Syntactic properties of natural languages and application to automatic parsing

Jacques VERGNE

Jacques.Vergne@univ-caen.fr LAIAC

Université de Caen F-14032 Caen cedex **France**

Communication: abstract

Research domain

This work concerns syntax of natural languages, with a double aim:

- discover syntactic properties of natural languages,
- and base automatic parsing on these syntactic properties.

Fundamental axiom: a natural language cannot be theorised as a formal language, nor processed as a formal language or a programming language.

This work is on French in a fine way on both aspects, linguistic study and parsing; a precise linguistic study has been done on Spanish (in collaboration with Eduardo López Gonzalo and Luis A. Hernández Gómez of the E.T.S.I. Telecomunicación, Universidad Politécnica Madrid); and punctual studies of transposability of concepts have been realised on some other natural languages. Until proof of the contrary, these concepts seems to be general properties of natural languages.

Syntactic properties of natural languages

A definition of syntax is given; then 2 new segments are observed in the hierarchy, between word and sentence: sequences (nominal or verbal), and blocks; then the problematics "to segment / to link" is explored; then relations between contiguity and dependency are studied.

Application to automatic parsing

This part shows how to deduce segmentation strategies from properties of segments and how to build **parsers of linear complexity in time**.

A co-operation between theoretical and applied research

This work illustrates a profitable co-operation between theoretical and applied research; it began by a purely operating work (automatic parsing without a dictionary); then came a phase of theoritical explicitation; at last, the theoritical concepts diffused again into the automatic parsing while increasing its efficiency.

•

Topics

3. Automatic parsing and linguistic bases

Syntactic properties of natural languages and application to automatic parsing

Jacques VERGNE
Jacques.Vergne@univ-caen.fr LAIAC
Université de Caen F-14032 Caen cedex **France**

Introduction: research domain

This work concerns syntax of natural languages, with a double aim:

- discover syntactic properties of natural languages,
- and base automatic parsing on these syntactic properties.

The method is scientific and experimental: work on corpus[1], hand observations, computer aided observations and statistics, to find regularities and build a theory which models them.

The computer is an experimental device, an observation and modelling tool which renews scientific approach of linguistics, as the telescope renewed astronomy.

It is a work about forms, exploring the importance and the redundancy of natural language forms.

Fundamental axiom: a natural language cannot be theorised as a formal language, nor processed as a formal language or a programming language (double problem with which met some syntactic theories and automatic natural language processing systems) (see [Vergne 92] page 181).

This work is on French in a fine way on both aspects, linguistic study and parsing; a precise linguistic study has been done on Spanish[2] (see [López 93], chapter 4, pages 115 to 169); and punctual studies of transposability of concepts have been realised on English, German, Polish, Basque, Japanese (with speakers of each natural language) and Latin. Until proof of the contrary, these concepts seem to be general properties of natural languages.

I. Syntactic properties of natural languages

I.1 A definition of syntax of natural languages

Syntax of natural languages: to characterise it apart from syntax of formal languages

The study of syntax of natural languages consists in:

- discover segments of the written or spoken chain, and build a hierarchy of these segments (the "constituents"); it is given a particular attention to the topology of the linear chain (segments hierarchy, topological relations of contiguity or inclusion between segments)

[1] The French corpus is made of two informative (scientific) texts: the preface of a book about pattern recognition, and a paper in a review about marine biology (7000 words, 250 sentences).

[2] In collaboration with Eduardo López Gonzalo and Luis A. Hernández Gómez of the E.T.S.I. Telecomunicación, Universidad Politécnica Madrid, on a corpus of 72 sentences from an economy review (see [López 93], chapter 4, pages 115 to 169).

- formulate the problematics: to segment / to link, or: constituency / dependency
- discover characteristics and properties of linguistic relations between segments (dependency, co-ordination, coreference), at different levels of the hierarchy.

This hierarchy of segments is to be considered in total generality from word to text, passing by sentence and paragraph (see [Lucas 92] on paragraph structure, and [Lucas 93] on book structure).

The present work concerns this hierarchy between word and sentence, and is to be placed in a larger research between word and text.

I.2 Hierarchy of segments (of constituents) between word and sentence

Between words and sentences, we observe 2 levels of segments:

- nominal sequences and verbal sequences are made of contiguous words, and are in a *contiguity* relation between them
- blocks are made of 1 to 3 contiguous sequences, and are in relation of *contiguity or inclusion* between them.

