chunks can include prepositional phrases.

In [GRE 95], the syntactic information is represented by means of a table: columns contain the different morpho-syntactic information (kind of chunk, categories, etc.) and lines lexical units.

There are currently seven versions (English, French, Spanish, Italian, Portuguese, Dutch, German) in which the morpho-syntactic information is represented by a bracketed (tagged) sentence with a set of dependency relations.

```
ADJ( health-care benefit )
ADJ( catastrophic benefit )
UNSURE( catastrophic health-care benefit )
NNPREP( repeal of benefit )
DOBJ( reflect repeal )
SUBJ( decrease reflect )
------------------------------------------------
 [NC The+AT *HeadN decrease+NN NC] [VC *ActV reflects+VBZ VC]
[NC the+AT *HeadN repeal+NN of+IN catastrophic+JJ health-care+JJ
*PrepN benefits+NNS NC] .+SENT (11)

Orig:  The decrease reflects the repeal of catastrophic health-care
benefits . (11)
```

The parser from GREYC (Caen University) [5] [GIG 97a] is a deterministic robust parser providing two different levels of representation for the syntactic information (built simultaneously by two interacting input-driven processes).

The parser identifies *non recursive phrases* (nrp) during the chunking process. It also identifies syntactic relations between the *npr* in order to built a functional representation of the sentence. Dependency relations are extracted using *linking rules* and *linking constraints* in a memory-based approach (the analysis is done by a set of memories, a "memory" being a sort of stack of linguistic objects -chunks- waiting for the correct valencies to be related to other linguistic objects). There is a first version of the parser for French but other versions have been developed (i.e. for English).

IFSP (*Incremental Finite State Parser*) [6] [AÏT 97] is a finite-state parser made of a cascade of transducers that apply incrementally. It is deterministic for the chunking process and non deterministic for certain dependency relations (those concerning prepositional phrase attachment).

The output of this parser is a bracketed sentence (with chunks) with a set of dependencies of the form *dependency(x,y)* (where *dependency* stands for the morpho-syntactic name of the relation and *x* and *y* are the linked words). Three versions exist for the IFSP: English, French and Spanish.

```
SUBJ(decrease,reflect)
DOBJ(reflect,repeal)
ADJ(health-care,benefit)
ADJ(catastrophic,benefit)
NNPREP(repeal,of,benefit)

    _The decrease reflects the repeal of catastrophic health-care
benefits .

   [SC  [NP _The decrease NP]/SUBJ :v reflects SC] [NP the repeal
NP]/OBJ [PP of catastrophic health-care benefits PP] .
```

2. EVALUATION OF SOME AVAILABLE ROBUST PARSERS

Most of the grammars of robust parsers are developed using newspaper corpora (i.e. GRACE action in France for the evaluation of taggers and parsers). Sometimes technical manuals are also used: together with newspapers these corpora are easily available from electronic resources. The following sections show the results obtained by some parsers (reported in the referenced articles) using this kind of corpora.

5. http://users.info.unicaen.fr/~giguet/syntactic.html
6. http://www.xrce.xerox.com/research/mltt/fsnlp/fsparsing.html

2.1. *SEXTANT*

Among the seven versions, only the English one is evaluated using Software Manuals [GRE 95]. The evaluation concerns binary relations on 130 sentences of the LOTUS technical manual corpus.

Kind of Corpora	*Precision*	*Recall*
Technical	70 %	64%

Table 1. Precision and recall rates for binary relations on SEXTANT (English) parser.

Problems encountered were: tagging errors that could not be recovered and thus are propagated throughout the parse, commas are ignored between a noun and a verb, interrogative clauses are not taken into account, subject of infinitive verb phrases is not extracted.

2.2. *Parser from GREYC*

The evaluation of Vergne's analyzer for French [GIG 97b] is focussed on subject-verb relations, obtained on a set of articles from the French newspaper *Le Monde* (about 474 sentences).

Kind of Corpora	*Precision*	*Recall*
Newspaper	96,4%	94%

Table 2. Precision and recall rates for French subject recognition.

Incorrect relations appear mainly as a result of ill-formed non recursive phrases, inverted subjects in reported speech, incorrect tags, coordinations not found.

2.3. *IFSP*

The evaluation of the IFSP for French [AÏT 97] is also on subject recognition over technical manual text (157 sentences) and newspaper articles from *Le Monde* (249 sentences). Figures on precision and recall are reported in the following table.

Kind of Corpora	*Precision*	*Recall*
Technical	99,2%	97,8%
Newspaper	92,6%	82,6 %

Table 3. Precision and recall rates for French subject recognition.

The reason for obtaining better results on technical texts is that this kind of corpora was used to develop the grammar.

IFSP for English shows about 84,5 % of precision for subject relations using newspaper texts [AÏT 98].

Finally, the figures for the evaluation of the IFSP for Spanish [GAL 99] on subject and object recognition using technical texts (292 sentences) and newspaper articles from *El Pais* (252 sentences) are the following:

Kind of Corpora	*Precision*	*Recall*
Technical	80,7 %	71,8 %
Newspaper	81,7 %	75,4 %

Table 4. Precision and recall rates for Spanish subject recognition.

Low results on subject recognition are mainly due to tagging errors, noun phrases containing time expressions in sentences where there is no real surface subject (very frequent in Spanish) and ambiguous coordinations.

2.4. *Discussion*

As we will show in this section, the reported results tend to decrease if the input text comes from different (i.e. more specialized) domains: reported speech, scientific reports, legal documents, etc. This observation reinforces the need of adaptative systems that, in order to obtain more accurate syntactic annotations, take into account the type of text [BIB 93], [ILL 00] and/or the particular characteristics of specific linguistic phenomena (see sections *3* and *4* of this article).

However, although the grammars of most existent robust parsers are corpus-based, they are not representative enough of the linguistic phenomena encountered in specialized raw corpora. To empirically show it, we have conducted an experiment on heterogeneous -specialized- corpora (about one hundred sentences for English and about seventy for French) using two existing parsers: SEXTANT and IFSP. The following evaluation is focused on precision and recall rates for constituency analysis (NPs) and subject extraction. The results obtained and some error analysis are presented.

2.4.1. *Constituency analysis (chunking)*

Although the articles describing the available parsers do not present results on the phrase structure analysis (accuracy of NP chunks), we have conducted an evaluation using a variety of raw corpora (financial report on South Africa, medical report on cancer, technical manual, reported speech of a conversation). The results of this expe-

riment will be later compared with the constituency analysis produced by the parser we are currently developing (see section *4.4*).

