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 the feasibility of the method** rather than to present an operationnal system.

I General objectives:

We intend to demonstrate that it is possible to parse texts with very few data: only determinants, 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 ASCOF (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 theorical 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 indexation, 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 theorical 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 1985 page 137). A more general theorical 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 spoken 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:

l'électricité cérébrale --> d?? (d=determiner) <the cerebral electricity>
maladies mentales et lésions cérébrales --> ??c?? (c=coordination)
<mental illnesses and cerebral lesions>

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, demontrative and indefinite adjectives), prepositions, coordinations, some punctuation signs (considered as words). These words are the **anchoring points** for pattern matching.

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 other words ? We have added the first numeral adjectives, indefinite adjectives, some very current adjectives often placed before the noun (*petit, autre, même*), adverb not derived from adjective (*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* ***la*** *de ma clarinette* .

For example, ***et*** can be:

- conjunction which coordinates adjectives (**b**):

valeur localisatrice ***et*** *pronostique* --> ??**b**? <localizing and prognosal value>

- conjunction which coordinates noun phrases (**c**):

maladies mentales ***et*** *lésions cérébrales* --> ??**c**??
<mental illnesses and cerebral lesions>

- conjunction which coordinates nouns (**e**):

création ***et*** *renouvellement lexicaux* -->?**e**?? <lexical creation and renewing>

- conjunction which coordinates prepositional phrases (**C**):

l'influence de l'inductance ***et*** *de la capacité* --> d?pd?**C**pd?
<the influence of inductance and capacity>

We have distinguished two categories of preposition according to the "attraction" between the two np: high attraction: *le système d'unités* <the unit system> (sort of compound nouns), or low attraction: *un chat sur un toit* <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 sur dans chez vers* (**p**).

So ***de*** can be:

- internal preposition (**o**) in:

une théorie ***de*** *la morphogénèse* --> d?**o**d? <a morphogenesis theory>

le système ***d'*** *unités internationnal* -->d?**o**?? <the international unit system>

- external preposition (**p**) in:

de *l'animal à l'homme* --> **p**d?pd? <from animal to human>

- preposition (**q**) in:

les différents moyens ***de*** *faire les mesures* --> d??**q**id?
<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*** : *-icité -ivité -abilité -ibilité -ubilité -arité -alité* ; we have a regular alternation adjective/noun: *électrique / électricité, combatif / combativité, portable / portabilité, particulier / 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 **action**, for example: *classification* comes from the noun *class(e)* + suffixe *-ification* , *nationalisation* comes from the adjective *national* + suffixe *-isation* , *climatisation* <air-conditioning> comes from the noun *climat* + *-isation* ; these words have been derived on the same way, with the same semantic aspect: the suffixe *-is-* + *-er* (verbal ending) = *-iser* or *-is-* + *-ation* (noun ending) = *-isation* has the property to make a verb or a noun which expresses an action, from adjectives (*national*) or nouns (*climat*); 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 *-ement* can be an adverb (derived from adjective) or a masculine noun: for example *lâchement* <slackly> is adverb and *relâchement* <slackening> masculine noun, but a more precise study brings the following information: *-ément* ==> adverb except 3 roots: *agrément complément incrément* and except the word *élément* ; *-ûment* ==> adverb: *assidûment* ; *-ublement -iblement -ivement* ==> adverb derived from an adjective:*indissolublement visiblement hâtivement* ; *-oment -rment -gment* ==> noun: *moment sarment fragment* ; *-issement -ionnement* ==> noun derived from a verb: *vagissement positionnement* .

At last, as far as neological production uses these elements and these rules to create new words, neologisms are analysed exactly as the other words (see Guilbert and Kokourek).

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 (cnp) is:

- either a simple noun phrase alone (**G**=snp)

- or a train of simple noun phrases separated by: an expernal preposition (**p**=*de , dans , pour*), or a conjunction co-ordinating snp (**c**=*et , ou*), or a conjunction co-ordinating prepositional phrase (**C**=*et , ou*) and followed by a preposition (**Cp**=*et de , ou avec*), or a preposition preceding an infinitive (**q**=*de , à , pour*) or a present participle (**r**=*-ant*).

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

2. the grammar of the simple noun phrase:

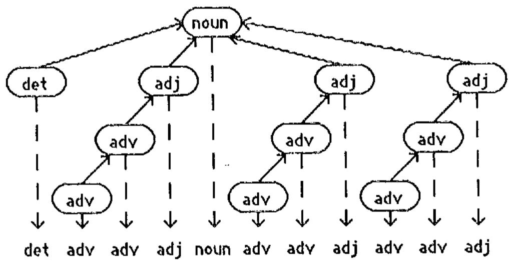
A snp is a train of words obeying two types of constraints:

- syntactical, when the phrase agrees with a dependency tree
- morphological, which 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 hierachical 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:



dependency relations: adj ----------> noun (depends on)
projection relations: det - - - - - > det (is projected as)

There is a snp pattern for every sub-stemma of the canonical stemma. The three canonical stemmas now used are **equivalent to about 2000 rewriting 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: ?b?->A Bb?->A ==> ?=adjectives

- 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: o?->B od?->B ==> ?=noun.