So, a sentence can be seen at these two levels:

- as a chain of contiguous sequences
- as a chain of contiguous or included blocks

I.2.a Nominal sequences, verbal sequences and clips

the term of "sequence" is chosen to emphasise contiguities around noun or around verb, and also to stress the nominal-verbal symmetry (on both linguistic and algorithmic aspects):

nominal sequence:

a nominal sequence is made of a noun and its immediate satellites: partitive, determiner, adjectives and adjective adverbs

verbal sequence:

a verbal sequence is made of a verb (in all its forms: conjugated, infinitive, participle) and its immediate satellites: auxiliary, negation, clitics, verb adverb

in both types of sequences, satellites depends on the central element: a noun or a verb

clip:

at the sequence hierarchic level, a clip can be seen as a generic category for prepositions, subordination and co-ordination conjunctions, relative pronouns, punctuation (mainly commas, brackets and colon) and block adverbs; a clip is most often a single word, but may be a segment: comma - co-ordination conjunction - block adverb - preposition (, e incluso a **reducir**)

view of a sentence at the sequence hierarchic level:

a sentence is a partition of 3 types of contiguous segments: clips, nominal sequences and verbal sequences: a sentence may be completely coloured in 3 colours according to the segment type;

in the corpus of scientific texts in French, the proportion of nominal sequences vs. verbal sequences is 70% vs. 30%

this ratio, peculiar to the type of corpus, could participate to characterise a corpus example of sentences segmented in contiguous sequences and clips:

conventions: nominal sequences, **verbal sequences**, clips

Dans les réseaux trophiques marins , de nombreuses molécules énergétiques **sont transférées** entre les différents niveaux d' organisation **structurant** les échanges .

En estas épocas **se nota más** que el comprador **pide** aplazamientos .

Comprobaremos , a la hora de **hacer** la declaración , que **también han subido** las tarifas , y con efecto retroactivo .

ya que **se van** a **reducir** los márgenes , **hay** que **mejorar** los costes de producción por encima de todo , y **diseñar** una excelente estrategia de marketing y **deben ser** cambios definitivos .

I.2.b Blocks and clips

clip:

at the block hierarchic level, clips mark the beginnings of blocks; clips then segment the sentence in blocks, they "clip", attach every block in the sentence structure; they increase the ability to segment while listening or reading; they are in a small number, and allow to segment an unknown natural language with few means;

block structure:

a block is made of a clip (En) and a block body (estas épocas): [En estas épocas]

the central block (from the structural point of view) of a sentence is the only block without a clip:

[Verano]

[**se nota más**]

[empresarios **empezaron**]

[Lo fundamental **es mantener** tu estrategia]

the block body is the result of the combinatory of a nominal sequence and a predicate:

a predicate is a verbal sequence with eventually an object or attribute nominal sequence

=> the 4 structures of a block body are: (central block / other blocks)

nominal sequence (without a predicate)	[Verano]	10% / **70%**
predicate (without a subject nominal sequence):	[**se nota más**]	0.5% / 20%
subject nominal sequence + predicate	[empresarios **empezaron**]	**90%** / 10%
predicate + subject nominal sequence		0.5% /0.5%

(the statistics are given on the corpus of scientific texts in French) some statistics are peculiar to a corpus, some others are very stable (peculiar to a natural language): for example, frequencies of block bodies, or of clips, are stable: central blocks are mainly verbal, and other blocks are mainly nominal

but the number of blocks per sentence is very dependent on the corpus (8 blocks per sentence in average in the corpus of scientific texts in French); while some artificial corpuses used to develop some NLP systems (a classical but wrong method) have most often 1 to 3 blocks per sentence

clips typology, and their proportions:

blocks and sequences subordination clip	70%	[a la hora]
sequences co-ordination clip	10%	[y miembros]
ending-clip (participles without an auxiliary)	10%	[**discutida**]
bracket-clip	6%	
blocks co-ordination clip	4%	[y con efecto retroactivo]

view of a sentence at the block hierarchic level:

a sentence is made of contiguous blocks, in which some blocks (about 1/6) include other blocks; the central block generally is the first block, but it may be preceded by some anteposed blocks;

example of sentences segmented in contiguous blocks, without included blocks:

conventions: nominal sequences, **verbal sequences**, clips, [blocks]

[Dans les réseaux trophiques marins] [, de nombreuses molécules énergétiques **sont transférées**] [entre les différents niveaux] [d' organisation] [**structurant** les échanges] .