Parser	*Precision*	*Recall*
Sextant E	89,2 %	88,5 %
Sextant F	87,1 %	91 %
IFSP E	87,7 %	83,3 %
IFSP F	90 %	91,4 %

Table 5. Precision and recall rates for different parsers on NP analysis in varied corpora.

Notice that, as mentioned in *1.3.3*, the definition of chunks is slightly different for both parsers (IFSP's are faithful to Abney's definition -the rightmost element is the head of the chunk- while SEXTANT's are closer to the traditional notion of *group* or *phrase*, including a head and their modifiers).

Errors in precision in both parsers are sometimes due to wrong tokenizing and tagging. Another important source of errors is punctuation, i.e. wrong constituency grouping in series of figures and lists containing comas or semi-colons. For recall, unidentified NPs correspond to tagging errors in ambiguous words and to titles and items of lists without punctuation (and thus marked as integrating a single chunk).

2.4.2. *Subject extraction*

The following table shows the results of our evaluation for subject extraction.

Parser	*Precision*	*Recall*
Sextant E	70,8 %	63,8 %
Sextant F	65,1 %	44,3 %
IFSP E	78,9 %	74,8 %
IFSP F	82,3 %	86,8 %

Table 6. Precision and recall rates for different parsers on SUBJECT recognition in varied corpora.

For English, SEXTANT results are slightly better than the ones mentioned by the author[7]. IFSP results on precision are lower, specially due to sentences presenting punctuation (i.e. quotes) or ambiguous structures (i.e. coordinations).

For French there is no published evaluation concerning SEXTANT. Some errors on precision come principally from wrong tokenization and tagging and the non identification of inverted subjects.

7. The original evaluation was carried out on technical manuals which already present complex phenomena from the point of view of parsing (imperative clauses, lists, etc.).

IFSP results on precision are lower due to tokenization, tagging and wrong links made between elements belonging to different sentences that the parser hasn't identified (i.e. lack of punctuation in a title). Recall is worse compared to the newspaper evaluation but better compared to technical manuals [8].

To conclude this section, table 7 shows precision rates for different parsers in English and French on subject recognition using different kind of corpora. The first column concerns the evaluation done by the authors when available (see tables 1 and 3) [9] using newspapers and some technical manual (a), the second one takes into account heterogeneous corpora with a variety of syntactic phenomena (b):

Parsers	*(a)*	*Parsers*	*(b)*
Sextant E	**70%**	Sextant E	**70,8 %**
IFSP E	**84,6 %**	IFSP E	**78,9 %**
Sextant F	(-)	Sextant F	**65,1 %**
IFSP F	**95,2 %**	IFSP F	**82,3 %**

Table 7. Average of precision rates for subject recognition by different parsers on different corpora.

3. VARIETY IN RAW CORPORA: THE BOTTLENECK

It is generally accepted that a significant problem for parsers is to obtain a high coverage of linguistic phenomena [ABE 99] without, on the one hand, loosing precision and, on the other hand, without restricting themselves to limited corpora (and thus becoming *ad hoc* systems).

As we have seen, precision rates are quite satisfactory on newspaper and technical texts. However, expanding the analysis to other kind of corpora decreases both precision and recall (tables 5, 6 and 7). Very often this is because some "non standard" phenomena encountered are simply not taken into account by the grammars of the parsers [CHA 00].

The more unrestricted text to parse from heterogeneous domains, the more we risk to encounter subtle phenomena. If we consider an heterogeneous corpus from newspapers, legal, finance and scientific documents we find about 4,7 % of sentences containing structures with numbers and letters, about 5,9 % of imperative sentences, about 8,4 % of ambiguous coordinations, etc. [GAL 00]. Here are three examples of

8. One reason could be that the version we have used to make this experiment is slightly better than the one mentioned in [AÏT 97], namely, imperative verbs are explicitly marked and thus subjects are not identified in imperative sentences.
9. For French IFSP (a) the figure is the average of precision rates in table 3.

some of these 'subtle' phenomena, respectively, series of figures in titles (A), lists (B) and imperative sentences (C):

(A) *Avis CMF 198C0355 (SBF 98-1501) 14/4/98*
(Notice CMF 198C0355 (SBF 98-1501) 14/4/98)

(B) *Electrodes et isolants :*

– *gris-brun : fonctionnement normal de la bougie,*

– *noir : mélange trop riche,*

– *gris clair-blanc : mélange trop pauvre.*

(Electrodes and isolators: gray-brown: normal plug working, / black : too rich mixture, / fair-white gray : too poor mixture.)

(C) *Dévissez les bougies.*
(Unscrew the plugs.)

Subtle non-standard phenomena can be very frequent in all kinds of texts (i.e. enumeration of items) or specific to a domain (i.e. imperative clauses in technical manuals): linguistic features are distributed differently across different kind of texts [BIB 93]. The variety of encoding of the information found in electronic sources provides enriched texts (with pragmatic marks i.e. in oral transcriptions) or impoverished forms (lack of significant punctuation, simplified syntax i.e. in titles, etc.).

The variation of linguistic features present in different kinds of corpora have obviously important implications for parsing. Grammars generalizing the linguistic description are likely to be inaccurate, because particular phenomena will not be taken into account. On the other hand, very specific parsers (sort of *ad hoc* analyzers for a precise domain) will be very accurate for a given domain but completely inappropriate for other kinds of domains.

A sort of "compromise" can be established if we consider the following hypothesis: any kind of corpora contains (a) "core" sentences ("simple" structures easy to capture by syntactic rules) and (b) "complex" phenomena (requiring particular syntactic treatment) [10].

Following this idea, a two-tier syntactic analysis is foreseen here. The aim is to analyze first the sentences that do not present complex phenomena (*first-tier sentences*): sentences containing phenomena that do not require a specific fine-grained syntactic treatment by the parser. In a second stage of the analysis, a more complex strategy has to be implemented to take into account sentences containing certain phenomena such as punctuation and structural ambiguities (*second-tier sentences*).

10. The number of "core" or "complex" sentences will depend on the corpus: the more a corpus is specialized, the higher the number of complex structures (compare plain text in a newspaper to plain text in a chemical report, to give an example).

The splitting of the analysis after a very precise diagnosis of the linguistic phenomena and, as a consequence, the application of specific strategies depending on the features of each sentence, are the keys to obtain very accurate results for most kinds of unrestricted text.

Thus for both tiers of analysis, we aim at obtaining very high precision and recall rates: around 95 % respectively, taking into account eventual errors from tokenization or tagging, for most kinds of corpora (see an evaluation in section *4.4*). The two-level analysis is then based on the fundamental distinction made between *fi rst-tier* and *second-tier* sentences.

