Linguistic Issues in Grace (Evaluation of Part-of-Speech Tagging for French)

Josette LECOMTE*, Nadine LUCAS, Martin RAJMAN*****

* INaLF (Institut National de la Langue Française), Nancy France
* LIMSI,CNRS, (Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur,
Centre National de la Recherche Scientifique, Orsay, France
** EPFL (Ecole Polytechnique Fédérale de Lausanne), Lausanne, Switzerland

Abstract

GRACE is the first large-scale evaluation program of taggers for French. This experiment allowed to compare the assignments of Parts-of-Speech tags by various different taggers, on a common corpus of literary and journalistic texts. The evaluation relied on the acceptance by the participants of a reference formalism for morpho-syntactic description (the reference tagset) used by an expert to tag the evaluation corpus, and by the participants to provide a description (mapping table) of their own tagset. The global strategy was to make the reference tagging and tokenization of the finest grain possible. The reference tags were decomposed in Parts of Speech (main category) and lists of additional attributes, thus defining detailed syntactic patterns. The steps of the GRACE program are described, and the main adjudication issues are reported. The linguistic issues encountered during this experiment were linked to the difficulty to project relevant information about the sentence structure at the token level. Information derived from local analysis may be accepted as well as information derived from a larger context. Although the limitations of the experiment are acknowledged, the GRACE program proved to offer interesting opportunities to assess the state of the art in PoS tagging.

1. GRACE OVERVIEW

Organized by the French CNRS, the GRACE project –Grammars and Resources for Analysers of Corpora and their Evaluation– (Adda et al.(1995)) aims at applying the Evaluation Paradigm to the task of morpho-syntactic tagging of French texts. GRACE is the first large-scale evaluation campaign specifically devoted to Part of Speech (PoS) tagging for French.

The GRACE evaluation campaign (Paroubek 1997) was organized in four phases: training, dry-run (followed by the Avignon workshop in April 1997), test, and adjudication. 20 participants, from academia or industry registered for the evaluation, and, as the project covered a longer span of time than expected, 13 took part in the final tests.

In GRACE, the basis for evaluation are the PoS tags attributed by an expert, allowing a comparison with PoS tags produced by the various systems in competition. The internal functioning of the taggers is not taken into account (black-box evaluation). Two main problems are encountered :

- How is it possible to compare systems using various tagsets and based on various theories.
- How will it be possible, from this comparison, to derive a correct appreciation of the quality of a tagger.

In this contribution, we are concerned with the first question.

2. THE EVALUATION PROCEDURE

The evaluation is based on a comparison between PoS produced by the evaluated tagging systems and the PoS tags manually produced by a linguist. It implies that the evaluation is done on the basis of the ”performance” of the various systems, as compared to a reference system.

Some implicitly acknowledged rules of the game should be pointed out. The tags are attached to tokens, a notion different from the classical word. Different systems define different tokens, i.e. different physical units which can be smaller or larger than the word. The choice on the physical unit (segmentation) is closely related to the choice of logical units for linguistic analysis, in other words the tagset is always related to the tokens.

Furthermore, the GRACE committee never intended to impose on participants one unique tagset, as it can be nonconsistent with the linguistic theories and linguistic approaches used by the participants. The solution was to define a detailed and comprehensive tagset (the GRACE tagset (Rajman 1997)), working as a pivotal description. All participants were required to produce a ”mapping table”, explicitly associating their own tags to corresponding GRACE tags. The mapping tables make it possible to convert participant tags into GRACE tags and vice versa.

Such an approach raises several problems, not only with tagging, but also with segmentation, tokenization, and all phenomena related to the evaluation itself.

In the following sections, we first describe the reference linguistic resources (corpora, tagset, lexicon) used in GRACE, and then review the problems raised during the evaluation campaign, and especially during the (hand-)tagging of the reference corpora.

