PASSAGE Syntactic Representation: a Minimal Common Ground for Evaluation

A. Vilnat[1], P. Paroubek[1], E. de la Clergerie[2], G. Francopoulo[3], M.L. Guénot[2]

(1) LIMSI: CNRS
Rue John von Neumann
F-91403 Orsay

(2) Alpage: INRIA Paris-Rocquencourt
Domaine de Voluceau
F-78153 Le Chesnay

(3) Tagmatica
126 rue de Picpus
F-75012 Paris

Abstract

The current PASSAGE syntactic representation is the result of 9 years of constant evolution with the aim of providing a common ground for evaluating parsers of French whatever their type and supporting theory. In this paper we present the latest developments concerning the formalism and show first through a review of basic linguistic phenomena that it is a plausible minimal common ground for representing French syntax in the context of generic black box quantitative objective evaluation. For the phenomena reviewed, which include: the notion of syntactic head, apposition, control and coordination, we explain how PASSAGE representation relates to other syntactic representation schemes for French and English, slightly extending the annotation to address English when needed. Second, we describe the XML format chosen for PASSAGE and show that it is compliant with the latest propositions in terms of linguistic annotation standard. We conclude discussing the influence that corpus-based evaluation has on the characteristics of syntactic representation when willing to assess the performance of any kind of parser.

1. Introduction

The work presented in the paper takes place in the context of PASSAGE[1] (Vilnat et al., 2008; de la Clergerie et al., 2008b), a 3-year French action with the following main tasks: (1) automatically annotating a French corpus of about 100 million words using 10 parsers; (2) merging the resulting annotations using a combination algorithm in order to improve annotation quality; (3) manually building a reference annotated subcorpus (around 400,000 words); (4) performing knowledge acquisition experiments from combined annotations; (5) running two parsing evaluation campaigns on the model of the EASy French evaluation campaign (Paroubek et al., 2006). The first campaign was run during October 2007, with 10 parsers and the test phase of second campaign during November 2009. The representation used in PASSAGE[2] is based on the EASy representation whose first version was crafted in an experimental project PEAS (Gendner et al., 2003), with inspiration taken from the propositions of (Carroll et al., 2002).

After a brief presentation of the state of the art in syntactic annotation, we present the Passage annotation schemes. Some linguistic phenomena are detailed in the following sections, and their PASSAGE representation is compared with their counterparts in other annotation schemes. Then, we explain the standardization effort made for the output XML format we use. The last section gives some information on the efforts made by the participants to export their parsing results to this format.

2. State of the art in syntactic annotation

The last decade has seen, at the international level, the emergence of a very strong trend of researches on statistical methods in Natural Language Processing. In our opinion, one of its origins, in particular for English, is the availability of large annotated corpora, such as the Penn Treebank (1M words extracted from the Wall Street journal, with syntactic annotations; 2nd release in 1995[3], the British National Corpus (100M words covering various styles annotated with parts of speech[4]), or the Brown Corpus (1M words with morpho-syntactic annotations). Such annotated corpora were very valuable to extract stochastic grammars or to parametrize disambiguation algorithms. For instance (Miyao et al., 2004) report an experiment where an HPSG grammar is semi-automatically aquired from the Penn Treebank, by first annotating the treebank with partially specified derivation trees using heuristic rules, then by extracting lexical entries with the application of inverse grammar rules. (Cahill et al., 2004) managed to extract LFG subcategorisation frames and paths linking long distance dependencies reentrancies from f-structures generated automatically for the Penn-II treebank trees and used them in an long distance dependency resolution algorithm to parse new text. They achieved around 80% f-score for f-structures parsing on the WSJ part of the Penn-II treebank, a score comparable to the ones of the state-of-the-art hand-crafted grammars. With similar results, (Hockenmaier and Steedman, 2007) translated the Penn Treebank into a corpus of Combinatory Categorial Grammar (CCG) derivations augmented with local and long-range word to word dependencies and used it to train wide-coverage statistical parsers. The development of the Penn Treebank have led to many similar proposals of corpus annotations[5]. However, the development of such treebanks is very costly from an human point of view and represents a long standing effort, in particular for getting of rid of the annotation

[1] ANR-06-MDCA-013: *Produire des annotations syntaxiques à grande échelle* (Large Scale Production of Syntactic Annotations), 2007–2009.