[En estas épocas] [**se nota más**] [que el comprador **pide** aplazamientos] .

[**Comprobaremos**] [, a la hora] [de **hacer** la declaración] [, que **también han subido** las tarifas] [, y con efecto retroactivo] .

[ya que **se van**] [a **reducir** los márgenes] [, **hay**] [que **mejorar** los costes] [de producción] [por encima] [de todo] [, y **diseñar** una excelente estrategia] [de marketing] [y **deben ser** cambios definitivos] .

blocks inclusion:

blocks inclusion is a major syntactic phenomenon, and its precise study clarifies syntactic structures

definition of inclusion:

the included block cuts the including block in 2 non-empty parts:

```
[ ——left part—— [ ———————————————— ] ——right part——
]
                  <--included block-->
```

<==================== including block
====================>

[empresarios [del sector público] **empezaron**]

this definition is only topological, and here does not concern dependency: dependency by contiguity (as in usual definition of a phrase) does not imply inclusion as defined here

inclusion points of a block in another block:

these inclusion points are in small number, because we observe that sequences are never cut

we can deduce these 3 inclusion points from the block structure:

[clip [3] subject nominal sequence [1] **verbal sequence** [2] object nominal sequence]

[1] between subject nominal sequence and predicate: 85% (the most frequent)
[2] between verb and object nominal sequence: 12%
[3] between clip and block body: 3%

example of sentences segmented in contiguous blocks, with some blocks included in depth 1:

conventions: [left part of the including block [included block] right part of the including block]

- blocks included between subject nominal sequence and predicate:

[La mesure *depth 0*
[de concentration] *depth 1*
[de chlorophylle] *depth 1*
est utilisée] *depth 0*
[pour **estimer** la biomasse phytoplanctonique]. *depth 0*

[Lo fundamental *depth 0*
[em momentos] *depth 1*
[de crisis] *depth 1*
es mantener tu estrategia] *depth 0*
[a largo plazo] . *depth 0*

[Las recetas *depth 0*
[que **piensan aplicar** los empresarios] *depth 1*
son] *depth 0*
[de lo más variado] *depth 0*

[empresarios *depth 0*
[del sector público] *depth 1*
[y miembros] *depth 1*
[del Gobierno] *depth 1*
empezaron] *depth 0*
[a **lanzar** mensajes] *depth 0*

[a favor] *depth 0*
[de la venta] *depth 0*
[a la vez] *depth 0*
[que sus activos más representativos *depth 0*
    [, el INI] *depth 1*
    [, Argentaria] *depth 1*
    [y Repsol] *depth 1*
, **se preparaban**] *depth 0*
[para un cambio] . *depth 0*

- blocks included between verb and object nominal sequence:

[Las empresas **debemos defender** *depth 0*
    [a ultranza] *depth 1*
la competitividad] ; *depth 0*

[Esto **ayudaría**] *depth 0*
[a **no subir** *depth 0*
    [, e incluso a **reducir**] *depth 1*
, los impuestos] . *depth 0*

inclusion depths of a block in another block:

depth 0 (non-included blocks) is the most frequent; then comes the inclusion in depth 1, nearly the only way to include a block: depth 2 is very rare, almost only for prepositional blocks; depth 3 has never been observed:

depth 0:	80%
depth 1:	20%
depth 2:	<1%
depth 3:	0%

hypothesis on understanding and inclusion depths:

when there is an inclusion, while listening or reading an included block, a memory effort has to be done to remember the first (or left) part of the including block, while waiting for its second (or right) part, to link its two parts, and then to recognise the including block;

if there is an included block (in depth 2) in the included block (in depth 1), we must successively stack in memory the first (or left) part of the including block, and the first (or left) part of the included block in depth 1, then wait for the second (or right) part of the included block in depth 1, to link its two parts, and at last wait for the second (or right) part of the including block, to link its two parts; it seems that this is an important memory effort, which seems to make impossible the inclusion in depth 3, or to make very rare the inclusion of a subordinated clause in depth 2.

example of a sentence segmented in blocks, with blocks included in depth 1 and 2:

[**Reducir** costes	*depth 0*
[**concentrando**	*depth 1*
[en una sola empresa]	*depth 2*
los servicios]	*depth 1*
[que **antes ofrecían** varias]	*depth 1*
es la táctica]	*depth 0*
[de Fernando Ocaña]	*depth 0*
[, de Tapsa] .	*depth 0*

included blocks typology:

the most frequent included blocks are prepositional blocks (65%), then co-ordinated nominal sequences (16%), then past participles without an auxiliary (10%), then subordinated clauses: relatives, then circumstantial;

we notice that the more a block is verbal, and the more it is difficult to include it in another block.