3. REFERENCE LINGUISTIC RESOURCES

3.1. Corpora

The corpora used in the GRACE experiment were exclusively made of French written texts. Since there is no established corpus for academic work in France, only two different genres were available, literary and journalistic. Two sources were used to build the corpora: the FRANTEXT corpus (INaLF), mainly containing literary texts from the XIXth century or the beginning of the XXth century, free of author or publisher copyrights; and contemporary extracts of the newspaper ”Le Monde” (a special agreement was signed for use of Le Monde texts).

The total amount of reference tagged text is about 140 000 occurrences (words). Both genres are roughly equally represented.

The texts were manually tagged as is usual, within the sentence frame. Both subsets of the corpus contain complex sentences, which may be considered as a bias in the experiment.

3.2. The tagset

As was said before, a reference was required concerning the tags. The GRACE committee decided to define a GRACE tagset, with brief guidelines for their use (Lecomte 1995; Lecomte 1997). The overall choice was to define a very detailed description, to fit the tokenization choice which aimed at the finest segmentation.

The GRACE tagset is directly inspired from the MULTEXT/EAGLES standard (Veronis et al. 1994; Veronis 1995), yet with some differences concerning some PoS or attribute tags. All modifications were raised from experience in tagging and from discussions with participants, under the supervision of a specific GRACE committee.

During the three phases of the GRACE action (training, dry-run and test), the current tagset slightly evolved : all participants were invited to suggest and discuss new modifications to the tagset. Some people were very active, others never reacted.

The current GRACE tagset contains 12 main categories (9 for Parts of Speech, 1 for punctuation, 1 generic class, and 1 evaluation blocker) :

PoS: A = adjectives
(+Attributes:Type, Degree, Gender, Number)
R = adverbs
(+Attributes:Type, Degree)
C = conjunctions
(+Attribute:Type)
D = determiners
(+Attributes:Type, Person, Gender, Number, Possessor, Quantification)
I = interjections
(no Attributes)
N = nouns
(+Attributes:Type, Gender, Number)
P = pronouns
(+Attributes:Type, Person, Gender, Number, Case, Possessor)
S = prepositions
(+Attribute:Type)
V = verbs
(+Attributes:Type, Mood, Tense, Person, Number, Gender)
F = punctuation
(no Attributes)
X = residual
(no Attributes)
? = extra-lexical forms
(no Attributes)

For each Part of Speech, there is a number (varying between 0 and 7) of additional fields : 4 for a noun, 7 for a verb. They all must be filled in with the corresponding value, either the sub-categorization (attribute value) or the ”irrelevant” value. If one attribute is not relevant for the word concerned, the value is ”-”, which means ”not relevant for this particular class or subclass of words”. For example, the attribute Case is allowed only for personal pronouns. For all other types of pronouns, the corresponding slot in the tag will be occupied by a ”-”.

The tagset comprises 312 different full tags, including the attributes.

3.3. The authorized patterns

Along with the list of tags, the organizers established for each PoS a list of authorized ”patterns” (i.e. feature combinations corresponding to linguistically sound tags). This was of great help for the building and validation of the mapping tables, as it allowed the creation of automatic verification tools. It gives another vision, another dimension.

Example : N [ckp] [mf] [sp]

which means that the tag for a noun will contain 4 compulsory fields:

1st : PoS
2nd : Type = c (common) or k (cardinal) or p (proper)
3rd : Gender = m (masculine) or f (feminine)
4th : Number = s (singular) or p (plural)

All these fields must be filled, with no possibility for indetermination or ambiguity or irrelevance for an attribute (no ”-” is allowed in the pattern). Hence, the tag for a noun is necessarily one of the following sound patterns :

Ncms Ncmp Ncfs Ncfp Nkms Nkmp Nkfs Nkfp Npms Npmp Npfs Npfp

From hypothesis, such problems as ambiguity, or indecidability were not supposed to exist when tagging in context. If a PoS does remain ambiguous in context, the only way to express it is by using a series of complete tags, thus allowing more than one correct answer. If the problem occurs for one attribute, one has to decide : either by choosing one value among others, or by using a series of complete tags to describe the various possibilities (below separated by —), or by using a shortcut for the attribute in question : a dot placed as a value for the attribute, meaning acceptance of ”all possibilities allowed by the pattern for this particular attribute”.

