Syntactic Analysis of Unrestricted French

Emmanuel Giguet and **Jacques Vergne**
GREYC — CNRS UPRESA 6072 — Université de Caen
14032 Caen cedex — France
{Emmanuel.Giguet,Jacques.Vergne}@info.unicaen.fr

Abstract

This paper describes a practical parser for unrestricted relations. These relations are computed between chunks and specified within a unique formalism. They are represented by means of labeled and directed links. The present implementation handles dependency, coordination and antecedence relations.

1 Introduction

This paper describes a robust system for syntactic parsing of unrestricted French. In this system, syntactic parsing means identifying constituents, called non-recursive phrases (*nr*-phrases), and linking them together.

Our work is based on the work of Lucien Tesnière (Tesnière 59) but we have derived our own concepts for specifying any kind of *nr*-phrases relations in order not to be restricted to dependencies anymore. In our research, we have emphasized the handling of relations interdependencies since mastering the propagation of linking constraints guarantees both the global coherence of the parse and the mastering of combinatorial explosion. In our system, we have implemented all major dependency relations, the coordination relation and the antecedence relation.

Hereafter, we first describe the architecture of the parser. Then, after this general presentation, we emphasize ways of linking *nr*-phrases with different kinds of relation using a unique formalism. Implementation details are then presented. Finally, a precise evaluation on subject-verb relations is carried out, empirically demonstrating the adequacy of the approach.

2 The Architecture

The architecture of the process combines two techniques: (1) Part-Of-Speech Tagging and Chunking techniques at word-level that build a constituent structure (each constituent is an *nr*-phrase); (2) linking rules at *nr*-phrase-level that link *nr*-phrases to build a functional structure. In our approach, both constituent and functional structures are build simultaneously by two interacting processes. The analysis is carried out as shown in figure 1.

Figure 1 shows two processes, labelled 1 and 2, managing respectively the word-level and the *nr*-phrase-level. The first process assigns tags to each part-of-speech and defines *nr*-phrase boundaries, shown as square brackets. The second process defines relations between *nr*-phrases. The two labels α and β show the interactions between word-level and *nr*-phrase-level. The execution of an entire basic cycle of deductions is successively: $1, \alpha, 2, \beta$.

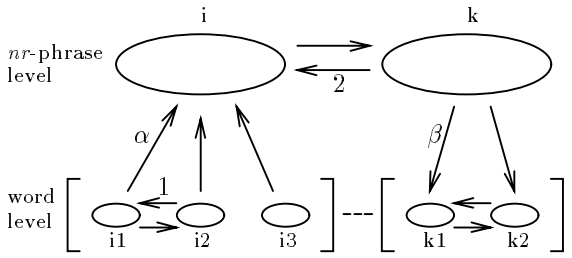


Figure 1: Process of analysis

The aim of this paper is to focus on the functional structure so that we concentrate on *nr*-phrase level, i.e., linking *nr*-phrases.

3 Linking Non-Recursive Phrases

The linguistic background of our research is based on the work of Lucien Tesnière (Tesnière 59) but it revises the notion of dependency as a relation between *nr*-phrases, and not between words. This feature distinguishes our approach from many other dependency-based parsing approaches (Covington 90; Sleator & Temperley 93; Tapanainen & Järvinen 97). As said in (Abney 96), "*By reducing the sentence to chunks [i.e.,* nr-*phrases], there are fewer units whose associations must be considered, and we can have more confidence that the pairs being considered actually stand in the syntactic relation of interest, rather than being random pairs of words that happen to appear near each other*".

3.1 Parsing as a Constraint Satisfaction Problem

Dependency grammar-based formalisms usually allow the specification of dependencies (1) using constraints on the two structures considered in the relation of interest (e.g., an NP and a VP for a subject-verb relation) and (2) using constraints existing between these two structures (i.e., agreement in person and number). This way of specifying relations leads to a failure since either these constraints are too relaxed and the noise is high, or they are too strict and the silence is high.