[2] `http://www.limsi.fr/Recherche/CORVAL/easy/PEAS_reference_annotations_v2.1.html`

[3] `http://www.cis.upenn.edu/~treebank/`

[4] `http://www.natcorp.ox.ac.uk/`

[5] `http://www.ims.uni-stuttgart.de/projekte/TIGER/related/links.shtml`

errors or inconsistencies, unavoidable with any kind of human annotation.

Despite the growing number of annotated corpora, the volume of data that can be manually annotated remains limited thus restricting the experiments that can be tried on automatic grammar acquisition. Furthermore, designing an annotated corpus involves choices that may block future experiments from acquiring new kinds of linguistic knowledge because they necessitate annotation incompatible or difficult to produce from the existing ones.

With PASSAGE (de la Clergerie et al., 2008b), we believe that a new option becomes possible. Funded by the French ANR program on Data Warehouses and Knowledge, PASSAGE is a 3-year project (2007–2009), coordinated by INRIA project-team Alpage. It builds up on the results of the EASy French parsing evaluation campaign, funded by the French Technolangue program, which has shown that French parsing systems are now available in good number, ranging from shallow to deep parsing. Some of these systems were neither based on statistics, nor extracted from a treebank. While needing to be improved in robustness, coverage, and accuracy, these systems have nevertheless proved their capacity to parse medium amount of data (1M words). Preliminary experiments made by some of the participants with deep parsers (Sagot and Boullier, 2006) indicate that processing more than 10 M words is not a problem, especially by relying on clusters of machines. These figures can even be increased for shallow parsers. In other words, there now exists several French parsing systems that could parse (and re-parse if needed) large corpora between 10 to 100 M words. Passage aims at pursuing and extending the line of research initiated by the EASy campaign by using jointly 10 of the parsing systems that have participated to EASy. They will be used to parse and re-parse a French corpus of more than 100 M words along the following feedback loop between parsing and resource creation as follows (de la Clergerie et al., 2008a):

1. Parsing creates syntactic annotations;
2. Syntactic annotations create or enrich linguistic resources such as lexicons, grammars or annotated corpora;
3. Linguistic resources created or enriched on the basis of the syntactic annotations are then integrated into the existing parsers;
4. The enriched parsers are used to create richer (e.g., syntactico-semantic) annotations;
5. etc. going back to step 1

In order to improve the set of parameters of the parse combination algorithm (inspired from the Recognizer Output Voting Error Reduction, i.e. ROVER, experiments), two parsing evaluation campaigns were planned in PASSAGE, the first already took place at the end of 2007, and the second at the end of 2009 (de la Clergerie et al., 2008b).

In the following, we present the annotation format specification and the syntactic annotation specifications of PASSAGE, and we explain how PASSAGE representation relates to other syntactic representation schemes for French (even adapted from Italian) and English, proposing extension to address English when needed. Second, we describe the XML format chosen for PASSAGE and show that it is compliant with the latest propositions in terms of linguistic annotation standard. We conclude discussing the influence that corpus-based evaluation has on the characteristics of syntactic representation when willing to assess the performance of any kind of parser.

Our comparisons will be made with the Stanford typed dependencies representation (SD) (de Marneffe and Manning, 2008), and, following these authors, with the GR (Carroll et al., 1999) and the PARC (King et al., 2003) schemes. We also draw some comparisons with TUT format (Bosco and Lombardo, 2006). TUT is a project for the development of a collection of morphologically, syntactically and semantically annotated Italian sentences; it includes: the definition of a native representation format (i.e. TUT format), which is dependency-oriented and aims at capturing the richness of the predicate-argument structure, and the conversion in Penn Treebank, in other constituency-based formats and in a format based on the Categorial Combinatory Grammar[6]. The authors of TUT took part in the EVALITA campaign series, with which PASSAGE is starting a collaboration. The first step has been to build a common corpus coomposed of 200 sentences with French and Italian aligned versions.