3.1. *First-tier sentences*

From the point of view of parsing, our intuition is that certain sentences -present in all kind of corpora- are easy to analyze. We attempt to formalize this intuition in what we call "first-tier" sentences. To do this, we use constituency-based notions.

We distinguish three kinds of segments: basic, additional and super-structure.

– six *basic segments*: main chunks (Abney-like [ABN 91], already used by [BAS 92], [AÏT 97], [GIG 97b]):

[NP] noun phrases

[AP] adjectival phrases

[PP] prepositional phrases

[FV] finite verb phrases

[IV] infinitive verb phrases

[GV] present participle verb phrases

– five *additional segments*: two of them for marking embedded clauses and three for grouping sets of chunks in specific positions [11]:

[BG] beginning of subordinate group

[SBC] subordinate clause

[ANP] group of chunks preceding a NP subject

[PNP] group of chunks following a NP subject and preceding a main verb

[PFV] group of chunks following the finite verb phrase of the clause

– one *super-structure* used to mark sentences with certain characteristics:

[S] first-tier sentence (the nature of its elements as well as their order depends on the rules defined by the first-tier sentence grammar).

11. ANP stands for "ante" NP, PNP for "post NP" and PFV for "post FV".

The simplest *first-tier* sentence[12] is a sentence with a noun phrase and a finite verb, eventually with an adverb or one basic constituent, ending with a relevant punctuation mark (full stop or semi-colon). Examples (the NP subject and the main FV are marked with square brackets; on the second example the adverbs are not delimited as there is no adverbial chunk in our approach and they are not included in the verb chunk because this ends by its head):

[Le différentiel] [sera alors bloqué].
(The differential will be then bloqued.)

[Leur tension] [ne se modifiera] plus par la suite.
(Their tension will not be further modified.)

These kinds of sentences have six words in average and constitute about 0,5 % of the whole sentences of an heterogeneous corpus (15 sentences over 4 034) [13].

First-tier sentences may also present additional segments such as ANPs or PNPs under certain conditions: *n* basic constituents are admitted after a main verb as well as a maximum two subordinate clauses and two commas in specific patterns, i.e. three NPs separated by two commas are not admitted as they might be part of an enumeration.

Here are some examples of first-tier sentences presenting commas, subordinate clauses and groups of chunks.

Effectivement, [les changements] proposés qui touchent le secteur agricole [sont] très importants.
(In fact, the proposed changes concerning the agricultural area are major.)

Pour que le moteur se refroidisse bien , [il] [est] donc nécessaire que la courroie soit en bon état et que sa tension soit correcte.
(For good motor cooling it is necessary that the belt is in good condition and that its tension is correct.)

As mentioned above, punctuation is not accepted in first-tier sentences (except for a full stop or semi-colon at the end of the sentence, commas in specific patterns or hyphen for compound words). Non-ambiguous coordinations are also admitted in first-tier sentences. In the following example coordinated elements are marked in parentheses. Notice that each one of these elements can be split into chunks, i.e. NP[les pertes], AP[sanguines].

(Les pertes sanguines) et (la durée d'hospitalisation) étaient en défaveur du curage D3.
(Blood loos and lenght of hospitalization were counterindications of D3 curetage.)

---

12. We are currently using French as working language to develop our grammar. Thus the definition of *first* and *second-tier* sentence takes French as a reference.
13. We have used a French corpus of 87 000 words from the following domains: oral transcription of an interview, general newspapers *Le Monde*, *Libération*; financial articles from *Les Echos*, scientific –medicine, physics, chemistry– and legal reports, technical manuals.

L'unique compilation critique des niveaux d'énergie des actinides neutres (a été effectuée ici) et (constitue une étape déterminante dans la spectroscopie de ces éléments).

(The only critical compiling of the neutral actinide energy levels has been done here and constitutes a decisive step in the spectroscopy of these elements.)

As a whole, first-tier sentences have an average length of 18 words and constitute about 20 % of the sentences of most kinds of corpora (in our experiment, 821 sentences over 4 034).

To sum up, the empirical observation of different kinds of corpora has lead us to formalize the structure -type and order of constituents- of what we call first-tier sentences. These sentences, as they do not present non-lexical nor ambiguous phenomena, can be easily parsed by a "core" grammar (see section *4.1*) with very high precision and recall rates (about 96 %), for most kinds of unrestricted corpora.

3.2. *Second-tier sentences*

From the point of view of parsing, second-tier sentences present specific features that have to be formalized carefully, i.e. taking into account very particular phenomena such as interrogative clauses in which the order of the elements is different from assertive clauses, different functions for a given punctuation mark (colon at the beginning of a list, before a quotation, etc.).

We have established a first difference between those *second-tier* sentences containing punctuation and those implying structural ambiguities.

3.2.1. *Phenomena linked to punctuation (non lexical marks)*

The analysis of punctuation is essential for accurately chunking [JON 94]; it also contributes to later resolving structural ambiguities [BRI 94].

(1) Interrogative and exclamative clauses.

They have a specific punctuation mark ("?" or "!") and can be constituted of groups of chunks without finite verb phrase, or they can be whole sentences. In this case, for interrogative clauses, the order of the elements of the entire sentence is modified.

Halte au massacre du peuple palestinien !
(Halt to the slaughter of Palestinian people!)

Quel est son véritable programme?
(What is his real program?)

The syntactic relations extracted from these kinds of clauses may take into account the lack of main finite verb or the particular order of the elements of an interrogative clause.

(2) "Structural" segments.

They are usually titles of sections and subsections and appear generally without punctuation at the end of the line. They may be composed of groups of chunks, frequently without main verb. They appear in all kind of corpora.

TABLE DES MATIÈRES:
1. APERÇU DE L'AFRIQUE DU SUD
(SUMMARY: 1. GENERAL SURVEY OF SOUTH AFRICA.)

L'emploi américain fait sortir le dollar de l'ornière
(American employment puts dollar out of the woods.)

The main difficulty in properly parsing these segments is to take into account the end of line character(s).

(3) Lists.

They are enumerations of items organized vertically (*'vertical lists'* [WHI 95]). There is sometimes a sentence introducing the list followed by a colon. The items of a list are often signaled by punctuation such as "-", "*", etc. The main difficulty for accurately parsing these structures is to take into account the overall sentences integrating the list.

3 - Desserrez la vis - pointeau de purge (un 1/2 tour à un tour) .
4 - Appuyez sur la pédale de frein
5 - Fermez la vis - pointeau de purge (...).
(3 - Loose the screw - breeding needle (from half to one turn). 4 - Press the brake pedal. 5 - Tighten the screw - breeding needle (...).)