I.2.c Two ways to make a segment perceptible

mark the "colour" common to all elements of a segment:

there is a same mark (gender, number) on the words of a nominal sequence

there is a same mark (number) on the words of a verbal sequence with verb "be"

there is a same mark (person, gender, number) on the subject nominal sequence and the predicate in a block

mark the cut between two segments:

the cut between 2 blocks is marked by an anteposed clip.

I.3 Problematics: to segment / to link, or constituency / dependency

I.3.a Definition

Every relation is based on a hierarchy of segmentation, at the minimum a hierarchy of two levels:

a relation (<––•) links two segments a (<-a->) and b (<-b->), which are included in a same segment C (<== C ==>) of the immediately superior level ;

every relation (<––•) **between** a (<-a->) and b (<-b->) takes place **inside** C (<== C ==>):

```
<=========== C ===========>
      <–––––––––––•
   <--a-->      <--b-->
```

The question of relations then is based on the answers to questions about segmentation.

I.3.b Application to sentence segments

While climbing up in the hierarchy of segments:

a relation **between** 2 words takes place **inside** 1 sequence

a relation **between** 2 sequences takes place **inside** 1 block

a relation **between** 2 blocks takes place **inside** 1 sentence

I.3.c Dependencies and positions of segments in a segment of the level above

between 2 words, **inside** 1 sequence:

dependencies to the central element of the sequence are directly function of words positions in the sequence, of words categories and of gender-number agreements; it is a position - category - agreement redundancy:

```
una excelente •––> estrategia          también  han •––> subido
 •––––––––––––––––>                      •––––––––––––––>
```

between 2 sequences, **inside** 1 block:

dependencies are also directly function of sequences positions in the block and of sequences types (nominal or verbal); a dependency between subject and predicate is marked by the person-gender-number agreements; it is still a position - category - agreement redundancy:

[empresarios <–-• **empezaron**]   [Lo fundamental <–-• **es mantener** <–-• tu estrategia]

in these 2 first cases, we see that agreement is together a mark of relation between 2 segments, and a mark of cohesion of the immediately superior segment;

between 2 blocks, **inside** 1 sentence:

dependencies are not directly function of blocks positions in the sentence, and are not marked by an agreement (but a block often depends on the preceding block).

[empresarios ... **empezaron**] [a **lanzar** mensajes] [a favor] [de la venta] [a la vez] [que ...

I.4 Contiguity and dependency

I.4.a Determination dependency and actancial dependency

Tesnière puts on a same level two types of dependency ([see Tesnière 59] pages 102 and 144):

the dependency of an actant on a verb: let us call it actancial dependency

the dependency of an adjective on a noun: let us call it determination dependency

Let us focus on determination dependencies.

In these determination dependencies: an adjective on a noun, a past participle on a nominal sequence, and a conjugated predicate on its subject nominal sequence, an agreement marks the determination dependency (see [Vergne 89] pages 8 to 17):

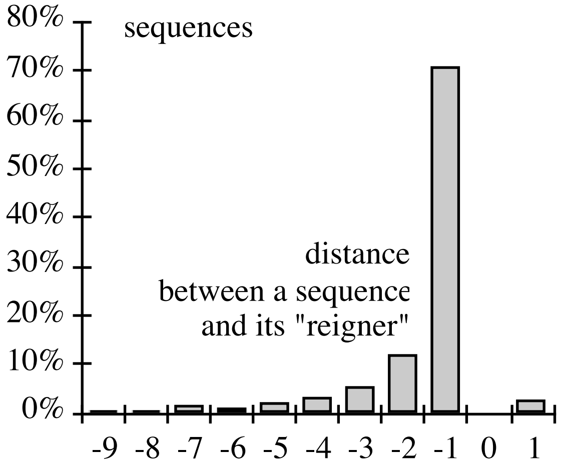
batallas <–-• públicas   [La privatización [de ...], [**discutida**]   [empresarios <–-• **empezaron**]
<-----------------•

I.4.b Determination dependency between sequences, and contiguity between sequences

If we observe a *same type of dependency* (the determination dependency) at a *same level of the hierarchy* of segments (sequences), we have a **homogeneity of sight**:

then we observe (on French) that the majority of sequences (70%) depends on the preceding sequence (its "reigner", Tesnière's "régissant");

in other terms, between 2 sequences, the most frequent mark of determination dependency is **contiguity**.