Examples:

philosophe	Ncms—Ncfs
philosophe	Nc.s
tarte	Ncfs—Afpfs

Along with the list of the GRACE/MULTEXT authorized tags and patterns, a document was established concerning the tagging guidelines. Some participants were interested in the way the reference texts would be tagged and these guidelines were extremely valuable to them. All participants received the first draft of a document which actually was not updated, as it appeared not to be a pre-requisite to hands-on work on corpora. However, the tagging guidelines were discussed throughout the adjudication phase after the dry-run. Needless to say it was really useful to the human expert who was in charge of preparing the reference tagged corpus, to check consistency.

3.4. The lexicon

A lexicon was first thought necessary, as a common tool for the participants, as well as for the human expert in charge of tagging reference text. The assumption was that tagging is done after the categories are checked in a reference lexicon. A version of MULTEXT lexicon was imported. It contained about 350 000 entries, which does not mean 350 000 different ”words”, as there were several entries for one word : one entry for one morpho-syntactic description.

Example :

token	lemma	description
mates	mat	Afpfp-
mates	mater	Vmip2s–
mates	mater	Vmsp2s–

The GRACE/MULTEXT lexicon was not distributed to participants. For at least two reasons :

- At the level of tokenization, this absence allowed the participants greater freedom in their strategy concerning segmentation, specially compound words. We quote one participant (J.Vergne, GREYC, 24 mars 1997):

  ”*Au moment des essais, personne n’a eu MULTEXT, il y a donc eu une liberté totale de choix sur les formes composées choisies par les participants. Et ceci est sain, dans le cadre de la liberté de choix des moyens pour réaliser une tâche précise et commune*”. ”At the time of the dry-run none of the participants had access to the MULTEXT [lexicon], therefore there was complete freedom of choice on compound forms, which could be selected as such by the participants. This is fair, in the perspective of freedom of choice concerning the means towards the accomplishment of a common well-defined task.”

- Because of the perpetual evolution of the tagset, even for minor changes, the update of the lexicon finally was always late... and thus not really available.

In fact, the participants worked with their own tools. Only one of them asked for the MULTEXT lexicon. The GRACE organizers nevertheless used this lexicon, mainly as a reference for the tokenization of the texts, in the reference tagged texts. It also helped in preparing the reference texts for the first two phases of the GRACE action (training and dry-run). But very soon the issue of the un-completeness of the lexicon arose. The lexicon provides grammatical categories, that is lists of potential tags. But these categories are derived mainly from morphological rules, and do not necessarily fit for tagging in context, where function prevails. This question was all the more acute because of the great number of ”transferred categories”, as will be discussed below. The lexicon was not adequate to the pursued goal, even for default tagging, because it was decided to tag in context, with the existing categories and patterns.

In order to help tagging, the lexicon would have been supposed to integrate all contextual behaviors. If all agree on the fact that too much information is equivalent to no information at all, what is the use of a lexicon? That is the reason why we did not use a lexicon in the third phase of the project..., and reference tagged texts were prepared with the help of ”patterns” only.

4. TAGGING THE REFERENCE CORPORA

4.1. Preliminaries

As already stated two corpora were used: the FRANTEXT corpus texts (INaLF) contains mainly literary texts from the beginning of the century. The second corpus is constituted of contemporary extracts of the newspaper ”Le Monde”. The total amount of reference tagged text is about 140 000 occurrences (words). 30 000 of them are lemmatized (phases 1 and 2). 110 000 are not lemmatized (phase 3).

The texts were supposed to be manually tagged, in context. In fact, automatic tools were used to segment and/or pre-tag texts which were then manually revised in context.

Only one person was in charge of the tagging task. No cross-tagging was possible, nor cross-revision for the dry-run, because of drastic cuts in the GRACE budget. Only a subset (about 10%) of the tagged corpus was submitted to cross-revision for the final test phase.