Le présent chapitre analyse certaines des possibilités commerciales existant au chapitre:

- des céréales ;
- des oléagineux ;
- des légumineuses à graines ;
- [...]
(The following chapter analyzes some of the commercial possibilities present in the chapter: cereals; oil-producing; leguminous plants; [...])

(4) Delimited units.

They are sets of words delimited by punctuation marks. Depending on their characteristics they may be considered quotations, named entities (titles of books, films, etc.), direct speech, comments or appositions.

All these units appear between parentheses, brackets or hyphens. They can be single words or acronyms but also nominal or other kind of chunks (APs, PPs, etc.) and even complete sentences. Example:

Parmi les autres valeurs à la baisse, on remarquait Pinault-Printemps, dont les ambitions dans le secteur -très gourmand en capitaux- des télécommunications semblent inquiéter le marché.

(Among other downwards values, Pinault-Printemps is to be highlined, whose ambitions in the telecommunications area -very greedy in capitals- seem to frighten the market.)

They can also appear in the middle of a sentence inside quotes and starting with capital letters (named entities). In this case they are often groups of nominal chunks.

Le thème "Structure et Dynamique des Atomes et des Ions" a une place toute particulière dans les activités du laboratoire.

(The subject "Structure and Dynamics of Atoms and Ions" has a particular importance within the activities of the laboratory.)

Another kind of "delimited units" are quotations appearing as direct object of a verb of saying (*affirm, assert, assure, say...*). The SVO order can be maintained or the quotation can be focalized as first position in the sentence. The quotation can also be part of the main clause.

"Le risque d'un échec n' est pas mince" , a assuré Romano Prodi lors d'une conférence de presse à Bruxelles, à une semaine de l'ouverture du sommet.

("Failure risk is not excluded", Romano Prodi assured during a press conference in Brussels.)

Les syndicats de greffiers et le ministère de la Justice ont "trouvé un accord sur les termes d' un protocole pour mettre fin au mouvement".

(Clerks' unions and the Ministry of Justice have "found an agreement on the conditions of a protocol in order to end up with the movement".)

A last case concerns specific words (one or a few) enclosed between quotes in specific usages.

Il évoque aussi , dans l' Express, les "disparitions" et ce qu'était l'état d'esprit des chefs militaires de l'époque.

(He also evoques in the *Express* the "disappearances" and what the mindset of the military leaders of that period was.)

The syntactic annotation of these segments within a sentence must take into account their structure as a unit, independently of their internal constitution. This has to be taken into account at the level of chunking and transmitted to the dependency extraction.

3.2.2. *Phenomena linked to structural ambiguity (lexical marks)*

The phenomena described in this section deal with the lexical units and not with additional structural marks. Here the words themselves are at the origin of syntactic structures that have to be analyzed specifically. On one hand, imperative verbs impose certain constrains to the general structure of the sentence; on the other hand, the presence of two lexical units sharing the same syntactic function in the sentence has to be

identified at the level of chunking to later extract a single dependency containing the coordinated elements.

(5) Imperative clauses.

The main characteristic of imperative clauses is that they have no NP subject. The knowledge by the parser that it is an imperative clause avoids marking a NP following the verb as its inverted subject (instead of its object). In French, the verb of imperative sentences is a second person of plural (finishing by *-ez*) or an infinitive.

Refermez l'ouverture latérale.
(Close the lateral opening.)

Abaisser le bouton de commande pour l'engagement.
(Push the lever down to start it.)

(6) Coordinations.

Ambiguous coordinations make the automatic extraction of syntactic relations very difficult. A coordination links two elements, single words, chunks, groups of chunks or entire clauses. The main characteristic is that the coordinated elements perform the same syntactic function. The elements can be identical (i.e. two noun phrases) or have different categories (adjective and past participle, noun phrase and subordinate clause, etc.) [BRU 98].

There are different coordination phenomena. The automatic identification of the scope of the coordinated elements requires specific analysis in each case.

Ellipsis is a coordination in which one of the elements of the coordination is absent (the subject in the following example).

Le débat concernant la nature juridique de la cession de 1940 [[est théorique en l'espèce] et [n'a pas à être tranché]].
(The debate concerning the juridical nature of the 1940 transfer is theoretical in the case and does not have to be settled.)

In *distributions* there are two parallel coordinations in the same sentence and the elements of these coordinations are linked the first with the first and the second with the second. There is usually an adverb or an adjective such as (in French) *"respectivement", "respectifs"* (*respectively, respective*).

La valeur totale des produits agricoles [[importés] et [exportés]] a été estimée, en 1993, [[à 4,3] et [à 5,45 milliards de rands], respectivement].
(The total value of agricultural products imported and exported has been estimated, in 1993, to 4,3 and 5,45 billions of rands respectively.)

For *correlations*, the two coordinated elements appear with a correlation element preceding each of them: in French (*ni... ni...*; *soit... soit...*; etc.) (neither... nor etc.).

La Loi sur l'éducation n'exige pas que les enfants fréquentent [[soit une école laïque] [soit une école catholique romaine]].

(The Law on Education does not demand children to go either to a laic school or to a catholic Roman school.)

Finally, *enumerations* are coordinations containing more than two elements separated by commas except the last two which are separated by a coordination conjunction. They often have a colon introducing the enumeration (note the similitude with the above mentioned vertical lists).

Présents: les juges La Forest, L'Heureux-Dubé, Sopinka, Gonthier, Cory, McLachlin, Iacobucci et Major.
(Present: Judges La Forest, L'Heureux-Dubé, Sopinka, Gonthier, Cory, McLachlin, Iacobucci and Major.)

Le réglage s'effectue dans l'ordre numérique des cylindres : 1, 2, 3 et 4.
(The tuning is made in the numerical order of the cylinders: 1, 2, 3 and 4.)

A sequence of noun phrases is not always an enumeration (i.e. apposition). Specific features have to be taken into account to provide accurate analysis for each phenomenon at the level of chunking as well as at the dependency extraction.

To recap, second-tier sentences present particular syntactic phenomena to be parsed with special mechanisms. Two main sub-groups of second-tier sentences can be distinguished, depending on the presence of non-lexical elements such as punctuation or of lexical phenomena appearing mainly in ambiguous coordinations or imperative clauses. A parser aiming at producing significant syntactic annotation (accurate annotation, that is at least >90 % of precision and recall rates) has to be adaptative to these kind of phenomena.