I.4.c A segment built on dependencies: the chain of contiguous dependent sequences

A sentence of a simplistic structure as "subject <–• verb <–• object" is made of a single chain of contiguous dependent sequences:

[La estrategia comercial **puede enfocarse**] [hacia la reducción] [de precios]

```
        <––––––––––––––•  <––––––––––––––––––––•  <–––––––––––•
```

when a sequence does **not** depend on the preceding sequence, a new chain begins (see just below)

the chain of contiguous dependent sequences is a new segment, and produces a new segmentation of the sentence

usually, a chain is made of more than one block (see just below)

but a chain also may be included in a block (see case 1 just below)

I.4.d Typology of cases where a sequence does not depend on the preceding sequence

It is the same thing than to make a typology of cuts between 2 chains of contiguous dependent sequences:

1) included block

=> the right part of the including block (most often a conjugated verbal sequence) depends on the left part (most often a subject nominal sequence), which precedes the included block

=> the cut (‖) is between the included block and the right part of the including block

[Lo fundamental [em momentos] [de crisis] ‖ **es mantener** tu estrategia] [a largo plazo].

```
    <––––––––––––•  <–––––––––––•
    <––––––––––––––––––––––––––––––‖–––––•
```

2) co-ordinated sequences or blocks

=> the second co-ordinated segment depends on the "reigner" (Tesnière's "régissant") of the first co-ordinated segment, which precedes this first co-ordinated segment

=> the cut (‖) is between the two co-ordinated sequences or blocks

... [al **afirmar**] [que el Gobierno **ha arruinado** su credibilidad] ‖ [y que **carece**] [de previsión] ...

```
    <–––––––––––––• <––––––––• <–––––––––––•
    <––––––––––––––––––––––––––––––––––––––––––‖–––––––•
```

3) two blocks (or chains) depend on the same block

=> the second block depends on the "reigner" of the first block, which precedes this first block

=> the cut (‖) is between these two blocks

[**Comprobaremos**][, a la hora] [de **hacer** la declaración] ‖ [, que **también han subido** las tarifas]

```
    <––––––––––––•  <–––––––• <––––––•
    <––––––––––––––––––––––––––––––––––––––‖–––––––––•
```

4) a block (or a chain) is anteposed to the central block

=> this block depends on the predicate of the central block, which follows its subject nominal sequence

=> the cut (‖) is between the anteposed block and the central block

```
[Entre batallas públicas] ║ [, Solchaga intento reconducirlo] [con medidas] [de eficacia discutida]
                              <------------●   <---------------● <------------●
          ●-------║-----------------> 
```

the cases 1) 2) and 3) correspond to the same dependencies diagram:

```
               <-----------●
               <----------------║-----●
          <======= chain 1 =======>  ,<==== chain 2 ====>
               <-seq.a->   <-seq.b->   <-seq.c->
```

the case 4) (the anteposed block) corresponds to the symmetrical of the preceding diagram:

```
                                        <-------------●
                    ●-----------║-------------------->
               <==== chain 1 ====>   ,<======= chain 2 =============>
               <-anteposed block->    <-subject-> <-predicate->
```

About the cut between 2 chains of contiguous dependent sequences, and comma in writing, pause in speech:

the more dependencies are long range dependencies (counted in number of sequences), the more a comma is frequent in writing (verified), and the more the pause is long in speech (see [López 93], chapter 4, pages 160 to 169).

II. Application to automatic parsing

II.1 A definition of automatic parsing

In parallel of the study of the syntax (see I.1 above), automatic parsing consists in:

- delimiting and identifying segments of the written or spoken chain, at different levels of the hierarchy: words, sequences, blocks and sentences
- linking segments: words inside sequences, sequences inside blocks, and blocks in sentences

II.2 History: parsing on the compiling model

Compiling, as an automatic translation of a formal language into another formal language, is the classical and habitual model of automatically parsing a natural language: it is a word by word process from left to right, in an unique pass, categorising and linking at once.