4.2. Drawing-up of the reference texts

First two phases (training and dry-run):

Texts were manually revised, in context, after having been automatically assigned the GRACE Grammatical Categories by means of a specific tool (called ”the segmentor”) designed by the GRACE committee. This tool put the text under the GRACE format, taking into account the Evaluation procedure requirements and the GRACE/MULTEXT Lexicon. Hence the text was presented with one token per line, lemmatized and pre-tagged (with GRACE tags). This text was then manually revised for disambiguation and/or rectification. For lack of time, no automatic tool was used to check the results.

No statistics were established concerning time spent in the various successive operations.

- As for the content of the reference texts, for the first phase of the GRACE action (training) : about 5 000 tagged occurrences sentences from the newspaper Le Monde
  + 200 coined sentences, representative of the main linguistic problems.

- For the 2nd phase (dry run) : about 25 000 tagged occurrences:
  five small files from the French corpus FRANTEXT = about 13 000 occurrences
  2 files from Le Monde = about 13 000 occurrences

- Third phase (It is the last, just completed ”test” phase.) : for this final phase, the reference tagged corpus amounts to about 110 000 occurrences :
  Le Monde (2 files) = 65 000 occurrences (32600 + 32700)
  Frantext (10 files) = 47 000 occurrences (4700 x 10)

The drawing-up of the reference texts was different. The automatic ”segmentor” used during the first two phases proved to be no longer adequate for segmenting and pre-tagging the texts. Lemmatization did not prove necessary, neither the GRACE/MULTEXT lexicon.

Hence the work started from scratch, with raw texts and following six steps :

1. raw texts
   FRANTEXT texts and Le Monde texts were used. Their size (number of occurrences given in words) was augmented, because the size of the corpus to be tagged by participants was augmented, and we wanted to keep a correct ratio between the two types of texts.

2. preparation of this text for automatic pre-tagging
   This pre-tagging was supposed to help ... We used for this operation the BRILL-tagger in use at INaLF. But for a correct pre-tagging, the text had to be ”cleaned” and normalized in a certain way. It cost a certain time.

3. pre-tagging (with tags different from GRACE tags)

4. segmenting of this pre-tagged text according to the GRACE format
   The text had to be put under the GRACE format.

5. mapping and manual disambiguation in context
   Tags had to be mapped into the GRACE tags, and verified in context.

6. last revision

These operations were conducted in various order, as texts arrived. For example :

1. the FRANTEXT texts (48 000 occurrences in 10 texts):
   A) Pre-tagging was done with BRILL-tagger
   B) Formatting of the text according to GRACE format. Semi-automatic work with Unix tools.
   C) Mapping of the tags, from BRILL to GRACE. Semi-automatic work, with Unix tools.
   D) Manual revision, word by word, in context. total amount of time for tasks B C D : about 200 man/hours

2. the first part of Le Monde (1 corpus of 50 articles, amounting to about 32 500 occurrences)
   A) manual preparation of the texts (taking off various unnecessary marks)
   B) Pre-tagging (BRILL)
   C) Formatting according to the GRACE format (semi-manually)
   D) Mapping (semi-manually)
   E) Manual revision, in context
   total amount of time : about 87 hours, decomposed into :
   for tasks A (manual preparation), B (automatic pre-tagging) : 20 hours
   for tasks C (formatting), D (Mapping and disambiguating) : 50 hours
   for task E (final revision) : about 17 hours

3. the second part of Le Monde (1 corpus of 50 articles, amounting to about 33 000 occurrences)
   The state of the text was different, as the human reviser got a text under the GRACE format, with BRILL tags. Tasks A, B, C were supposed to be done. Globally speaking, one can say that 18 or 20 hours were gained.

Task D was done in the following semi-manual way, with no specific tool:

- sometimes, automatic mapping, either from the category, or from a verbal ending, etc...

- sometimes a word by word approach, (for grammatical words,) Task D took 45-50 hours.