We recognize here Tesnière's "*translation* " concept: the "*translation* " 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*
sens usuel ou ***selon la théorie***
recherche heuristique ***dans les graphes*** *appliquée 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?), we can deduce that this unknown word is a masculine singular noun for two reasons: it matches the pattern: determiner - noun, and the whole snp inherits its gender and number from the determiner. Here is an ambiguous case: [unknown word 1] [unknown word 2] (??); 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 each word:

*l'électr**icité** cérébrale* --> d?? and ***icité*** ==> singular feminine noun
==> noun - adjective
*une nouvelle concep**tion*** --> d?? and ***tion*** ==> singular feminine noun
==> adjective - noun
les ondes alpha --> d?? and no number agreement
==> noun - apposed noun

In the texts now processed (content tables of scientific books), we have noun - adjective in about 97 % cases, probably because French is a centrifugal language: the governor first, then the dependants (see Wagner et Pinchon page 155, Tesnière 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 analysises and when morphology does not implies the category:

- what does ***et*** coordinate ?

valeur [(localisatrice) ***et*** *(pronostique)]* noun [(adjective) ***et*** (adjective)]
[(valeur localisatrice) ***et*** *(pronostique)]* [(snp) ***et*** (noun)]

two possible analysises 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 choosed, and that may be wrong in some rare cases.

But at the end of the parsing of the text, the lexicon is extracted and it is possible to consult it to reparse 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 lighted by morphology:

. at the beginning, we look if it is possible to deduce its category, gender and number, and the deduction is marked sure or not sure, for example:
-icité ==> feminin singular noun, sure (*électricité*)
-ement ==> masc. singular noun (*enregistrement*) or adverb (*purement*)
-ant ==> present participle, not sure (*concernant* or *passant*)

. in the study of each snp, every category and some genders and numbers are known and the gender-number agreement is verified, for example:
-al and adjective (deduced by syntax) ==> masculine singular (*principal*)
-ives 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 *l'* or *d'* as one word, *autour de* --> *autour-de* (one word), *du* --> *de le* , *des* --> *de les* , *au* --> *à le* , *aux* --> *à les* .

Then for each word, the lexicon is consulted, and if not found, the first morphological study is made (see above), whence the set of 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 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 8.

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

The **test of one pattern** is made in the following way: the **pattern by snp** is calculated: *l'electricité cérébrale* --> d?? --> G (G=snp); *maladies mentales et lésions du cerveau* --> ??c?od? --> GcG (**o**=preposition internal to snp); *l'activation par fermeture des yeux* --> d?p?od? --> GpG (the activation by closing eyes) (**p**=preposition external to snp).

We verify that this pattern by snp can constitute a **cnp** (top-down aspect). The patterns by snp may be for example: G (snp) GcG (co-ordinated snp) GpG (sub-ordinated snp) GrG (two snp separed by a present participle).

We try to apply the agglutination 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, whence 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 determinant and adjectives; this is done by a set intersection algorithm and by getting gender and number of the determinants from the lexicon, 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*
bcC *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: ??b? ??c?
test of ??b?
calculation of the pattern by snp: ??b? --> G possible complex phrase
agglutination: applicable rule: ?b? --> A (co-ordinated adjectives)
localisatrice : ?= adjective
pronostique : ?= adjective
bracketed structure: *valeur* +A(*localisatrice* +*et* +*pronostique*)
new pattern: ?A and end of agglutination
study of the single snp:
syntactical constraint: ?A matches with the stemma 1 (noun-adjective)
valeur : ?=noun
morphological constraints:
valeur : singular (by morphological study)
localisatrice : feminine singular (*-trice* by morphological study)
pronostique : singular (by morphological study)
gender-number agreement: feminine singular
this snp is correct
and of course the cnp is correct:
valeur >noun/f/s can be adjective elsewhere
localisatrice >adj/f/s can be noun elsewhere, qualifies *valeur*
pronostique >adj/f/s can be noun elsewhere, qualifies *valeur*
bracketed structure: G(*valeur* +A(*localisatrice* +*et* +*pronostique*))
if we ask for all possible analysises, we get also:
G(*valeur* +*localisatrice*)+*et*+ G(*pronostique*)

V Conclusion:

In the texts now processed, tables of contents and diagrams in scientific books and articles (about 10 000 words), the recognition of categories is correct to 99 %, and the lexicon 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 et 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:

Blewer, Féneyrol, Ritzke, Stegentritt *ASCOF - a modular multilevel system for French-German translation* 1985 Computational Linguistics, special issue Slocum (ed.)

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

Louis Guilbert *La créativité lexicale* 1975 Larousse Paris

Claude Hagège *La structure des langues* 1982 Que sais-je? PUF Paris
L'homme de paroles 1985 Fayard Paris

Rostislav Kokourek *La langue française de la technique et de la science* 1982 Brandstetter Verlag Wiesbaden

Laurent Miclet *Méthodes structurelles pour la reconnaissance des formes* 1984 Eyrolles Paris

Pascale Pagès *Analyse morphologique automatique du français, extraction des verbes et mise en valeur morpho-sémantique de la dérivation* 1984 thèse de doctorat de 3^{ième} cycle en traitement automatique des langues Inalco Université de Paris III

Patrice Pognan *Analyse automatique du tchèque* 1979 Université de Paris III

Lucien Tesnière *Eléments de syntaxe structurale* 1982 Klincksieck Paris

Jacques Vergne *Symbiose de la syntaxe et de la morphologie dans l'analyse automatique du français avec un minimum de données* 1986 TA Informations Paris

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

*