Synergy of syntax and morphology in automatic parsing of French language with a minimum of data

Feasibility study of the method

Jacques Vergne Pascale Pagès
Inalco Paris

Abstract:

We intend to present in this paper a parsing method of French language whose particularities are: a multi-level approach: syntax and morphology working simultaneously, the use of string pattern matching and the absence of dictionary. We want here to evaluate feasibility of the method rather than to present an operational system.

I. General objectives:

We intend to demonstrate that it is possible to parse texts with very few data: only determiners, indefinite and numeral adjectives, conjunctions and prepositions, that in such an analysis, the consecutive application of morphology and syntax is insufficient, but that the use of syntax and morphology simultaneously is very efficient, and at last that the notion of a grammatical or lexical category of the word is not attached to the word, but depends on the local situation of the word in the sentence.

II. Comments upon objectives:

A. Comparison with the classical parsing strategies:

In nearly every contemporary automatic parsing system, we have a chronology of several steps: one step by linguistic level: the morphological step, then the syntactical step, and then eventually the semantic step. The morphological step uses an exhaustive dictionary of forms or sometimes of lemmas.

But we know that the human understanding uses simultaneous information coming from all levels. Some parsers now begin to use two levels simultaneously: for example the system ASOF (SFB 100 Sarrebrück, see reference) uses semantic information coming from semantic nets in the syntactical analysis and semantic constraints in the syntactical rules.

We propose to use morphology and syntax simultaneously in a parser with no dictionary.

B. Assumptions and theoretical aspect:

By "parsing" we mean to recognize adjectives and nouns with their gender and number, verbs in the infinitive or in the present participle and the adverbs derived from adjectives, to produce the lexicon of the text and to determine syntactical relations between words.

The possible applications of such a parser are: automatic indexing and, also as a first step in every system which uses a parser: inquiry system, automatic translation, filling knowledge bases from texts, etc., particularly when the parser must work in open semantics.

The main practical interests are to avoid consulting an exhaustive dictionary in the first step of the parsing and to process the neologisms exactly the same way as the other words. Dictionaries are expensive to update and in fact are never completely updated; consulting a dictionary produces many artefact-ambiguities locally to the word, which are cancelled as soon as the immediate context is examined; these ambiguities produce a combination-explosion and much redundant processing.

There are also some theoretical aspects. If the lexical category can be deduced with no dictionary, is this category really lexical? We could rather name it a contextual or functional category or even a function: a word gets its category from the flow of the text and the dictionary gives the usual categories. The lexical category then is only the regularity of the function.

Further, any word can potentially have any category (perhaps more in English than in French). Claude Hagège: "ce sont des fonctions, non des parties du discours, qu'il convient d'abord de poser" (L'homme de paroles 1965 page 137). A more general theoretical aspect lies in the use of computer as an experimental tool which allows now to consider linguistics as an experimental science and to use the experimental method to test linguistic theories.

III. The means:

A. General method:

The main method is pattern matching. The principle is the following: the shape to recognize is compared to a set of patterns until a match is found.

But what exactly is a shape in a natural language?

The classical terms are "morphology" for the word, and "syntax" for the sentence. We could say that morphology is the shape of the word and syntax the shape of the sentence; and more, we propose here to fill conceptually the gap between the level morphology-word and the level syntax-sentence: morphology + syntax = shape of the text.

We can also remember that our habit of the written word properly delimited by spaces or punctuation makes us forget that the syntax string is a continuum that we cut while understanding: we use simultaneously morphological, syntactical, semantic and pragmatical information with which we make deductions, inferences, deadlocks and use intuition (about the word see Tesnière page 27, §11 to 15).

It is possible to classify the pattern matching methods in two categories: the statistic and the structural methods (see Miclet and Fu). More precisely, we use here a string pattern matching method: every word of the sentence is replaced by its category, coded by a character, and question marks for unknown words: the cerebral electricity: "shape of the text." - > d?? (d=determiner) <the cerebral electricity> <mental illnesses and cerebral lesions> - > ??c?? (c=coordination) - > ?? (o=co-ordination)

Let us call this string the pattern by word that will be used in the grammar and in the parsing.

The information used in the parsing is composed of three types of data: a small lexicon of the words in finite number (about 80 forms): morphological deduction rules for each word, a set of patterns of the noun phrase for pattern matching, and of course the text to analyse itself.

The first step of this study is the noun or prepositional phrase. The following steps are the recognition of these phrases in the sentence, and the whole parsing of the sentence.

This work is implemented on Apple Macintosh, and the programming language is Pascal UCSD which is suitable to develop such a parser whose algorithms are rarely recursive.