4. TWO-TIER ROBUST SYNTACTIC ANNOTATION

Though some existing parsers are claimed to be modular (i.e. often based on a multi-agent architecture as in [SAB 90]), for linguistics applications such as machine translation or information retrieval it is generally held that the modularization should be linguistically rather than computationally motivated.

In this sense, modularity (the strategy of dividing a problem into sub-problems until there is a collection of small relatively simple problems) should arise after having given evidence of the different linguistic phenomena present in a given corpus.

In our approach, the notion of "modularity" (i.e. a two-tier analysis) arises from the variety of linguistic phenomena to be tackled even if it keeps some ideas consistent with the computationally motivated systems, that is the relative independence of the modules and, at the same time, the inter-connection between them: all the modules are part of a whole application (the parser) aiming at obtaining accurate syntactic information from any kind of corpora given as input.

Thus modularity must be here understood as the possibility of activating a certain set of grammar rules -specialized in a specific phenomena- if a given sentence presents what we call a “complex” structure (from the point of view of parsing).

The system in which we are working at the Xerox Research Centre Europe is based on a parser designed by C. Roux, S. Aït-Mokhtar and J.P. Chanod. We use this system as a platform to develop a new (modular) architecture (figure 2) made of new grammar rules for a two-tier phrase-structure analysis and dependency extraction.

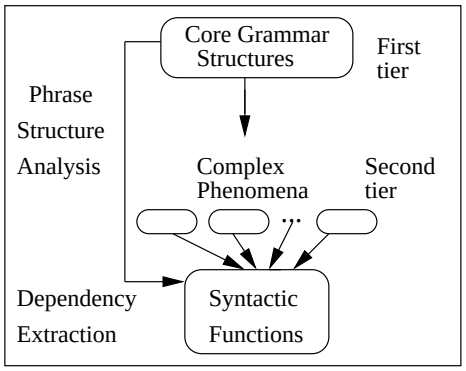

FIG. 2 General architecture.

Following the principle illustrated in section *3*, our parser is thus made of different kinds of modules, namely a first one for *first-tier* sentences and several others for *second-tier* sentences. A third kind of module is devoted to the dependency extraction from already constituent-analyzed sentences.

4.1. *Core grammar module*

The Core grammar is a set of rules permitting (1) the division of a sentence in different basic segments, (2) the division into additional segments, (3) the marking up of first-tier sentences.

The identification of basic segments defines non recursive chunks [ABN 91]. Eventual structural ambiguities are avoided by delimiting the chunk by its head (the head being the rightmost element of the chunk: nouns for NPs and PPs; adjectives for APs; verbs for FVs, GVs and IVs).

The following annotated sentence corresponds to the output produced by the core grammar at different levels of the analysis: basic segments, additional segments, superstructures (identification of first-tier sentences). At their beginning there is always a number and the tag “MAX” corresponding to a maximal structure.

Example: *"Le Canada pourrait saisir cette occasion pour accroître ses exportations de blé."*(Canada could grasp this chance to increase its wheel exports.)

```
335>MAX{NP{Le Canada} FV{pourrait} IV{saisir} NP{cette occasion}
 IV{pour accroître} NP{ses exportations} PP{de NP{blé}}.}
```

Note that in this example, a traditional grammar would have marked *"ses exportations de blé"* as a unique constituent. However such an structure, NP PP following an infinitive, is ambiguous because both chunks could potentially be attached to the infinitive as in *"Une économie en plein essor poussera à augmenter la production de 5 %*. (A booming economy will push to increase production by 5 %.).

```
672>MAX{NP{Une économie} PP{en NP{AP{plein} essor}} FV{poussera}
IV{à augmenter} NP{la production} PP{de NP{5 \%}}.}
```

Here the NP *"la production"* and the PP *"de 5%"* that follow the finite verb are two different constituents both depending on the verb.

The step of marking basic chunks can be considered as a *"shallow parsing"* of the text and it is fundamental for the following stages of the analysis. When chunks are identified, subordinate clauses and basic chunks are considered main units when marking additional segments. This second stage consists of the grouping of different chunks following very precise criteria (i.e. only two subordinates clauses are allowed after the verb, two commas are allowed in specific patterns, etc.).

Example: *"Les exportations, par contre , ont presque doublé, atteignant 9,3 millions de rands en 1994, dont 7,8 millions attribuables aux céréales préparées."*(Exports, on the other hand, have nearly doubled, reaching 9,3 millions of rands in 1994, 7,8 millions of which attributable to prepared cereals.).

```
603>MAX{S{NP{Les exportations} PNP{, par contre ,} FV{ont
presque doublé} PFV{, GV{atteignant} NP{9,3 millions}
PP{de NP{rands}} PP{en NP{1994}} , BG{dont} NP{7,8 millions}
AP{attribuables} PP{aux NP{céréales}} préparées} .}}
```

Finally, sentences already marked with additional segments are filtered into first or second-tier sentences. First-tier sentences are those containing specific patterns, i.e. only basic segments, additional segments with or without two subordinates, non ambiguous coordinations, etc. This kind of sentence is marked with a "S".

Examples: *"Les deux filiales rachetées sont Morgan Stanley Trust Company et Morgan Stanley Bank Luxembourg SA."*(The buyed subsidiaries are Morgan Stanley Trust Company and Morgan Stanley Bank Luxembourg SA.) and *"Il a jugé qu' il y avait [TRADUCTION] atteinte acceptable aux droits des appelants."*(He has judged that there was [TRANSLATION] acceptable attack to the rights of appellants.).

```
1467>MAX{S{NP{Les deux filiales} PNP{rachetées} FV{sont}
```

```
PFV{NP{Morgan Stanley Trust Company} et NP{Morgan Stanley Bank
Luxembourg SA}} .}}

601>MAX{NP{Il} FV{a jugé} PFV{SBC{BG{qu'} NP{il} FV{y avait}}}
[ NP{TRADUCTION} ] atteinte AP{acceptable} PP{aux NP{droits}}
PP{des NP{appelants}} .}
```

The first example shows a first-tier sentence containing a non ambiguous coordination. The second sentence is an example of second-tier sentence marked with basic and additional segments but not considered "S" due to the brackets and quotes (phenomena to be analyzed specifically).

To sum up, the core grammar is the first module of the parser. Its aim is to mark basic and additionnal constituents in all the sentences of a given text -when possible- and to split all the sentences into first and second-tier, depending on the structure they have. First-tier sentences are sent to the dependency extraction module (see section *4.3*) and the result of their analysis is very high in terms of precision and recall rates (see section *4.4*) for most kind of corpora.