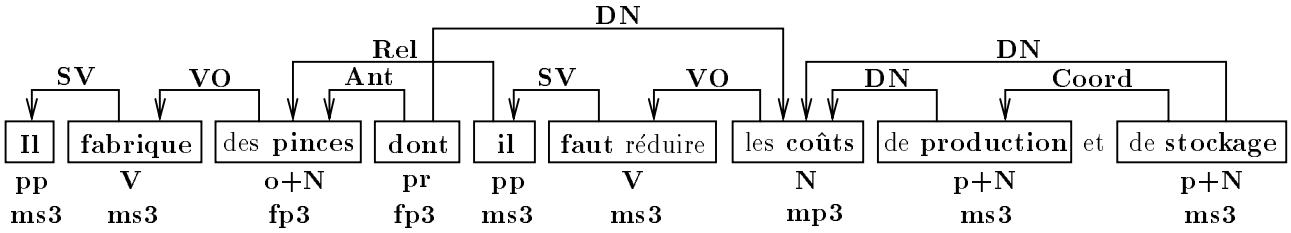


Figure 2: Syntactic Analysis with Dependency, Antecedence and Coordination Relations

As such static constraints on structures are unavoidable, the parsing process needs other knowledge to become effective. Several proposals have been made to reach this goal, such as: introducing possible or impossible occurences of structures between the considered items; or using a maximum distance between the items. All these proposals can be proved inadequate within any unrestricted corpus.

Specifying a relation solely with constraints on structures is not enough to get an effective parser. Parsing implies specifying and dynamically handling linking contraints: items selected to instanciate a new relation depend on the linking constraints defined by the previously computed relations; this new relation generates new linking constraints that have to be taken into account when instanciating further relations. Parsing can be seen as a Constraint Satisfaction Problem.

3.2 Propagating Linking Constraints

In our research, we have emphasized the handling of relations interdependencies which has become the predominant feature of our system. In other words, we have studied how the instanciation of a relation reduces the complexity of further decisions by discarding potential choices.

An example illustrates this general concept. Considering the sentence: "*[The flight] [from Paris] [is cancelled] [because of a strike].*"

By instanciating a subject-verb relation from *is cancelled* to *The flight*, a constraint is generated: *from Paris* can not be the governor of any other relation. Thus, the governor of *because of a strike* can only be *The flight* or *is cancelled*.

3.3 Handling Linking Constraints

Mastering the propagation of linking constraints guarantees both the global coherence of the parse and the mastering of combinatorial explosion. Even, if these goals seem important, handling or not handling linking constraints depends on the aim of the research:

(Covington 90) who deals with free word order languages is not concerned with linking constraints since none can easily be pointed out in such languages. (Sleator & Temperley 93) handle linking constraints with an input restriction on parseable sentences called "*planarity*": links do not cross when drawn above the words. Therefore, the formalism and the parsing process are designed according to this restriction.

The aim of our research includes being able to deal with any kind of syntactic phenomena, including non-projective[1] sentences (figure 2). Thus, we are not concerned with restrictions such as "*planarity*" or "*projectivity*". Our parser deals with these natural phenomena with the help of linking constraints propagation.

Furthermore, such restrictions which often lead to in-built parser restrictions are an impediment to the future processing of other kinds of relations.

3.4 From Dependency Relations to Unrestricted Relations

From Tesnière's first approach to dependency definition "*Between a word and its neighbours, the mind foresees some connections.*", we have derived our own concepts to process any kind of relations, in order to put in evidence different kinds of syntactic phenomena such as coordination and in order not to be restricted to dependencies anymore.

To be able to parse other kinds of relations, we have extended our rule-based declarative control in order to flexibly handle the requirement of each kind of relations such as head uniqueness restricted to dependency relations. Thus, the formalism which specifies the computation of the different kind of relations is unique.