B. The small lexicon:

It contains about 80 forms (not lemmas): determiners (articles, possessive, demonstrative and indefinite adjectives), prepositions, coordinations, some punctuation signs (considered as words). These words are the anchoring points for pattern matching.

269
But we realize that it is impossible to keep the general position to have only the words in finite number in this lexicon, and that the problem becomes pragmatic: what is the minimum of data necessary to recognize correctly the adjectives, some very current adjectives often placed before the noun (petit, autre, même), adverbs not derived from adjectives (bien, mal, très).

Every form has its possible categories, eventually gender and number; the list of the possible categories can be open: a form can have another category in a particular sentence: *le bien et le mal*, *le fa de ma clarinette*.

For example, *et* can be:
- conjunction which coordinates adjectives (b);
- value localitoisches and pronominal -> Tfi7T <localizing and pronominal value>
- conjunction which coordinates noun phrases (c):
  - maladies mentales et lésions cérébrales -> ?TcP? (lexical creation and renewing)
- conjunction which coordinates prepositional phrases (C):
  - influence de l'incidence et de la capacité -> c?P?Cdpf?

We have distinguished two categories of proposition according to the "attraction" between the two np: high attraction: *le système d'unités* <the unit system> (sort of composition), or low attraction: *un chat sur un toit* <at a cat on a roof> (facultative "circumstant"), whence two kinds of prepositional phrases: *internal* to the np: à de en (o) or *external* to the np: à de en dans chez vers (p).

So can be:
- internal preposition (o) in:
  - une théorie de la morphogenèse -> d?o?? <a morphogenesis theory>
  - le système d'unité international -> d?o?? <the international unit system>
- external preposition (p) in:
  - de l'animal à l'homme -> po??d <from animal to human>
  - proposition (q) in:
    - les différents moyens de faire les mesures -> d?q??d <the different ways to make measures>

C. The morphological approach:

Our attitude is to explore all possibilities to extract, to deduce information from the mere morphology of the word, without dictionaries, information which can be used at any time of the syntactical analysis.

For example, let us observe the words ending with -ité: -ité -ité -ité -ité -ité -ité; we have a regular alternation adjective/noun: électricité - électricité, combativité - combativité, portabilité - portabilité, particularité - particularité; from these endings, we can deduce that the word means a quality (semantic aspect) and is a singular feminine noun (category).

On the semantic opposite, endings as -ification, -isation suggest an example, for classification comes from the noun class(e)-suffix -ification, -isation: it comes from the semantical adjective + suffix -ification, -isation, -a rification, -isation: these words have been derived on the same way, with the same semantic aspect: the suffix -es -er (verbal ending) = iser or iser - iser or - iser has the property to make a verb or a noun which expresses an action, from adjectives (national) or nouns (cünstat); words ending with -ification or -isation are always feminine nouns.

In some cases, at first sight, the morphology does not give reliable information: a word ending by -ment can be an adverb (derived from adjective) or a masculine noun: for example élancement (clakly) is adverb and relâchement (câkly) masculine noun, but a more precise study brings the following information: -ément = adverb except 3 roots: agrémént complément increment and except the word élément; -ément = adverb: attenement + septem + -ement + -ment + -ment = noun: moment sarment fragment + -ment + -ment + -ment = noun derived from a verb: vigesrement positionnement.

At last, as far as geological production uses these elements and these rules to create new words, paleologists are analysing exactly as the other words (see Guilbert and Koiourek).

These morphological properties of each word are the second kind of anchoring points for pattern matching.

D. The grammar:

1. The grammar of the complex noun or prepositional phrase:

The phase is considered as a three level hierarchical structure (finite number of levels); the grammar is not recursive (on that point joining Tesnière and leaving Chomsky): phrase = complex noun or prepositional phrase which is composed of simple noun phrases which are composed of words or "agglutinated" words.

A complex noun phrase (cpn) is:
- either a simple noun phrase alone (G-cpn)
- or a train of simple noun phrases separated by: an external proposition (p=de, dans, pour), or a conjunction co-ordinating prepositional phrase (C-cpn, ou), or a conjunction co-ordinating prepositional phrase (C-cpn, ou), followed by a proposition (P=c, et c, ou avac), or a proposition preceding an infinitive (t=de, à, pour) or a present participle (f=ant).

These snp have between them relations of subordination or co-ordination.

2. The grammar of the simple noun phrase:

A snp is a train of words obeying two types of constraints:
- syntactical, when the phrase agrees with a dependency tree.
- morphological, when the word is usually named gender-number agreement.

A pattern of a snp is a horizontal projection of a sub-stemma of a canonical stemma. Let us remember that a stemma (word introduced by Tesnière) is a dependency tree. The canonical stemma represents an abstract of all possible patterns of a snp.

A sub-stemma is:
- either the unchanged canonical stemma,
- or the canonical stemma without a leaf,
- or a sub-stemma without a leaf.