The sentences already marked with basic segments (and some additional segments if possible) but not considered first-tier -due to their structure or to the phenomena they contain- are sent to the specialized grammar modules.

4.2. *Syntactic diagnosis and specialized grammar modules*

The second-tier of the analysis is done by a set of specialized modules. After splitting the corpus into two kind of sentences, all the syntactic marks (chunks and groups of chunks) given by the core grammar are removed from second-tier sentences. Though some of these marks correspond to right limits of structures, some others do not due to the specific phenomena. A new marking up is done at this stage of the analysis taking the "non-standard" phenomena into account [14].

There are at present three kind of modules which contain specialized grammar rules. A first module performs a sort of *syntactic diagnosis* using formal criteria to identify certain phenomena. For instance, quotations and parentheses are here marked before any other syntactic treatment. Note that the analysis is made in different levels (each one is specialized in a specific structure) and that this permits to analyze different phenomena present in the same sentence.

Example: *"Avant les élections de 1994 , le Congrès National d'Afrique (ANC) , a élaboré un document intitulé " Plan de reconstruction et de développement " (RDP) qui a servi de plateforme électorale à ce parti pour faire connaître ses objectifs."*

14. Note that the marking of chunks and groups of chunks is an intermediate step of the parsing (essential for the extraction of syntactic dependencies) and that the dependencies are the final result of the parser.

(Before the 1994 elections, the African National Congress (ANC) has made a document entitled "Reconstruction and development plan" (RDP) that has served as a electoral basis to this party to made public their objectives.)

```
601>MAX{Avant les élections de 1994 , le Congrès National
d'Afrique PR[( ANC )] , a élaboré un document intitulé
QT[" Plan de reconstruction et de développement "]
PR[( RDP )] qui a servi de plateforme électorale à ce
parti pour faire connaître ses objectifs .}
```

A second module is in charge of marking basic syntactic units (chunks and groups of chunks) taking into account the already identified phenomena. The specialized rules in this module analyze specific phenomena dealing with the nature and the order of the elements of the sentence: interrogative and exclamative clauses, lists, imperative clauses, different kind of coordinations etc. In each sentence, every specific phenomena is given a special mark: CD (coordination), LST (list), etc.

Finally, a third module refines the analysis already performed in the previous stages. This module is a sort of *garbage-collector* which verifies that no specific phenomena has been forgotten by the precedent modules. It also analyses phenomena that have not been caught by the previous grammar rules such as, to give an example, sentences ending (but not starting) by a quotation (being part of a direct speech).

Here is an example of the annotated text after the syntactic analysis performed for *"Aussi, les éléments en notre possession font en effet apparaître un total de cession d'actifs en Europe, dans le monde et en France d' environ 800 milliards »."* (Furthermore, the elements we have show in fact a total in Europe, worldwide and in France of about 800 billions as far as the transfer of assets is concerned ».).

```
20>MAX{QT[Aussi, NP{les éléments} PNP{PP{en NP{notre possession}}}
FV{font} PFV{en effet IV{apparaître} NP{un total} PP{de NP{cession}}
PP{d' NP{actifs}} PP{en NP{Europe}} , PP{dans NP{le monde}}} et
PP{en NP{France}} PP{d' environ NP{800 milliards}} >>.]}
```

At the end of this second-tier analysis all the sentences have been precisely segmented into different kind of basic (NP, PP, etc.) and specific (QT, PR, etc.) chunks and groups of chunks. They are then sent to the dependency extraction module.

4.3. *Dependency extraction module*

After the first and second-tier modules, the last stage of the parsing consists of the extraction of syntactic relations between the words (usually the heads of the chunks). The extraction of dependency relations is performed by a set of rules that apply a syntactic pattern over the chunking tree and verify the order and the morphological features of each word involved in the potential relation.

There are at present thirteen dependencies which are extracted when applicable from all sentences (both first and second-tier), among others: SUBJ (subject of a verb), COMP (direct object), VPP (prepositional complement of a verb), COORDFV (coordinated FV) [15].

There is also a specific set of dependencies concerning "non-standard" phenomena which are applied to second-tier sentences when applicable, among others: IMP (imperative clause), COORD (coordinated elements when potential ambiguity), SGM (segment), etc. The final output of the parser is thus the segmented sentence followed by a list of syntactic relations as in *"Voir page 240 pour plus de détails".* (See page 240 for more details.):

```
IMP(voir)
COMP_INF(voir,page)
VPP(voir,pour,plus)
NPP(plus,de,details)

226>MAX{IV{Voir} NP{page 240} PP{pour NP{plus}} PP{de NP{détails}} .}
```

"Le geste chirurgical doit se limiter à une lobectomie ou une pneumonectomie , l'intervention doit être la plupart du temps simple". (The chirurgical gesture has to be a lobectomy or a pneumonectomy, the operation should be simple most of the times.):

```
SUBJ(doit,geste)
SUBJ(doit,intervention)
COMP_INF(doit,se_limiter)
COMP_INF(doit,être)
VADJ(être,simple)
VPP(limiter,à,COORD)
COORD(lobectomie,ou,pneumonectomie)
NADJ(geste,chirurgical)

1651>MAX{NP{Le geste} PNP{AP{chirurgical}} FV{doit} PFV{IV{se limiter}
CD[PP{à NP{une lobectomie}}} ou NP{une pneumonectomie}], NP{l'intervention}
FV{doit} PFV{IV{être} AP{la plupart du temps simple}} .}
```

4.4. *Evaluation*

We have conducted an evaluation of the results obtained by the core grammar, by the specialized modules (for the analysis of quotations and parentheses only) and by the dependency extraction module.

15. Some dependencies have a feature added to be more explicit (subject of a passive verb SUBJ_PASS, object of an infi nitive COMP_INF etc.).

The corpora used is a set of heterogeneous texts extracted from the Web coming from the following sources: general newspapers (*Le Monde*, *L'Humanité*); financial newspaper (*Les Echos*); technical manual for a motorbike; scientific report on energy; oral transcription of a dialog; legal reports. The overall corpora contain about 80 000 words and about 3 600 sentences (with an average of 22 words per sentence) [16].

For our evaluation, we have selected 245 sentences at random (representative of all the mentioned domains). After parsing this sample, 54 sentences have been marked as first-tier sentences (22 % of the corpus). We have also parsed the overall corpora and obtained 18,9 % of coverage by the core grammar: about 700 sentences are considered first-tier.

Taking these figures into account, the coverage of the core grammar can be guaranteed at around 20 % for any set of heterogeneous corpora from different domains (except corpora not respecting certain structural constrains, i.e. oral transcriptions with no punctuation at all).