Generalization of dependencies to other kinds of relations leads toward the design of an open architecture for parsing unrestricted relations. In our system, we have implemented dependency relations, coordination relation (labeled Coord in fig 2) and antecedence relation (labeled Ant in fig 2).

4 The parser

In this section, we are about to explain the parsing process: the creation of new relations, the propagation of linking constraints. The analysis is carried out from left to right and is deterministic. The parser input is a sequence of *nr*-phrases which may be ambiguous. For incremental design and robustness purposes, partial parser outputs are valid.

[1] "*Projectivity*" defined in (Mel'čuk 88) includes "*planarity*": a sentence is projective if it is planar and if no dependency covers its head. Mel'čuk points out that "*most sentences of a language are projective*".

if the current *nr*-phrase
is a nominal *nr*-phrase and
is not object and
is not already subject and
is not attached to a preposition
then
it is stored as possible subject
into the subject-verb memory.

if the current *nr*-phrase
is a verbal *nr*-phrase and
there are possible subjects
in the subject-verb memory
then
retrieve the best-fit subject from the memory
attach the verb to this subject,
discard this subject from the memory,
discard items located between the subject and
the verb from every memory.

Figure 3: A concrete example: Handling a subject-verb relation with two rules

4.1 Implementation of the linking process

The process is both data-driven and declarative: condition-action rules do not describe syntactic structures but the linking process. These rules manage both relations instanciation and propagation of linking constraints. Relation instanciations are achieved in two distinct steps by two distinct kinds of rule actions:

1. *store* an *nr*-phrase as a candidate for some particular relations of interest in relevant memories,
2. *attach* one *nr*-phrase to another located in a memory and *discard* some particular items which are possible candidates for some particular relations from the relevant memories.

Figure 3 gives a concrete example of two rules written to handle a subject-verb relation.

Building up the syntactic structure is constrained by the interactions of the rules through memories:

Instanciating a particular relation between two *nr*-phrases is only possible if one of the two *nr*-phrases is stored in a memory. To be stored in a memory, (1) the item must have been considered as a potential candidate for a future relation, but moreover (2) it should not have been discarded by an other rule propagating linking constraints generated by the instanciation of a new relation. In fact, discarding items in memories is a propagation of linking constraints: this corresponds to the death of potential relations. For instance, in figure 3 rule 2, no more relation will cross the new subject-verb relation since all the items located between the subject and the verb are discarded from every memory. Adding such a constraint to every rule that creates a new dependency relation would lead to implement the "*planarity*" constraint (see section 3.3).

The rules conditions allow the manipulation of: (1) relations in the syntactic structure in progress; (2) heads of *nr*-phrases; (3) features of *nr*-phrases; (4) and status of the memories.

Rules actions are: (1) actions on a memory (storing one *nr*-phrase and linking two *nr*-phrases, discarding an item from a memory, erasing the content of a memory), (2) actions on an *nr*-phrase (changing/adding a feature).

The coherence of the structure which is built up can be controlled with the help of a query language on the whole syntactic structure in progress. For instance, in figure 3 rule 1, a basic query is used to check if the nominal *nr*-phrase is neither an object, nor a subject. Any kind of complex queries can be written in order to navigate in the structure in progress and check properties.

The current implementation requires the system to store candidates into memories for possible expectations (e.g, nominal *nr*-phrase possibly expecting a verb for a subject-verb relation) but also to retrieve the best-fit candidate from a memory. This ability is provided by the memory-based framework.

4.2 Memory-Based Framework

4.2.1 Memories as favoured places to perform relations

The process is based on a set of memories. Each memory is dedicated to the management of one specific relation (e.g, subject-verb, verb-object, coordination, PP attachment). A memory contains *nr*-phrases whose association with a future *nr*-phrase must be considered. For instance, the memory that manages the subject-verb dependency relation contains nominal *nr*-phrases which can be involved in a future relation with a verbal *nr*-phrase.