A stemma is a two dimensional diagram: the vertical dimension of the hierarchical levels and the horizontal dimension of the written words; a stemma can be horizontally projected to obtain the one dimensional train of words as they are written. Here is an example of a possible canonical stemma of the snp, and its projection:

```
| noun | det | adv | adj | noun | adv | adj | noun | adv | adj | noun | adv | adj |
|------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|      |     |     |     |     |     |     |     |     |     |     |     |     |
```

There is a snp pattern for every sub-stemma of the canonical stemma. The three canonical stemmas now used are equivalent to about 2000 reworking rules.

The "agglutination" rules are applied inside the snp from right to left and are the bottom-up aspect of the algorithm:
- every adjective can be an agglutinated adjective (A) as the result of the co-ordination of several adjectives: *t=a A B t=A A ->> A + B adj
- in the forms: noun de noun or: noun à noun, we have [internal prepositional phrases working like adjectives (B), which are included in the snp, and processed as adjectives in the parsing of the snp: d?y->B d?f->B = d?y-noun.

We recognize here Tesnière's "transformation" concept: the "transformation" of the noun into adjective (see Tesnière pages 443 and seq.). The prepositional phrases which can be considered as adjectives are not only preceded by: de à en, but potentially by every preposition, for example in:
- numérotation
- par domaine ou lexicographique
- seraux
- ou selon la théorie
- rechercher
- neurologique dans les graphes applique à la reconnaissance
- We can remark here that the two co-ordinated objects have fundamentally the same function that is often but not always marked by the same category.
- We shall now deduce linguistic information from the form of the snp. For example, if we analyse the form: *un* [unknown word] (d?y), we can deduce that this unknown word is a masscle singular noun for two reasons: it matches the pattern: determiner - noun, and the whole snp retains its gender and number from the determiner. Here is an ambiguous case: [unknown word 1] [unknown word 2] (d?y); we have here three solutions: either noun - adjective, or adjective - noun, or noun - apposed noun. It is often possible to decide by a morphological study of the corresponding cpn:

```
| letricité cérébrale | d?y | et | letricité feminine noun |
|---------------------|-----|----|------------------------|
| nérole nouvelle conception | d?y | and | feminine feminine noun |
| les extra alpha | d?y | and | no number agreement |
```

In the texts now processed (content tables of scientific books), we have noun-adjective in about 97% cases, probably because French is a cantrifugal
language: the governor first, then the dependants (see Wagner et Pinchon page 155, Teinture pages 33 and 147). For example, la linguistique informatique and l'informatique linguistique which are morphologically ambiguous, are both understood by a native speaker as a form: noun - adjective.

If we choose to obtain only one analysis to get one deduction, we must have an order of trial, barring syntactical or morphological impossibility: this order is now: noun - adjective, then adjective - noun, then noun - apposed noun, with three stemmas tried in this order.

3. some parsing difficulties:

Wrong deductions upon the category come in the cases we have several possible analyses and when morphology does not imply the category:

- what does et coordinate?
  - valeur [localisatrice et pronostique] noun [adjectival] and [adjectival] 6
  - valeur localisatrice and pronostique

Two possible analyses according to whether et co-ordinates two snp or two adjectives; then pronostique can be analysed as noun or adjective;

- if the pattern is ?? without any possible morphological deduction, the form noun - adjective will be chosen, and that may be wrong in some rare cases.

But at the end of the parsing of the text, the lexicum is extracted and it is possible to consult it to repeat the ambiguous phrases.

IV. The analysis algorithm:

A. How syntax and morphology work together:

In such a parser, the parsed language implies a parsing strategy: in French, syntax gives more information than morphology; for example, in English morphology is poor and syntax becomes more important; in German, morphology is richer because of declensions and the three genders. So, in French, the parsing is guided by syntax and sometimes lightened by morphology:

at the beginnings, we look if it is possible to deduce its category, gender and number, and the deduction is marked sure or not sure, for example:

- 6 = feminine singular noun, sure (éléctricité)
- 7 = masculine singular noun (pronostique) or adjective (enthalpe)
- 8 = feminine plural (qualitatives)

- in the study of each enp, every category and some genders and numbers are known and the gender-number agreement is verified; for example:
  - 7 and adjective (deduced by syntax) --- masculine singular (prise)
  - 7 and adjective --- feminine plural (qualitatives)

If a snp does not agree in gender and number, the analysis fails and the next stemma is tested.

B. General case:

First, some replacements are made in the phrase submitted to the analysis, for example: space inserted after the apostrophes to isolate 7 or 8 as one word, about do -> about (one word), the do -> the, dos -> de los, seu -> a be, naux -> a les.