3. PASSAGE syntactic annotation

PASSAGE uses 6 kinds of syntactic groups. Each can be assimilated to a set of minimal non recursive constituency local constraints. They are : the noun phrase (**GN**), the prepositional phrase (**GP**), the verbal nucleus (**NV**), the adjective phrase (**GA**), the adverb phrase (**GR**) and the verb phrase introduced by a preposition (**PV**). Figure 1 to Figure 4 give annotation examples, where the group labels may be found in the bottom right corner of the boxes enclosing the groups.

The 14 PASSAGE syntactic relations establish all the links between the groups described above and/or word forms. We have the classical subject-verb (**SUJ-V**), auxiliary-verb (**AUX-V**), attribute-subject/object (**ATB-SO**). Then we distinguish among 3 kinds of dependencies between the verb and complements or modifiers: direct object-verb (**COD-V**), complement-verb (**CPL-V**) in case of adjuncts or indirect objects [7] and modifier-verb (**MOD-V**) for optional modifiers. The complementor (**COMP**) is used to link the introducer and the verb kernel of a subordinate clause or a preposition and a noun phrase when they are not contiguous. We also use different modifier relations to link to the noun (resp. adjective, adverb or preposition) all the chunks which modify it: **MOD-N**, **MOD-A**, **MOD-R** and **MOD-P**. The three last relations are: coordination (**COORD**), to relate the coordination and the coordinated elements, apposition (**APP**), to link the elements which are placed side by side, when they refer to the same object and juxtaposition (**JUXT**), to link chunks which are neither coordinated nor in an apposition relation. Figure 1 to Figure 4 give annota-

[6] `http://www.di.unito.it/~tutreeb/`

[7] As in SD, we do not try to differentiate adjunct from argument

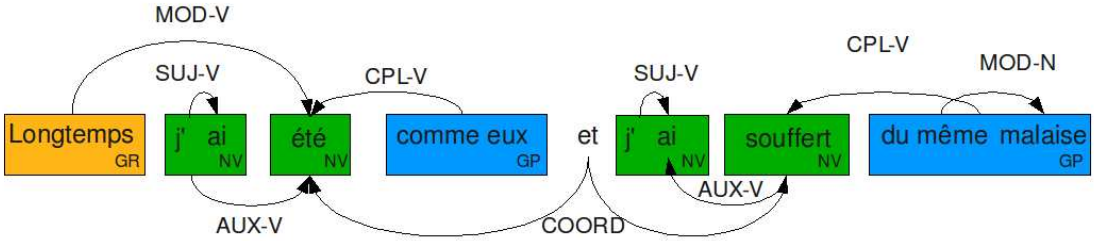

Figure 1: Annotation of a sentence extracted from the literary corpus. Tentative translation : *For a long time, I have lived as they do, and I suffered the same illness*

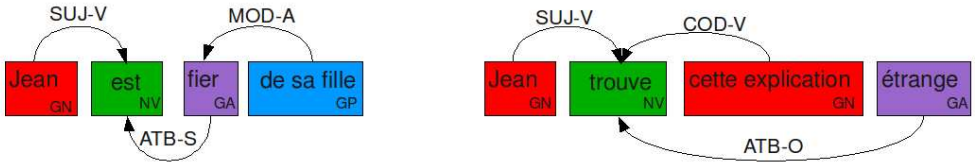

Figure 2: The attribute relation - *Jean is proud of his daughter* and *Jean finds this explanation strange*

tion examples, where these relations are represented by the arrows linking groups, or words.