The power of such an approach is that all relevant candidates are together in a single location when the relation has to be computed (a memory is a limited search-space): for a specific relation, the relevant knowledge sources can choose a successful candidate more accurately (see section 4.2.2).

Moreover, when the selection has to be performed, the process does not have to consider the past of the analysis but the current state of the memories. Therefore, far discontiguous relations are handled the same way as contiguous relations (if necessary, there are ways to distinguish them).

An other interesting point is that memories contain candidates for an association with a future *nr*-phrase. No requirement is made on the presence of this *nr*-phrase. If such an *nr*-phrase does not occur before

the end of the sentence, the memory is erased: the candidates are forgotten. In other words, when a new *nr*-phrase is added to a memory, no explicit expectation on structure is done, only implicit expectations are described by the rules. For instance, when parsing the title "*Selection in a memory*", *Selection* is stored into the subject-verb memory but no verb will occur so that it will be discarded at the end of the sentence. This kind of behaviour is to be related to tagging techniques and is fundamental to deal with unrestricted text.

4.2.2 Selection in a memory

Each memory is dedicated to the management of a specific relation. It is obvious that the knowledge required for selecting a candidate in the different memories is not always the same. In this system, every memory has its own specific method for choosing the successful candidate.

For instance, in our system, syntactic knowledge is involved for constraining the search space (i.e., the memory) depending on number, person and gender in a subject-verb dependency relation; similarity of structures is considered for coordination relation; psycholinguistic knowledge constrains the distance between the future associated *nr*-phrases.

It is interesting to point out that the above-mentioned knowledge sources are not sufficient to deal with complex phenomena. In memories, semantic and pragmatic knowledge sources can also interact with other knowledge sources to constrain the search space.

Furthermore, barriers can bound searches to stop-words such as relative pronouns or subordination conjunctions. These barriers are dynamically activated and disactivated depending on the syntactic structure in progress.

4.2.3 Focusing on the Subject-Verb memory

It is interesting to show in a concrete way how modularity of memories leads to flexibility, and to clarify how it helps us mastering the triggering of adequate knowledge sources and which items the triggered sources will act upon.

The subject-verb memory is an example of such a memory where several kinds of knowledge are combined in order to handle the corresponding relation in a reliable and robust way. We will see that the relevant knowledge which deals with subject selection is clearly located in a single place:

- Syntactic constraints on agreement: these constraints are based on coordination relations, person and number of *nr*-phrases.

- Structural constraints on *nr*-phrase: they are involved in specific configurations in order to favour subject with determiner rather than subject without determiner.

- Basic semantic constraints are used to prevent some particular temporal NP from being taken as subject.

- This memory selects the leftmost possible subject close to the first active barrier located on the left-hand side of the verb. This models the linking process of a subject with its verb, taking into account embedded clauses.

The latter shows the tight links between memories and the dynamic linking process which feeds them.

Selection in memories is usually achieved with the help of a standard constraints relaxation mechanism.

5 Evaluation

The evaluation we offer is restricted to subject-verb relations since no french treebank is available yet. However, it is possible to use our syntactic parse viewer on internet at http://www.info.unicaen.fr/~giguet (for Java-enabled browsers) in order to have an idea of the parser reliability for other relations.

5.1 Corpus Metrics

The evaluation of the parser has been carried out on a set of articles from the newspaper "Le Monde". This corpus has not been used to build up the parsing rules. This set is made of 24 articles (dealing with politics, economics, fashion, high-technology, home news, ...) representing 474 sentences (max. length: 82 words, avg. length: 24.43 words). The definition of sentence is standard but includes two additional boundaries ";" and ":". Figure 4 gives an overview of the sentences length.