While doing this, we make the implicit or explicit hypothesis that natural languages are formal languages. This hypothesis has been useful in its time for the first steps in the field, but we know now it is false:

for example, here are some important different features:

the form/meaning mapping is many-to-many for natural languages, and one-to-one for formal languages;

the redundancy of forms is very high in natural languages, and equal to zero in formal languages;

segmentations of natural languages and formal languages are completely different.

So, it is now an obstacle to overcome (see introduction above, and [Vergne 92] page 181) while basing segmentation strategies on linguistic properties of segments of natural languages.

II.3 Linguistic properties of segments => segmentation strategies

Segmentation strategies have to be directly based on linguistic properties of observed segments:

Defining word categories according to sequence tripartition

word categories are based on the sentence tripartition in: words of nominal sequences, words of verbal sequences, and other words: clips, punctuation (processed as words) and block adverbs

for example, an adverb may be either an adjective adverb in a nominal sequence, or a verb adverb in a verbal sequence, or a block adverb outside sequences

Segments recognition while climbing in the segment hierarchy

the segments are recognised while climbing up in the segment hierarchy:

1) words recognition
2) sequences recognition
3) blocks recognition

it has to be noticed that only sequences structures (in terms of word categories) and blocks structures (in terms of clip and sequences) are expected, but not sentences structures (in terms of blocks), which are only observed

II.4 Towards parsers of linear complexity in time

Here is described the algorithm to build parsers which do not use any tree process and which are of linear complexity in time according to the number of words.

The text is processed sentence by sentence:

II.4.a Word recognition

Cutting the sentence in words

amalgams are separated:	*du* -> *de le*	*del* -> *de el*	(2 words)
elided words are separated:	*l'eau* -> *l'* *eau*		(2 words)
locutions are grouped:	*à partir de* -> *à-partir-de*		(1 word)

Building the possible categories list of every word without an exhaustive dictionary

the main lexical problem to solve is neology and scientific and technical sub-languages

in relation with neology, we can make a typology of 3 types of words, with a special solution for every type:

1) grammatical words, in which neology is almost nil, which make small and finite sets (except some prepositional locutions): clips, determiners, auxiliaries, pronouns, quantifiers
   solution: a little lexicon (400 forms in French), with any homography with other categories
2) verbs, in which neology is very low, which make an almost finite set

solution: a lexicon of verb roots (about 7000 verbs in French ≈ 50 Kb) with codes for homographic forms with nouns or adjectives, and some ending rules to handle neology

3) nouns, adjectives (and adverbs derived from adjectives), in which neology is very rich, which make an almost infinite set

solution (for alphabetic languages only): extract ending rules from a base of forms; an ending rule allows to deduce possible categories, genders, numbers from an ending (about 500 rules are enough); at this step, if a word does not match with any ending rule, it is a noun or an adjective with unknown gender and number

Sieve => shortening possible categories lists

by automatic and systematic generation of blocks patterns at word level (including every case of block inclusion), from sequence and block patterns, we can get all authorised contiguities between 2 categories, and then build a sieve which allows to shorten possible categories lists of every word

II.4.b Sequence recognition

the sequence recognition is a symmetrical process for nominal sequences and verbal sequences

the strategy is to make appear the tripartition in clips and sequences, while colouring the words in 3 colours, spreading the colours inside sequences, and trying to get one colour for every word:

1) colouring words in 3 colours: nominal sequences, verbal sequences, and clips

giving colours to every word, according to its possible categories

to every category, corresponds a place in the tripartition, therefore a colour: context free rules are applied; a given word gets 1 to 3 colours: *le* may be a determiner (nominal sequence colour: –), or a clitic (verbal sequence colour: /)

giving colours to every word, according to its neighbours (categories or colours)

context sensitive rules are applied (about 30 in the French parser): for example, in French, we find very often 2 contiguous words which can be either determiner-noun or clitic-verb: *la ferme* (the farm, close it), *le bois* (the wood, drink it); if the word before is a preposition, they are determiner-noun: *à la ferme* (at the farm), *dans le bois* (in the wood), and the colour is nominal; if the word before is a subject pronoun, they are clitic-verb: *il la ferme* (he closes it), *je le bois* (I drink it), and the colour is verbal; these rules are found by corpus observation, and/or on parsing errors

marking borders between colours

when colours of two contiguous words are single and different, borders are marked: {} for nominal sequences, <> for verbal sequences

propagating colours from sequence borders

starting from left marked borders of sequences ({ <), colours are propagated from left to right while possible, and the same from right to left

verifying sequence closures

for every open sequence ({ <), we verify that the sequence is closed (} >)

choosing the category of every word according to the type of its sequence for every word, its category is chosen in its possible categories list, according to the type of its sequence: for example, in a nominal sequence, only nominal sequence categories are chosen