4. Linguistic phenomena

PASSAGE was designed with the aim of addressing only the essential level of syntactic functions, leaving aside finer grain relations like specification and information more related to lexical issues like those addressed by SD with *element of compound number* or *abbreviation* or by PARC with verb tense and aspect, noun number and person and named entities' types. PASSAGE is not designed to address semantics, since the aim of its creation was intrinsic parser evaluation. As a consequence, the PASSAGE scheme allows to retrieve a syntactic dependency structure based on the relations in and between chunks, but as the semantic structure does not necessarily maps on it we do not want to annotate what is specific to the latter.

4.1. Syntactic head *vs.* Semantic head

For instance, the PASSAGE annotation of (1) will all be the same, indicating that the second chunk modifies the first, and then that resp. "*président*", "*guise*" and "*imbécile*" are the syntactic heads of the three constructions. But PASSAGE will not precise that in (1a) "*président*" is also the semantic head, whereas in (1b) the semantic head is not "*guise*" but "*récompense*", and in (1c) it is "*Pierre*" and not "*imbécile*".

(1) a. *le président des États-Unis* [8]
Groups : [le président]$_{GN1}$ [des États-Unis]$_{GP2}$
Relations : MOD-N(GP2,GN1)

b. *en guise de récompense*[9]
Groups : [en guise]$_{GP1}$ [de récompense]$_{GP2}$
Relations : MOD-N(GP2,GP1)

c. *cet imbécile de Pierre* [10]
Groups : [cet imbécile]$_{GN1}$ [de Pierre]$_{GP2}$
Relations : MOD-N(GP2,GN1)

The syntactic head is responsible for the syntactic constraints' satisfaction with the rest of the sentence (e.g. agreement). The semantic head is responsible for the semantic constraints' satisfaction, e.g. lexical selection restriction.

4.1.1. Valency *vs.* Transitivity

Syntax and semantics are also distinct when dealing with transitivity and valency. Valency is a lexical semantics property which defines the arguments which are expected by a given lexeme (often a *verb*, but not always); transitivity represents the effective syntactic relations between a verb (in this case) and its subject and components in a given sentence. In PASSAGE, we annotate transitivity (syntactic domain) but we do not annotate valency (semantic domain). Both are not always identical in sentences.
For example with *manger*, a bivalent verb :

(2) a. *Je mange de la soupe* [11]
Groups : [Je mange]$_{NV1}$ [de la soupe]$_{GN2}$
Relations : SUJ-V(Je, mange), COD-V(GN2, NV1)
Valency (argument structure) : *manger (je, soupe)*
→ Identical structures

b. *Il mange mais ne grossit pas*[12]
Groups : [Il mange]$_{NV1}$ mais [ne grossit]$_{NV2}$

[8] *the president of the United States*
[9] *by way of reward*
[10] *this fool Pierre*
[11] *I am eating soup*
[12] *He eats (a lot) but does not become fat*

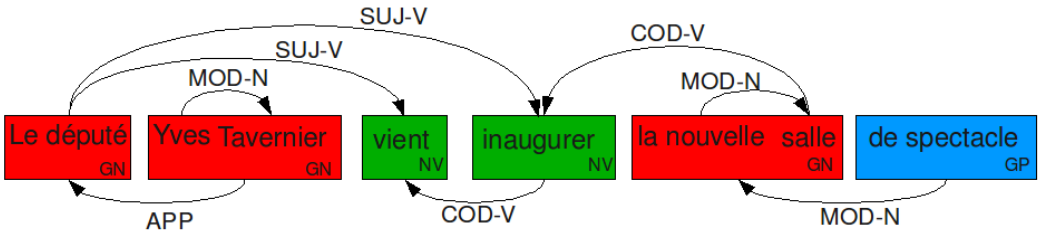

Figure 3: The apposition relation - *Member of Parliament Yves Tavernier comes to inaugurate the new theatre.*