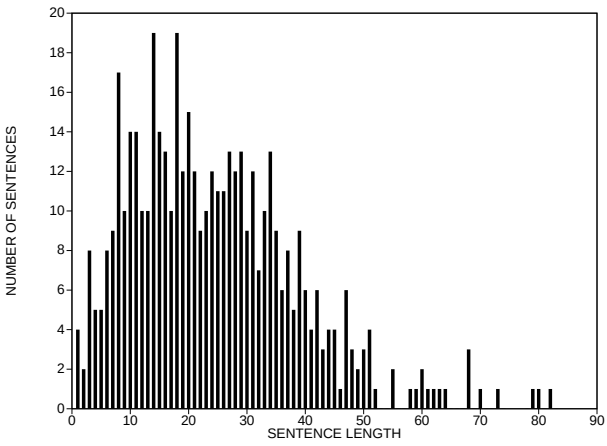
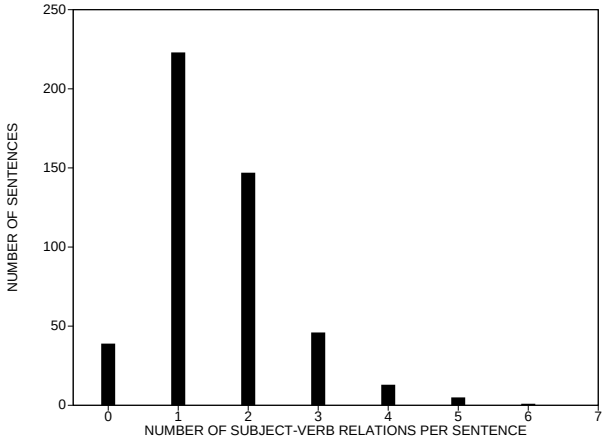
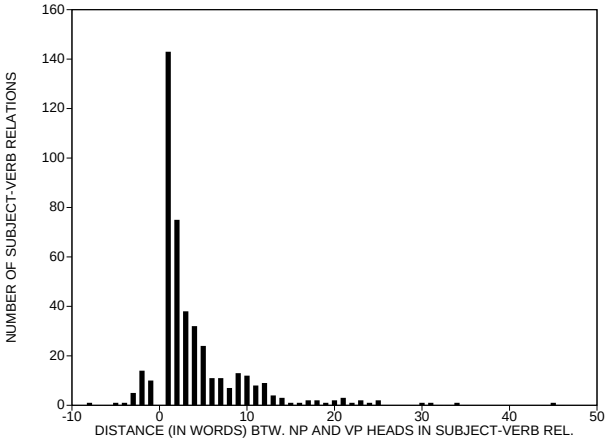


Figure 4: Sentence Length



nature of subject	number
NP	458
Infinitive VP	2
Relative Pronoun	85
Personal Pronoun	193
Total	738

Figure 5: Subject-Verb Relations



Max. dist. btw. NP-subject and VP heads	
in standard relations	45 words
in inverted relations	8 words

Figure 6: NP-subject in Subject-Verb Relations

5.2 Relation computation evaluation

5.2.1 Subject-Verb relations in the corpus

In this corpus, there are 738 Subject-Verb relations. Figure 5 shows the span of subject-verb relations in the sentences: 39 sentences do not have subject verb relations and the maximum number of subject-verb relation per sentence is 6. According to the nature of the subject, we distinguish 4 kinds of SV relations in the corpus: relations involving (1) an NP subject, (2) an Infinitive VP subject, (3) a Relative Pronoun subject and (4) a Personal Pronoun subject.

An other interesting metric is the distance between the verb and its subject. Figure 6 illustrates this phenomenon only for NP-subject. This metric is less relevant for other relations, even if several sentences contain personal pronouns and relative pronouns that are far subjects, for instance in cases of verb enumeration or prepositional phrase insertion. The figure shows that the distance between an NP-subject and its verb can reach up to 45 words.

5.2.2 Evaluation on Subject-Verb Relations

The evaluation function is based on the following principle: every verb has to be attached to no more than one subject. From this starting point, 3 cases exist: it is a *correct* relation if the verb is attached to the expected subject (the two *nr*-phrases heads also have to be correct), *incorrect* if not and a *silence* if no subject is provided but one was expected.