Task E, devoted to a last verification of the internal consistency and consistency with other texts, took about 16 hours.
total amount of time : 67 hours, decomposed into :
task C was automatically done.
for task D (mapping and disambiguating) : 50 hours
for task E (final revision) : 17 hours
Some conclusions :
total amount of words : about 112 000/120 000
total amount of time for preparing them :
machine time is unknown
human time to complete the work : 200 + 67 + 87 = 354 hours
(no other tool than common Unix tools)

Total human time to map, disambiguate, revise, can be amounted to about 360 hours, which means that processing speed is about 6 words per minute .

5. ADJUDICATION ISSUES

Discussions reported here took place during the two first phases of the GRACE action. As was said before, as the tagset evolved, the tagging guidelines evolved too. It can be interesting to have a look on the main questions which were raised, and why.

The ”why” question is easy to solve : it was mainly because participants made decisions on a particular linguistic problem different from those of the GRACE committee. Sometimes they went deeper in the examination of a problem , and wanted their efforts be visible in the evaluation; sometimes the GRACE decisions went deeper, and some participants feared to be penalized in the evaluation. When suggestions for modifications were made, they were examined, and the committee accepted or rejected the proposals. In all cases, a modification had an impact on the reference tagged text in preparation, and on the lexicon. This was not always easy to manage!

The second reason is that the tags were established also with a special goal in mind: reference tags should be easy to check and should provide a convenient basis for comparison. Hence a link was established between ”tags” and what was called the ”DMS” (= morpho lexical description of the tokens). This DMS was the support to be evaluated. The GRACE committee discussed tags and descriptions, having in mind the evaluation procedure.

The linguistic questions raised concerned Parts of Speech as well as attributes as well as particular words. These issues can be subdivided in two, those raised by participants themselves, and those raised by the committee.

5.1. Problems raised by participants

Participants had questions which directly concerned linguistic issues, tags and criteria for their use.

- The three Cases proposed in MULTEXT for personal pronouns appeared difficult to manage to some participants, who preferred the four traditional cases. The committee opted for 4.

- Numerals , particularly Cardinals, also raised questions: there were long and sterile discussions as for their status which is not clearly specified in traditional grammar Riegel94. The GRACE committee, following MULTEXT, proposed to merge all Cardinals into one specific category, relying on graphical discrimination. Some participants did not agree, because they made their own decisions on that particular point, and they wished that their results could be appraised. Hence, it was agreed that Cardinals must be tagged on a syntactical basis according to the four main Parts of Speech: Noun, Pronoun, Adjective, Determiner. The indication of Cardinality is given as an attribute.

- Contracted articles, adjectives, relative pronouns also raised many issues . Were they to be considered as one unit or as several units? After having proposed a minimalistic approach, the committee finally adopted a solution consisting in going as far as possible in the description.

  One participant suggested to discriminate between ”plural indefinite determiners” (=”de”, ”des”) and ”prepositional articles”. After having gone deeper into the problem, the committee decided to go farther, and to discriminate between ”prepositional” articles (linked to a question of valence), and ”partitive” articles. The French ”de” is not only a preposition, not only a plural indefinite article (for ”des”) but also a former partitive marker. In the reference manually tagged corpus there exists a discrimination between partitive articles, fused prepositional articles and indefinite article ... It was thought interesting to see if an automated tagger could emulate this analysis, and to what extent.

  It meant that the committee adopted compound tags describing :

preposition + definite article :	”des”	Sp+Da-mp-d
partitive marker + definite article :	”des”	Da—-i+Da-mp-d
as well as simple tags :		
indefinite article :	”des”	Da-mp-i

- Auxiliaries and Past Participles, were also discussed at length.The question rose when a sequence was met, containing past participles and adjectives in a relation of coordination. Are the past participles adjectives as well? How is to be considered the French verb ”être”, either as a main verb (because of the presence of adjectives), or as an auxiliary verb (because of the presence of past participles)? The good solution would have been to redefine what is to be considered as an auxiliary, what is a verbal participle, and what is an adjectival participle. J.Veronis suggested to re-think the problem entirely. This solution was not adopted, because of lack of time. We simply informed participants of what was decided for the reference tagged texts :

  ”avoir” and ”être” are auxiliary verbs if followed by a past participle
  a past participle is verbal when preceded by one of these two auxiliairies .
  A past participle is adjectival in all other cases.
  In fact, the linguistic problem was eluded.....