2) studying every sequence: verifying sequence structure and observing internal agreements

nominal and verbal sequences structures are verified, and agreements are observed and computed:

in nominal sequences, gender-number agreement between noun, determiner and adjectives

in verbal sequences with conjugated verb "be", number agreement between "be" and attribute

inside sequences, dependencies on the central element (noun or verb) are computed

3) if unexpected structure, finding another segmentation of the sequence

the most frequent case is that two sequences of the same type have been delimited as a single sequence: the structure then is not recognised, and the two sequences are separated into two pieces, and their structures are verified again

another case occurs with locutions made of a verb and a noun without a determiner: *tenir compte* , *faire partie* ; these locutions have to be cut into a verbal sequence and a nominal sequence

4) updating the lexicon of the text

at the end of this step, the sequences are delimited, the sentence is tagged, and the lexicon of the text is updated

most frequently, a complete and right colouring of sentences is obtained

II.4.c Block recognition

1) compressing sequences into a single code

the words of a nominal sequence are coded with the single code N

verbal sequences are typed: conjugated: V infinitive: I present participle: R

2) cutting this pattern into blocks or block-parts

locating beginnings of blocks

as blocks begin with clips, every clip marks the beginning of a block: [p [P [,

(p = preposition P = subordination conjunction)

closing complete blocks according to block patterns

blocks which does not include another block can be recognised at this step, according to block patterns, because they are not cut in parts: [p -> [pN] [P -> [PNV]

relations in these blocks are computed at this moment

at the end of this step, the sentence is represented as a chain of blocks or block-parts:

[PNV] [**N**] [pN] [**V**] [pN] [**N**] [pN]

3) putting block-parts together

blocks which include other blocks are cut in 2 to 4 parts (see inclusion points above), and these parts have to be put together:

computing how every block-part can be continued

for a given block-part, and according to block patterns, it is computed how it can be continued:

N: Ø V VN means that a nominal sequence (N) can stay alone (Ø) or can be continued with a conjugated verbal sequence alone (V), or with a conjugated verbal sequence with its object (or attribute) nominal sequence (VN)

P: NV NVN means that a subordination conjunction (P) cannot stay alone (no Ø) and can be continued with subject and verb (NV), or with subject, verb and object (NVN)

pN: Ø means that a prepositional (preposition = p) block can only stay alone (Ø)

putting block-parts together

the chain of blocks or block-parts is processed from left to right:

every block-part which can be continued is stacked

for every block-part, if it can be the right continuation of a stacked left block-part (the stacked block-parts are tested backward: the most recent first, because of block inclusion), constraints on what can separate two block-parts are verified: subject-verb, verb-object, clip-sequence;

if the two block-parts can be put together (the right part is a possible continuation of the left part, and constraints are satisfied), they are linked, the left part is unstacked, and the depth of blocks included between the 2 parts is augmented of one level

if the two linked parts are subject and verb, number-person (eventually gender if verb "be") agreements between subject and verb are not used as constraints but only observed: it is possible because agreements are redundant with positions and patterns

II.5 Dependencies computability according to segments type

Inside a sequence, as dependencies to the central element of the sequence are directly function of words positions in the sequence and of words categories, these dependencies can be algorithmically computed after the sequence recognition.

In same way, inside a block, as dependencies are also directly function of sequences positions in the block and of sequences types, these dependencies can also be algorithmically computed after the block recognition.

But, between two blocks, as dependencies are not directly function of blocks positions in the sentence, these dependencies can only be heuristically computed, by valuation functions for example.