Then, for each word, the lexicum is consulted, and, if not found, the first morphological study is made (see above), whereas the possible categories of each word; this set is classed in the order of trial.

Then the set of all possible patterns by word is made from the combinations of the possible categories of each word, and from contextual constraints of each letter of the pattern; these constraints are as severe as possible to reduce the number of combinations as much as and as soon as possible: for example, for the phrase: évolution de l'électro-encéphalogramme d'un malade atteint de paralysie générale selon les effets du traitement, the number of possible patterns is reduced from 1250 to 85.

Then, each pattern is tested until the first successful analysis, except if there are possible adverbs, infinitives or present participles. In that case, a measure of the quality of the analysis is made to get the best analysis.

The test of one pattern is made in the following way: the pattern by snp is calculated: l'électro-encéphalogramme is a noun and the rest is an adjective.

We verify that this pattern by snp can constitute a snp (top-down aspect). The patterns by snp may be for example: G (snp) GoG (co-ordinated snp) GoG (sub-ordinated snp) GoG (two snp separated by a present participle).

We try to apply the grammatical rules (bottom-up aspect see above).

Then we study each snp; we test if it is possible to find a match with one of the three stemmas tested in the order: noun - adjective, then adjective - noun, then noun - apposed noun, with a deduced or confirmed category (noun or adjective) for every question mark; we test if we have a gender-number agreement between the governing noun and its eventual depending determinants and adjectives; this is done by a set intersection algorithm and by getting gender and number of the determinants from the lexicum, and by a morphological study (see above) of adjectives and nouns whose category has just been deduced.

At any moment, if a constraint is not satisfied, the test of this pattern is stopped and the next one is tested.

A bracketed phrase gives the history of the analysis.

C. A parsing example:

valeur localisatrice et pronostique

process by word:

- valeur
- ? localisatrice
- GoG et co-ordinates adjective (b), snp (c) or prep. phrases (C)
- ? pronostique

C is impossible because et is not followed by a preposition possible patterns: ?/ 7/ 8?

test of ??

calculation of the pattern by snp: ?? 7/ 8/ 9/ 10.

appositional role: GoG a (co-ordinated adjectives)

pronostique : ? — adjective

bracketed structure: valeur +A [localisatrice et + pronomstique]

new pattern: ?A and end of appositional study of the single snp:

syntactical constraint: ?A matches with the stemma 1 (noun-adjective)

- ? noun
- morphological constraints:
  - valeur :
    - singular (by morphological study)
  - localisatrice :
    - feminine singular (f - feminine by morphological study)
  - pronostique :
    - singular (by morphological study)

gender-number agreement: feminine singular

this snp is correct

and of course the snp is correct

valeur = noun

forms can be adjective elsewhere

localisatrice = adjective elsewhere, qualifies valeur

pronostique = adjective elsewhere, qualifies valeur

bracketed structure: GoG [valeur + [localisatrice et + pronostique]]

if we ask for all possible analysis, we get also:

GoG [valeur + localisatrice] + et + GoG [pronostique]

V. Conclusion:

In the text now processed, tables of contents and diagrams in scientific books and articles (about 10 000 words), the recognition of categories is normal to 95%, and the lexicum of the text is correctly extracted, but the deduction of the hierarchy of the snp and of relations between snp cannot be realised only by using syntactical and morphological data because semantic and pragmatic information is lacking.

The original assumptions are verified:

- it is possible to deduce categories of words by string pattern matching, with no dictionary and with very few data, by simultaneous use of syntactical and morphological information.

the concept of category is really a functional concept.

VI. References:

Blower, Fénérol, Ritzke, Stegentritt. ASCOF - a modular multilingual system for French-German translation 1985. Computational Linguistics, special issue. Blochum (ed.)

K.S. Fu Synaptic pattern recognition and applications 1982. Prentice-Hall

Louis Giffard. La créativité locative 1975. L'actualité des verbes et raise en valeur morpho-sémantique de la dérivation 1984. Thesis of the doctorate of 3ème cycle en traitement automatique des langues informatique linguistique

Inalco Université de Paris III

Pascale Pagès. Analyse morphologique automatique du français, extraction des verbes et mise en valeur morpho-sémantique de la dérivation 1984. Thesis of the doctorate of 3ème cycle en traitement automatique des langues informatique linguistique

Institut Université de Paris III

Patrice Poggen. Analyse automatique du tchaké 1979. Université de Paris III

Lucien Teinture. Éléments de syntaxe structurale 1982. Klimskoog Paris

Jacques Verme. Systèmes de la syntaxe et de la morphologie dans l'analyse automatique du français avec un minimum de données 1968 TA Informations Paris

R.L. Wagner et J. Pinchon. Grammaire du français classique et moderne 1962. Hachette Paris

271