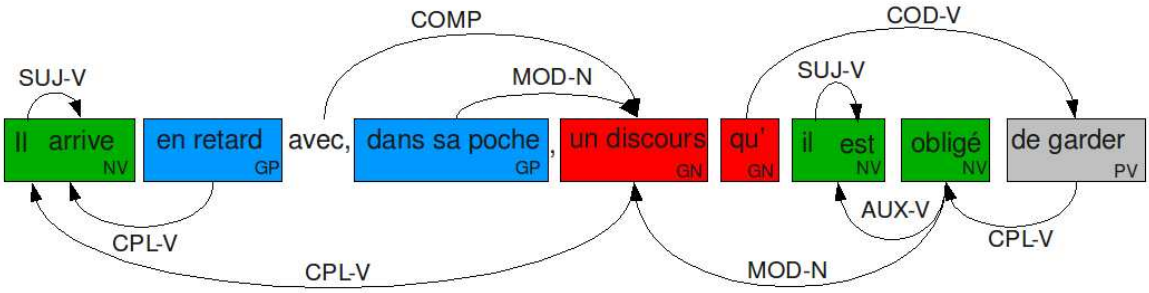

Figure 4: The complementor relation - *He arrives late with, in his pocket, a speech he has to keep.*

[pas]$_{GR3}$
Relations : SUJ-V(Il, mange), no COD-V
Valency (argument structure) : *manger (il, ∅)*
→ PASSAGE does not annotate the lack of a relation which is semantically expected but syntactically not realised (the second argument).

And with *souffler*, a monovalent verb :

(3) a. *Le vent souffle.*[13]
Groups : [Le vent]$_{GN1}$ [souffle]$_{NV2}$
Relations : SUJ-V(GN1, NV2)
Valency (argument structure) : *souffler (vent)*
→ Identical structures : the subject is the first semantic argument

b. *Il souffle un vent à décorner les bœufs.*[14]
Groups : [Il souffle]$_{NV1}$ [un vent]$_{GN2}$ [à décorner]$_{PV3}$[les bœufs]$_{GN4}$
Relations : SUJ-V(Il, souffle), COD-V(GN2, NV1),...
Valency (argument structure) : *souffler (un vent)*
→ the COD-V is the first argument.

4.2. Subject relation : Control, Passive and Compound Tenses

In case of infinitive or gerundive forms, the subject is generally absent of the proposition. But, as soon as this subject is present in the sentence, we annotate this relation. For example, in case of control verbs, the subject or the object of this verb is also the subject of the infinitive, and the relation is written. The example of Figure 3 illustrates this annotation : *Le député* is related to *vient* and *inaugurer* by a SUJ-V relation (see also example 4a. When the infinitive clause is an adjunct, the subject may also be identified, and is annotated, when it is present, as in the example 4b. On the opposite, when the infinitive form appears in a locution, or when the subject is clearly absent from the sentence, the subject is not annotated (see example 4c. For these annotations, we go further from what is done for example in the Tut format (Bosco and Lombardo, 2006) (as well in the native one and in the Tut-Penn translation). In these annotation formats, the fact that there is an absent subject in the clause is annotated, without trying to identify it in the embedding sentence.

(4) a. *Pierre propose à Paul de venir.*[15]
Groups : [Pierre]$_{GN1}$ [propose]$_{NV2}$ [à Paul]$_{GP3}$ [de venir]$_{PV4}$
Relations : SUJ-V(GN1, NV2), SUJ-V(GP3, PV4)

b. *Avant de partir, Marie éteint la lumière.*[16]
Groups : [Avant de partir]$_{PV1}$ [Marie]$_{GN2}$ [éteint]$_{NV3}$ [la lumière]$_{GN4}$
Relations : SUJ-V(GN2, NV3), SUJ-V(GN2, PV1)

c. *Fumer tue.*[17]
Groups : [Fumer]$_{NV1}$ [tue]$_{NV2}$

[13] *The wind blows*
[14] Approximate translation *It is blowing a gale*
[15] *Pierre proposes Paul to come*
[16] *Before leaving, Marie swithches off the light*
[17] *Smoke kills*

Relations : SUJ-V(NV1, NV2) →The verb *fumer* has no subject in this case.