The evaluation of the linguistic accuracy of the parser has focused on the set of 245 randomly extracted sentences. First, we have manually checked the precision and recall rates of the 22 % of first-tier sentences. For chunking we have evaluated the accuracy of the chunks and groups of chunks; for dependency extraction we have taken into account the presence of subject relations: 115 dependencies extracted (3 wrong), 11 missed.

As for the other 88 % of the sample corpus (second-tier sentences), we have analyzed with specialized grammars those sentences presenting parentheses and/or quotations: 53 sentences (that is 21,6 %) of the sample corpus. The manual evaluation of these sentences has focused on the precise chunking of the punctuation marks (quotes and parentheses) as well as on subject extraction taking into account the presence of two-tier phenomena.

The following table summarizes the results obtained. Last row shows the present coverage of the parser as well as the overall precision and recall rates for constituency (*const.*) and subject relation (*subj.*):

	Coverage	*P const.*	*R const.*	*P subj.*	*R subj.*
first-tier	22 %	97,3 %	98,6 %	97,4 %	91,2 %
second-tier (QT and PR)	21,6 %	96,5 %	80,4 %	95,9 %	79,7 %
total	**43,6 %**	**96,9 %**	**89,5 %**	**96,7 %**	**85,4 %**

Table 8. Results of our modular parsing approach for first and second-tier sentences.

Compared to the results obtained by some parsers (see *2.4*) the results on precision and recall for phrase structure analysis as well as for subject extraction are much better.

16. Note that the corpora used for the development of the grammars is another set of varied corpora of about 87 000 words.

The coverage of the grammar is still insufficient but other specific phenomena are currently under development to obtain a maximal coverage for most kinds of corpora.

Some work is also currently being done to improve recall rates, specially in second-tier sentences, as our goal is to push them at the same level of acceptation than precision rates (at least 95 %).

5. CONCLUSIONS

After giving a general overview of robust parsing, this paper describes different available robust parsers focusing on their linguistic performance for syntactic annotation of corpora. The evaluation of some of these systems shows that their precision and recall rates decrease when the input text comes from specialized domains presenting "non-standard" phenomena, i.e. phenomena which are not precisely enough (or not at all) taken into account by the grammars of these existent parsers.

In this paper, in order to tackle variety in raw corpora, we have proposed an approach based on the decomposition of the grammar modules. This approach relies on the splitting of the corpora in first and second-tier sentences depending on the features these sentences present (particular punctuation, order of the constituents, etc.). A specific treatment is given to second-tier sentences in order to handle phenomena requiring concrete chunking and dependency extraction.

The evaluation shows very encouraging results on precision and recall rates for the different corpora used for the experiment (coming from a variety of both general and specialized domains).

As far as future research is concerned, it will be devoted to increasing the coverage of the parser by creating the specialized modules for other phenomena mentioned in this paper. Another line of research will be the use of the accurate results obtained from the core grammar module as training corpora to improve the extraction of non deterministic syntactic relations (PP attachment) by applying symbolic machine learning methods.

6. Acknowledgements

This work is part of a PhD research financed by a CIFRE grant. The author wants to express her gratitude to C. Jacquemin, G. Grefenstette and S. Aït-Mokhtar for their continuous help and detailed insights. The author alone is responsible for any errors.

7. Bibliographie

[ABE 99] ABEILLÉ A., BLACHE P., « Grammaires et analyseurs syntaxiques », *Traité IC2, Volume Ingénierie des Langues,* , 1999.

[ABN 91] ABNEY S., « Parsing by chunks », BERWICK R., ABNEY S., TENNY C., Eds., *Principle-Based Parsing*, Academic Publishers, 1991.

[ABN 94] ABNEY S., « Partial Parsing », *Tutorial given at ANCL-94, http://www.sfs.nphil.uni-tuebingen.de/~abney/Papers.html*, Stuttgart, 1994.

[ABN 96] ABNEY S., « Partial Parsing via Finite State Cascades », *Proceedings of the ESSLLI 96 Robust Parsing Workshop*, Prague, 1996.

[AÏT 97] AÏT-MOKHTAR S., CHANOD J., « Incremental Finite-State Parsing », *Proceedings of ANLP-97*, Washington, 1997.

[AÏT 98] AÏT-MOKHTAR S., GALA PAVIA N., « A description of the English Finite-State Parser Architecture », rapport, 1998, Xerox Research Centre Europe (XRCE), France.

[APP 93] APPELT D., JERRY H., JOHN B., ISRAEL D., TYSON M., « 'FASTUS': a finite-state processor for information extraction from real-world text », *Proceedings of IJCAI-93*, Chambéry, France, 1993.

[BAS 92] BASILI R., PAZIENZA M. T., « A shallow syntactic analyzer to extract word associations from corpora », Cargese, Corse, 1992, p. 114-124.

[BAS 99] BASILI R., PAZIENZA M. T., ZANZOTTO F. M., « Lexicalizing a shallow parser », *Proceedings of TALN'99*, Cargese, Corse, 1999.

[BIB 93] BIBER D., « Using Register-Diversified Corpora for General Language Studies », *Computational Linguistics 19(2)*, 1993, p. 219-241.

[BLO 33] BLOOMFIELD L., *Language*, Holt, (second edition 1961), New York, 1933.

[BRI 94] BRISCOE T., « Parsing (with) Punctuation etc. », rapport, 1994, Rank Xerox Research Centre, Grenoble.

[BRU 98] BRUN C., « Etude et Implantation de la coordination en vue de l'analyse automatique du français écrit dans le cadre de la Grammaire Lexicale Fonctionnelle », PhD thesis, Université Grenoble III et XRCE, France, 1998.

[CHA 94] CHANOD J. P., « Développements en Analyse Syntaxique Automatique », *Proceedings of TALN-94*, Marseille, 1994.

[CHA 96] CHANOD J.-P., TAPANAINEN P., « A Robust Finite-State Parser for French », *ESSLII'96 Robust Parsing Workshop*, Prague, 1996.

[CHA 00] CHANOD J.-P., « Robust Parsing and Beyond », NOORD G. V., JUNQUA J., Eds., *Robustness in Language Technology*, Kluwer, 2000.

[CHU 88] CHURCH K., « A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text », *Proceedings of the 2nd Conference on Applied Natural Language Processing, ,* 1988, p. 136-143.

[DAE 99] DAELEMANS W., BUCHHOLZ S., VEENSTRA J., « Memory-based shallow parsing », *Proceedings of CoNLL*, Bergen, 1999.