- Capitalized nouns, which are not really ”proper” nouns, are preceded by a definite article, but are not ”common” nouns. During the second phase of GRACE-1 (dry run), we tried to use a solution proposed by J.Veronis : a special Type ”d” for such nouns. At the Avignon workshop, this was categorically rejected by all the participants, who preferred the traditional division between ”proper” and ”common”, using the EAGLES criteria to discriminate.

- A special attribute for ”typing” was tried :
  It was meant for items written in numbers, or abbreviations, etc... This attribute was rejected by participants too.

- ”Que”, ”comme”, etc.. were also discussed, as special grammatical words. For these items, it is difficult to define a proper tag unless the different layers of the sentence are taken into consideration. The rules were minutely described in the guidelines.

5.2. Issues raised inside the committee

These issues were more oriented towards segmentation of the text and evaluation. Tags were discussed, not specially in their relation to ”linguistics”, but with regard to the evaluation procedure. For example :

- Punctuations were first considered as Residual. The first idea was to tag all ”separators”, including the ”blank”, in the same way. It was supposed to be a good idea as far as the evaluation procedure was concerned. But all punctuations are not ”separators” :

  hyphens and apostrophes are not ”separators” in the same way. Nor the blanks ... The questions about punctuations were closely related to the more crucial problem of ”segmentation” of the text (Marcus et al. 1993; Mathieu-Colas 1994): what is a token? what is to be considered as a token? What is a word? Are ”compounds” allowed? What is a ”compound”? etc... For instance, as the choice of tokenization was already made, and as the apostrophe was considered as a punctuation mark, in words such as ”l’heure” there are three tokens (here separated by a dot) : l.’.heure. Thus, the letter l was considered as a token and had to be tagged.

  The committee finally opted for a special category (”F”), for all punctuations, though not entering into the details of their function in context. This decision is still a subject of controversy.

- Irrelevance, Indecidability, Indetermination, Ambiguity, Factorization are real linguistic problems.

  Some patterns authorize an indication of ”irrelevance” for some attributes linked to the particular PoS concerned (with a hyphen as a value of this attribute). For example, the attribute ”person” is valid for possessive determiners, but is not relevant for articles or demonstrative determiners. The hyphen is used in the corresponding field of the tag. These problems were met during the manual tagging of the reference corpus, and discussed with participants as well as inside the committee, but not with the same preoccupations. Inside the committee, the question was : To what extent could the tagset be modified or the description be modified or the evaluation be modified if the committee decided to take the above mentioned linguistic problems into account?

  – Nothing existed for ”invariable” words, for which none of the proposed value for an attribute was relevant. Could the hyphen be used as a value of one attribute in that case? or another marker?

    Eventually, it was decided to assimilate ”no” value with ”all” values. It is not possible to use the hyphen in that case, but only a punctuation mark, or, better, a series of complete tags.

    Example :

    the French word ”velours” is invariable in number.
    velours Ncms—Ncmp or velours Ncm.

  – Nothing existed to mark factorization, linked to coordination phenomena. It is not really a question of irrelevance nor of indetermination, but rather a question of cumulative information. No pattern authorizes cumulative information for one attribute. None of the existing values represents cumulative information. Could it be possible to add new non-univocal values for existing attributes? The decision was the following : if a word factorizes an information, one can choose either one preferential tag, or a series of tags.

    For example, in a sentence such as ”ils sont déchirés et sales”, the word ”sont” is at the same time a ”main” verb and an ”auxiliary” verb (according to the GRACE criteria detailed above)
    sont Vmip3p-—Vaip3p-

  – Ambiguity is not supposed to exist when coding in context. Yet it exists... The solution to this problem is to give a series of complete tags in the description... As a result, there is not ”one” meaning for ”one” description containing a series of complete tags : the ”—” in a description means sometimes ”this” OR ”that” and sometimes ”this” AND ”that”.