As PASSAGE leans towards syntax, we chose to annotate the subject relation between the groups which are concerned with the agreement constraints. It is the reason why the subject is annotated between the noun phrase and the auxiliary, and not the main verb in case of compound tenses. The same choice is made for the passive forms. Examples are given in the Figure 1 : the clitics *j'* is the subject of the auxiliary *ai*, with which the agreement constraints have to be respected. The fact that the main verb is *été* in the first clause or *souffert* in the second one, may be retrieved by following the AUX-V relation between the auxiliary and the main verbs. If we need to get this information, we have to follow the two links (SUJ-V + AUX-V), and not only the first one. Following the same principle, in case of passive forms, the subject is annotated between the surface subject and the deep subject (which is a semantic notion) is annotated as a CPL-V (related to the main verb), when it is present.

(5) a. *Pierre est applaudi.*[18]
Groups : [Pierre]$_{GN1}$ [est]$_{NV2}$ [applaudi]$_{NV3}$
Relations : SUJ-V(GN1, NV2), AUX-V(NV2, NV3)
→The verb *applaudi* has no deep subject.

b. *Le livre est applaudi par la critique.*[19]
Groups : [Le livre]$_{GN1}$ [est]$_{NV2}$ [applaudi]$_{NV3}$ [par la critique]$_{GP4}$
Relations : SUJ-V(GN1, NV2), AUX-V(NV2, NV3), CPL-V(GP4, NV3)
→The verb *applaudi* has a deep subject annotated as CPL-V.

4.3. Coordination

The fact that SD leans toward semantics and PASSAGE does toward syntax can also be seen in the example (6)[20] for which SD gives direct *amod* links between *products* and *electronic* (or *products* and *computer*,...), while this information can only be accessed through indirect links with PASSAGE and GR as shown in figure 5. Following the same principles, the fact that the noun *products* is the direct object of both verbs *makes* and *distributes* is directly annotated in SD, the fact that the verbs are coordinated is no more indicated. In GR and Passage, the relation is annotated on the coordinator *and* : to obtain the relation between *products* and *makes* two relations have to be composed.

(6) *Bell, based in Los Angeles, makes and distributes electronic, computer and building products.* (WSJ-R)

[18] *Pierre is applaused*

[19] *The book is applaused by critics*

[20] This example and the annotations are given in (de Marneffe and Manning, 2008) for SD and GR, and we annotate it following Passage principles

4.4. Apposition and adposition

Note that with PASSAGE, intra-chunk relations such as the MOD-N relation between *nouvelle* and *salle* in Figure (3, can only address single word forms and not chunks as is the general case.
This is because PASSAGE does not allow the nesting of chunks. In the case the of MOD-N relation, we preferred in PASSAGE to have a nominal constituent holding the anteposed adjective and an intra-chunk MOD-N relation instead of an adjectival and nominal chunk linked by a MOD-N relation because adjectives occurring before the noun are much less frequent in French than those occurring after and the corresponding syntactic structure is generally straightforward. In case of apposition, the PASSAGE annotation is very similar to the one annotated in SD. If we consider the 7 example, in PASSAGE, the annotation includes *agleeful Alex de Castro* in a *GN*, and *a car salesman* in another *GN*. *MOD-N* relates *gleeful* to *Castro* and *Alex* to *Castro*. In the other *GN*, *car* and *salesman* are related by a *MOD-N*. An *APPOS* relation links the two *GN*s. This annotation is rather similar to the one presented in (de Marneffe and Manning, 2008)

(7) *I feel like a little kid, says a gleeful Alex de Castro, a car salesman, who has stopped by a workout of the Suns to slip six Campaneris cards to the Great Man Himself to be autographed.* (WSJ-R)