In this heuristical computation, it is possible to use a topological constraint on the chain, which selects candidates:

a dependency established between 2 sequences a and e delimits an upper segment S made of these 2 sequences and of the sequences between them:

```
<============== S ==============>
  <---------------------------●
<--a--> <-b-> <-c-> <-d-> <--e-->
```

these segments S delimited by dependencies can be only contiguous or included; this property is a topological constraint for every new dependency, according to already computed dependencies:

for example, if we look after the sequence on which depends a sequence f, placed after e, it may depend on e (making an upper segment contiguous to S), on a or on a sequence placed before a (making an upper segment which includes S), but f cannot depend on d, c or b:

```
    <============== S ==============>
       <-------------------------●
    <--a--> <-b-> <-c-> <-d-> <--e--> <-f->
<------<-------------------------<------●
```

(sequences a b c d build a chain of contiguous dependent sequences - see above I.4.c and I.4.d)

Dependencies inside a block or inside a sequence are about 3/4 of the total of dependencies; this reduces to 1/4 the heuristically computed dependencies (then with a lower precision, because of the lack of totally deterministic criteria to decide between several candidates).

Conclusion: a co-operation between theoretical and applied research

This work illustrates a profitable co-operation between theoretical and applied research; this co-operation has been more in alternation than simultaneous: it began by a purely operating work (automatic parsing without a dictionary), and syntactic properties of the studied natural language progressively emerged from familiarity with the corpus; then came a phase of theoritical explicitation, lighted by experiences, observations, automated measures on the corpus; at last, the theoritical concepts diffused again into the automatic parsing while increasing its efficiency.

References

[**López** 93] Eduardo **López Gonzalo**: *Estudio de técnicas de procedado lingüistico y acústico para sistemas de conversión texto-voz en español basados en concatenación de unidades* tesis doctoral, E.T.S.I. de Telecomunicación, Universidad Politécnica de Madrid, julio de 1993

[**Lucas** 92] Nadine **Lucas**: *Syntaxe du paragraphe dans les textes scientifiques en Japonais et en Français* Colloque international: Parcours linguistiques de discours spécialisés, Université Paris III, septembre 1992

[**Lucas** 93] Nadine **Lucas**, Nishina **Kikuko**, Akiba **Tomoyoshi**, K.G. **Suresh**: *Discourse analysis of scientific textbooks in Japanese: a tool for producing automatic summaries* Departement of Computer Science, Tokyo Institute of Technology, March 1993

[**Tesnière** 59] Lucien **Tesnière**: *Eléments de syntaxe structurale* Klincksieck (Paris) 1982

[**Vergne** 89] Jacques **Vergne**: *Analyse morpho–syntaxique automatique sans dictionnaire* thèse de doctorat de l'Université Paris 6, juin 1989

[**Vergne** 90] Jacques **Vergne**: *A parser without a dictionary as a tool for research into French syntax* communication-demonstration at CoLing 90 International Conference on Computational Linguistics vol. 1 pp. 70-72, Helsinki, Finland, August 1990

[**Vergne** 92] Jacques **Vergne**: *Syntax as clipping blocks: structures, algorithms and rules* communication and demonstration at SEPLN 92 congress (Sociedad Española para el Procesamiento del Lenguaje Natural), pp. 179-197 and 467, Granada, Spain, September 1992

-

A parser based on syntactic properties of natural languages:

towards parsing of linear complexity in time

Jacques VERGNE

Jacques.Vergne@univ-caen.fr LAIAC

Université de Caen F-14032 Caen cedex **France**

Demonstration

Research domain

This work concerns syntax of natural languages, with a double aim:

- discover syntactic properties of natural languages,
- and base automatic parsing on these syntactic properties.

Fundamental axiom: a natural language cannot be processed as a formal language or a programming language (as compiling).

This work is on French in a fine way on both aspects, linguistic study and parsing; a precise linguistic study has been done on Spanish (in collaboration with Eduardo López Gonzalo and Luis A. Hernández Gómez of the E.T.S.I. Telecomunicación, Universidad Politécnica Madrid).

Syntactic properties of natural languages

Two new segments are observed in corpus and defined in the segments hierarchy, between word and sentence: contiguous nominal or verbal sequences, and contiguous or included blocks.

Towards parsers of linear complexity in time

Segmentation strategies are directly based on linguistic properties of observed segments:

- word categories are based on the sentence tripartition in nominal and verbal sequences, and clips
- the segments are recognised while climbing up in the segment hierarchy:
    - words recognition
    - sequences recognition
    - blocks recognition
- only sequences structures (in terms of word categories) and blocks structures (in terms of clip and sequences) are expected, but not sentences structures (in terms of blocks), which are only observed

A co-operation between theoretical and applied research

This work illustrates a profitable co-operation between theoretical and applied research.

Hardware and software required

Mac II, system 7