- Transferred categories were also a big problem. It raises from the fact that linguists use category as a means to classify words, and function to characterize the role of sets of words in context. Tags being attached to tokens do not allow such a differentiation. We speak of ”transferred categories” (after Tesnire (Tesniere 1959)) when, for example, a noun behaves in context as an adjective (ex. ”un air *vache*”), when an adverb is used as a noun (”un *aujourd’hui* vaut mieux qu’un *demain*”), when an adjective or a past participle behaves as a noun (ex. ”un *aigri*”, ”un *frileux*”) etc... There was some uncertainty whether the participants would rely on dictionaries and categories based on morphology, or compute a syntactical function. The possibility to allow tagging from either morpho-lexical clues or contextual clues was considered necessary by the committee, although the participants did not raise this issue.

- ”Residual” versus ”Unknown” tags. Are they real ”PoS” and to be considered as such, or special categories not to be evaluated as others?

  There is no problem for the linguists : ”X” (Residual) is a PoS, though a waste basket. The ”?” (Extra-lexical) is the absence of a category, and reserved to unknown and undecidable items, which are not evaluated. All French words, or tokens assimilated to French words (for instance a word with a typing error, but with a French verbal ending, and recognizable in context), are tagged with the help of the existing categories, ”X” included. All foreign words or foreign acronyms or unknown symbols, or formulae are to be tagged ”?”.

  Since the ”?” tag blocks evaluation, it should be scarcely used by the expert. However, some previous decisions on tokenization, specially the choice to tag apostrophes as punctuation mark, did entail absence of consistent category for the preceding token (see example below). There remains a problem in the way the evaluation takes this inadequacy into account.

- The treatment of ”Compounds” was also a major point of discussion. This issue was raised at every moment ... The choice of the smallest token had to be balanced by the acceptance of compounds. As for segmentation : what is a token, are existing ”compounds” to be dissociated? Is it possible to create ”compounds”? As for their description : How can we imagine special tags for them, or special description? What about internal existing punctuation? What about the blanks between the components, will they receive a special tag? A compound may be ”known” as a lexical entry though each component is extra-lexical: what is to be done in that case? As for their evaluation : some participants will recognize strings of tokens as compounds, others not. What about the measures?

The final decision was to allow compounds, and to tag them as such. In the reference manually tagged corpus, when a compound is met, each of its components will receive a multiple description : first as a constituent of a compound annotated with numbers (see below 1.2, 2.2), and second, as an individual token.

Example:

alors que :

alors	:	first element of a subordinating conjunction OR (individually) adverb
que	:	second element of a subordinating conjunction OR (individually) subordinate conjunction

⇒ alors que :

alors	Cs/1.2—Rgp
que	Cs/2.2—Cs

and other examples :

ad hoc :	ad	Rgp/1.2—?
	hoc	Rgp/2.2—?
compte-rendu :	compte	Ncms/1.3—Ncms
	-	Ncms/2.3—F
	rendu	Ncms/3.3—Afpm
tout à l’heure :		
	à	Rgp/2.5—Sp
	l	Rgp/3.5—Da-fs-d/1.2—X
	’	Rgp/4.5—Da-fs-d/2.2—K
	heure	Rgp/5.5—Ncfs
tout au plus :	tout	Rgp/1.5—Rgp
	au	Rgp/2.3—Sp+Da-ms-d
	plus	Rgp/3.3—Rgp

6. CONCLUSION

As a conclusion, we can say that this experience was the most enriching one : discussions with participants proved that the decisions made as for the tagset were acceptable and not too far from the traditional approaches. Yet, one can regret that, because of lack of time, some aspects could not be rethought or amended. Theoretical rules found in grammars are not always adequate when working on corpora, in TAL.

At least two reasons can be evoked for this discrepancy: changing conventions in written usage and the type of corpora used in the GRACE action. As for changing usage, among other points, we can note the extensive use of numbers written in Arabic figures and of acronyms, in contemporary French.