In cases of subjects coordination, each verb depending on the coordination has to be attached to the head of this coordination, that is, to the head of the first item. In cases of verbs coordination, one correct relation counts for each verb attached to the expected subject and one incorrect relation for each verb attached to an unexpected subject.

The results are listed in Figure 7, on the following page. *Precision* is the ratio of correct links over the number of computed links. *Recall* is the ratio of correct links over the number of expected links. The reported rate (96.39% precision and 94.04% recall) empirically validates our approach.

Our results can still be improved since this evaluation was the first on large corpora. The 42 silences and incorrect relations can be classified in 5 categories: (1) incorrect implementation of agreement check, (2) illformed *nr*-phrases, (3) coordination not found, (4) inverted subject in reported speech, (5) incorrect *nr*-phrase tags. We have pointed out better ways of solving the three first classes. The fourth and fifth classes require further studies to be carried out in a general way.

6 Conclusion and Future Work

We have described a system for syntactic parsing of unrestricted French. The analysis is carried out while (1) maintaining the global coherence of the syntactic structure and (2) mastering combinatorial explosion. This is achieved thanks to the propagation of linking constraints and the use of a query language on the structure in progress.

The result is a flexible architecture which has the ability to put in evidence different kinds of syntactic phenomena described within a unique framework. Dependency, coordination and antecedence relations are implemented.

Running on a collection of newspaper articles from "Le Monde" (11583 words, 474 sentences and 738 sub-

Nature of subject	number	correct	incorrect	silence	precision	recall
NP	458	418	26	14	94.14%	91.27%
Infinitive VP	2	2	0	0	100.00%	100.00%
Relative Pronoun	85	85	0	0	100.00%	100.00%
Personal Pronoun	193	191	0	2	100.00%	98.96%
Total	738	694	26	16	**96.39%**	**94.04%**

Figure 7: Evaluation on Subject-Verb Relations

ject verb relations) where complex structures appear, we get 96.39% precision and 94.04% recall for subject-verb relations. These first results empirically validate the approach and we can say the parser is reliable for this relation. Moreover, it is robust since one parse is always provided (sometimes a partial parse). The present version of the linking process is efficient: it is deterministic and it has a linear complexity in time. Today, we are working on a slightly modified version of the parsing process in order to enable new knowledge to change past deductions. In this case, these deductions and their consequences are discarded.

We now have to continue precise evaluation of our parser for all the other kinds of relations and to continue improving the parser. A demo is available at http://www.info.unicaen.fr/~giguet.

References

(Abney 96) Steven Abney. Part-of-speech tagging and partial parsing. In Ken Church, Steve Young, and Gerrit Bloothooft, editors, *An Elsnet Book*, Corpus-Based Methods in Language and Speech. Kluwer Academic, Dordrecht, 1996.

(Covington 90) Michael A. Covington. A dependency parser for variable-word-order languages. Technical Report AI-1990-01, Artificial Intelligence Programs, The University of Georgia Athens, Georgia 30602 USA, January 1990.

(Mel'čuk 88) Igor A. Mel'čuk. *Dependency Syntax: theory and practice*. State University of New York Press, Albany, 1988.

(Sleator & Temperley 93) Daniel D. Sleator and Davy Temperley. Parsing english with a link grammar. In *Proceedings of the Third International Workshop on Parsing Technologies (IWPT'93)*, pages 277–292, Tilburg, Durbuy, August 1993.

(Tapanainen & Järvinen 97) Pasi Tapanainen and Timo Järvinen. A non-projective dependency parser. In *Proceedings of the fifth Conference on Applied Natural Language Processing (ANLP'97)*, pages 64–71, Washington, DC USA, April 1997. Association for Computational Linguistics.

(Tesnière 59) Lucien Tesnière. *Éléments de syntaxe structurale*. Klincksieck, Paris, 1959.