[DEB 82] DEBILI F., « Analyse Syntaxico-Sémantique Fondée sur une Acquisition Automatique de Relations Lexicales-Sémantiques », PhD thesis, Université Paris XI, France, 1982.

[DIN 99] DINI L., DI TOMASO V., SEGOND F., « Ginger II, an example-driven word sense disambiguator », *Computers and Humanities, special issue*, 1999.

[EJE 93] EJERHED E., « Nouveaux courants en analyse syntaxique », *T.A.L.*, vol. 34, n° 1, 1993.

[FIL 68] FILLMORE C. J., « The Case for Case », BACH E., HARMS R., Eds., *Universals in Linguistic Theory*, p. 1-88, New York, 1968.

[GAL 99] GALA PAVIA N., « Using the Incremental Finite-State Architecture to create a Spanish Shallow Parser », *Proceedings of XV Congres of SEPLN*, Lleida, Spain, 1999.

[GAL 00] GALA PAVIA N., « Hétérogénéité des corpus: vers un parseur robuste reconfi gurable et adaptable », *RECITAL-00 (TALN student session)*, Lausanne, 2000.

[GAU 00] GAUSSIER E., GREFENSTETTE G., HULL D., ROUX C., « Recherche d'information en français et Traitement Automatique des Langues », *Tal vol. 41, n. 2*, , 2000, p. p. 473-493, Hermes.

[GIG 97a] GIGUET E., VERGNE J., « From Part-of-Speech tagging to Memory-based deep syntactic analysis », *Proceedings of IWPT-97*, 1997.

[GIG 97b] GIGUET E., VERGNE J., « Syntactic Analysis of unrestricted French. », *Proceedings of the International Conference on Recent Advances in NLP, RANLP-97*, Tzigov Chark, Bulgaria, 1997.

[GRE 92] GREFENSTETTE G., « SEXTANT: Exploring Unexplored Contexts for Semantic Extraction from Syntactic Analysis », *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics, ACL-92*, Newark, Delaware, 1992.

[GRE 95] GREFENSTETTE G., « Low-Level Parsing applied to Technical Manuals », *Proceedings of IPSM'95, Industrial parsing of Software Manuals*, Limerick, Ireland, May 4–5, 1995.

[GRE 99] GREFENSTETTE G., « Shallow Parsing Techniques Applied to Medical Terminology Discovery and Normalization », *Proceedings IMIA WG6, Triennial Conference on Natural Language and Medical Concept Representation*, Phoenix, USA, 1999.

[HAL 94] HALLIDAY M. A. K., *Introduction to Functional Grammar*, Edward Arnold, second edition, London, 1994.

[HIN 83] HINDLE D., « User manual for Fidditch, a deterministic parser », *Naval Research Laboratory Technical Memorandum 7590-142.*, , 1983.

[HUD 94] HUDSON R., « Word Grammar », ASHER R., Ed., *The Encyclopedia of Language and Linguistics*, Pergamon Press, 1994.

[ILL 00] ILLOUZ G., « Typage de données textuelles et adaptation des traitements linguistiques. Application à l'annotation morpho-syntaxique », PhD thesis, Université Paris-Sud, UFR Scientifi que d'Orsay, 2000.

[JAC 99] JACQUEMIN C. ET TZOUKERMANN E., « NLP for term variant extraction: synergy between Morphology, Lexicon and Syntax », STRZALKOWSKI T., Ed., *Natural Language Information Retrieval*, Kluwer Academic, 1999.

[JAR 97] JARVINEN T., TAPANAINEN P., « A dependency parser for English », rapport, 1997, TR-1, Department of General Linguistics, University of Helsinki, Montreal.

[JEN 88] JENSEN K., « Why computational grammarians can be skeptical about existing linguistic theories », *Proceedings of COLING-88*, Budapest, 1988.

[JON 94] JONES B. E. M., « Exploring the role of punctuation in parsing natural text », *Proceedings COLING-94*, Kyoto, Japan, 1994.

[JOS 96] JOSHI A., « A Parser from Antiquity: An Early Application of Finite-State Transducers to Natural Language Parsing », *Proceedings ECAI '96, workshop on extended finite state models of language*, Budapest, 1996.

[KAH 00] KAHANE S., « Les grammaires de Dépendance », *Traitement Automatique des langues*, vol. 42, n° 1, 2000.

[KAR 95] KARLSSON F., VOUTILAINEN A., HEIKKILÄ J., ANTTILA A., *Constraint Grammar. A Language-Independent System for Parsing Unrestricted Text*, Mouton de Gruyter, Berlin and New York, 1995.

[MEL 88] MEL'CUK I., *Dependency Syntax: theory and practice*, Albany, 1988.

[MUN 99] MUNOZ M., PUNYAKANOK V., ROTH D., ZIMAK D., « A learning approach to shallow parsing », *Proceedings EMNLP-WVL'99*, 1999.

[PER 83] PERLMUTTER D. M., *Studies in Relational Grammar 1*, University of Chicago Press, Chicago, 1983.

[ROC 93] ROCHE E., « Analyse syntaxique transformationnelle du français par des transducteurs et lexique-grammaire », PhD thesis, Université Paris 7, 1993.

[SAB 90] SABAH G., « CARAMEL: un système multi-experts pour le traitement automatique des langues », *Modèles Linguistiques*, vol. 12, n° 1, 1990, p. 95-118.

[SLE 93] SLEATOR D., TEMPERLEY D., « Parsing English with a Link Grammar », *Proceedings of the 3rd International Workshop on Parsing Technologies, IWPT-93'*, 1993.

[SRI 97] SRINIVAS B., « Complexity of lexical descriptions and its relevance to partial parsing. », PhD thesis, University of Pennsylvania, 1997.

[TAP 97] TAPANAINEN P., JARVINEN T., « A non-projective dependency parser », *Proceedings of ANLP-97*, Washington, 1997.

[TAP 99] TAPANAINEN P., « Parsing in two frameworks: finite-state and functional dependency grammar », PhD thesis, University of Helsinki, Language Technology, Department of General Linguistics, Faculty of Arts, 1999.

[TES 59] TESNIÈRE L., *Elements de syntaxe structurale*, Hachette (3e éd. 1979), Paris, 1959.

[VOU 96] VOUTILAINEN A., JARVINEN T., « Using the English Constraint Grammar Parser to analyse a software manual corpus », *Industrial Parsing of Software Manuals*, Language and Computers: studies in practical linguistics, 17, Rodopi, Amsterdam, 1996.

[WHI 95] WHITE M., « Presenting Punctuation », *Proceedings from European Workshop on Natural Language Generation*, Leiden, Pays Bas, 1995.