5. Standard XML format

The aim is to allow an explicit representation of syntactic annotations for French, whether such annotations come from human annotators or parsers. The representation format is intended to be used both in the evaluation of different parsers, so the parses' representations should be easily comparable, and in the construction of a large scale annotation treebank which requires that all French constructions can be represented with enough details.
The format is based on three distinct specifications and requirements:

1. MAF (ISO 24611)[21] and SynAF (ISO 24615)[22] which are the ISO TC37 specifications for morpho-syntactic and syntactic annotation (Ide and Romary, 2002) (Declerck, 2006) (Francopoulo, 2008). Let us note that these specifications cannot be called "standards" because they are work in progress and these documents do not yet have the status Published Standard. Currently, their official status is only Draft for an International Standard (DIS).

2. The format used during the previous TECHNOLANGUE/EASY evaluation campaign in order to minimize porting effort for the existing tools and corpora.

3. The degree of legibility of the XML tagging.

From a technical point of view, the format is a compromise between "standoff" and "embedded" notation. The

[21] `http://lirics.loria.fr/doc_pub/maf.pdf`

[22] `http://lirics.loria.fr/doc_pub/N421_SynAF_CD_ISO_24615.pdf`

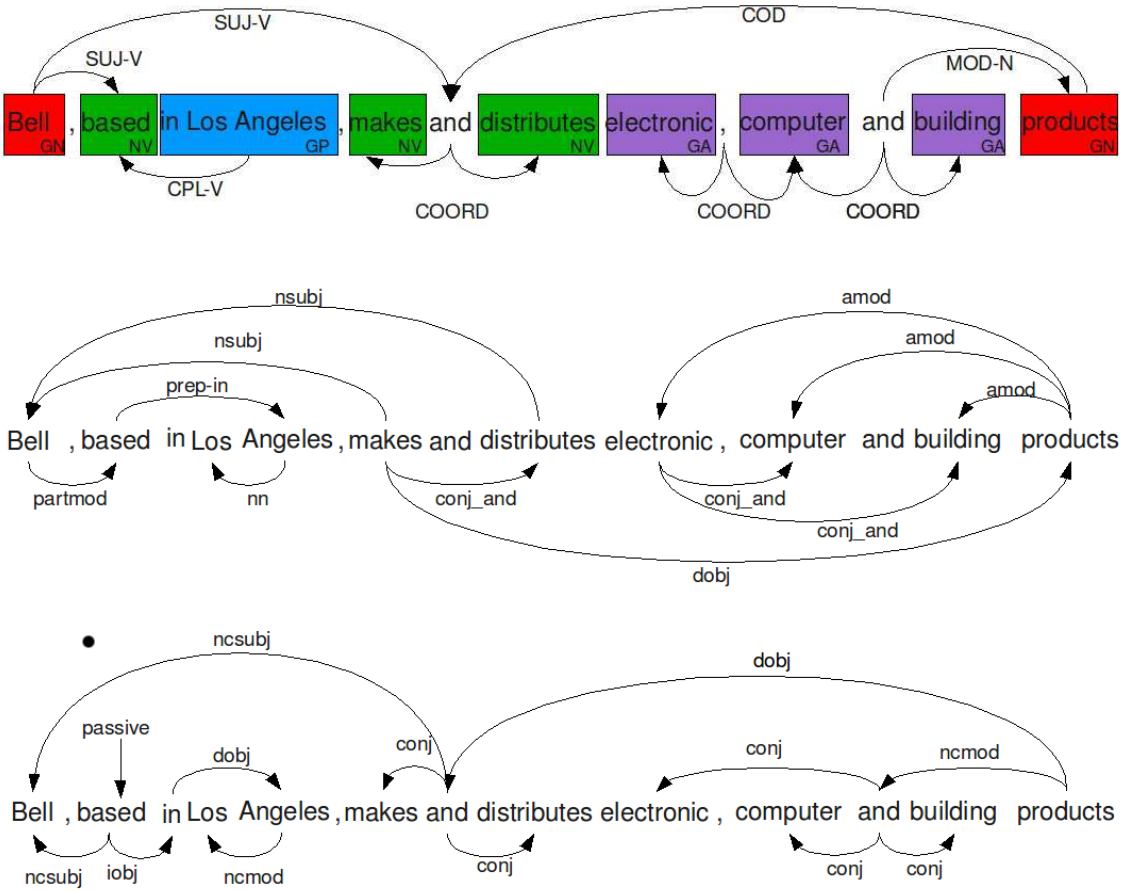

Figure 5: Annotations of the same sentence by respectively Passage, SD(de Marneffe and Manning, 2008) and GR(Carroll et al., 1999)

fine grain level of tokens and words is standoff (wrt. the primary document) but higher levels use embedded annotations. A standoff notation is usually considered more powerful but less readable and not needed when the annotations follow a (unambiguous) tree-like structure. Let us add that, at all levels, great care has been taken to ensure that the format is mappable onto MAF and SynAF, which are basically standoff notations.

The structure of a PASSAGE annotated document may be summarized with the UML diagram in figure 6.