At a deeper level, we can also highlight the fact that most grammars, such as (Le Goffic 1993), deal with ”minimal sentences”, or single-clause sentences, set as schematic models. On the contrary, the ”real text” corpora contain long and complex sentences, with multiple layers. Therefore, the choice of tokenization at the finest possible level and the consequent choice to tag individual tokens conflicted with the human feeling of ”syntactical value”. Discussions about compounds words and compound expressions, or cumulative information and factorization were very vivid. They indeed reflect the need to accept as valid various descriptions projected to the word level but based on a more or less extended notion of context, ranging from contiguous words through phrase up to sentence.

A very important work was needed to establish the guidelines for tagging. Although most participants accepted them as they were, it should not be forgotten that the number of tags is very important and that providing explicit rules for each and every case is by no way an easy task. Discussions with computer linguists and scientists helped to keep logic and consistency. They had a different ”vision” of problems, and in some cases, put into perspective special phenomena.

The tagset and criteria for tagging were fixed at a given moment for the experiment, but are in fact still under discussion (Vergne 1998).The whole, though not perfect, proved workable and useful.

As for the perspectives which remain to be developed for the future experiments, the establishment of a reference corpus containing more different genres is certainly important. For syntax, the most crucial point lays in the clear definition of linguistic units. The manual tagging relies on functional analysis, where syntactical function is beared by a set of words, either phrases (syntagmas), or clauses. Linguists tend to describe words as elements in a set, for instance a verb as a VP, in which case the set has only one element. An adequate frame should be established for complex sentences, where the functional set can be described as a combination of phrases and clauses.

The GRACE action is still a forum, and its results are not an end in itself, but a basis for new developments, advances and progress.

7. References

Adda G., Blache Ph., Mariani J., Paroubek P., and Rajman M.. Action GRACE - Mise en place du paradigme d’Évaluation - Application au domaine de l’analyse morpho-syntaxique. In *Proceedings of the Conférence sur le Traitement Automatique du Langage Naturel (TALN’95),* Marseille, France, juin, 1995.

J. Lecomte, “Recommandations pour l’étiquetage morpho-syntaxique manuel de textes”, internal note, GRACE, September 1995.

J. Lecomte, “Codage MULTEXT-GRACE pour l’action GRACE. Les étiquettes morpho-syntaxiques et les critères d’assignation”. Draft, internal technical note, GRACE, 1997.

P. Le Goffic, “Grammaire de la phrase française”, Hachette, Paris, 1993.

M. Marcus and B. Santorini and M. A. Marcinkiewicz, “Building a Large Annotated Corpus of English: The Penn Treebank”,*Computational LInguistics*, 19. 2 1993, 313–330.

M. Mathieu-Colas, “Les mots a traits d’union. Problèmes de lexicographie informatique”, Didier Erudition, Paris, 1994.

Patrick Paroubek, Gilles Adda, Joseph Mariani, Martin Rajman, “Les procédures de mesure automatique de l’action GRACE pour l’évaluation des assignateurs de Parties du Discours pour le français”, Actes des 1ères Journées Scientifiques et Techniques du Réseau Francophone de l’Ingénierie de la Langue de l’Aupelf-Uref, Avignon, April 1997.

M. Riegel and J.-C. Pellat and R. Rioul, “Grammaire méthodique du français”, PUF, Paris, 1994.

M. Rajman, “Format de description lexicale pour le français, Partie 2 : Description morpho-syntaxique”, internal technical report GTR-3-2.1, GRACE, June 1997.

L. Tesnière, “Eléments de syntaxe structurale”, Klincksieck, Paris, 1959.

J. Vergne, E. Guiguet, “Regards théoriques sur le Tagging”, In *Proceedings of the Conférence sur le Traitement Automatique du Langage Naturel (TALN’98)*, Paris, France, june 1998.

J. Véronis et al., “Common Specifications and Notation for Lexicon encoding”, MULTEXT report LRE-62-050, WO1.6 Deliverable, preliminary version, 1994.

J. Véronis and L. Khouri, “MULTEXT Lexical Specifications : application to French”, Document MULTEXT-LEX1, Draft. Version O.3, last modified 22-02-1996.