The document begins by the declaration of all the morpho-syntactic tagsets (MSTAG) that will be used within the document. These declarations respect the ISO Standard Feature Structure Representation (ISO 24610-1). Then, tokens are declared. They are the smallest unit addressable by other annotations. A token is unsplittable and holds an identifier, a character range, and a content made of the original character string. A word form is an element referencing to one or several tokens. It has two mandatory attributes: an identifier and a list of tokens. Some optional attributes are allowed like a part of speech, a lemma, an inflected form (possibly after spelling correction or case normalization) and morpho-syntactic tags. The following XML fragment shows how the original fragment *"Les chaises"* can be represented with all the optional attributes offered by the PASSAGE annotation format:

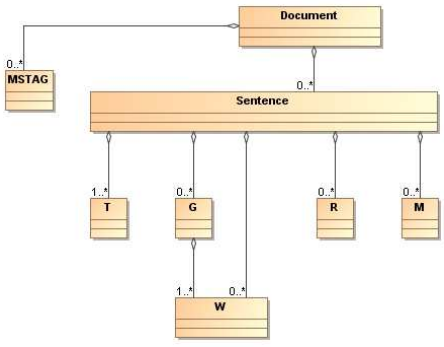

Figure 6: UML diagram of the structure of an annotated document

```
<T id="t0" start="0" end="3"> Les </T>
<W id="w0" tokens="t0"
   pos="definiteArticle"
   lemma="le"
   form="les"
   mstag="nP"/>
<T id="t1" start="4" end="11"> chaises </T>
```

```
<W id="w1" tokens="t1"
   pos="commonNoun"
   lemma="chaise"
   form="chaises"
   mstag="nP gF"/>
```

Note that all parts of speech are taken from the ISO registry[23] (Francopoulo et al., 2008). As in MAF, a word may refer to several tokens in order to represent multi-word units like *"pomme de terre"*. Conversely, a unique token may be refered by two different words in order to represent results of split based spelling correction like when *"unetable"* is smartly separated into the words *"une"* and *"table"*. The same configuration is required to represent correctly agglutination in fused prepositions like the token *"au"* that may be rewritten into the sequence of two words *"à"* + *"le"*. Contrary to MAF, cross-reference in token-word links for discontiguous spans is not allowed for the sake of simplicity. Let us add that one of our requirements is to have PASSAGE annotations mappable onto the MAF model and not to map all MAF annotations onto PASSAGE model. A **G** element denotes a syntactic group or a constituent (see details in section 3.). It may be recursive or non-recursive and has an identifier, a type, and a content made of word forms or groups, if recursive. All group type values are taken from the ISO registry. Here is an example:

```
<T id="t0" start="0" end="3"> Les </T>
<T id="t1" start="4" end="11"> chaises </T>
<G id="g0" type="GN">
   <W id="w0" tokens="t0"/>
   <W id="w1" tokens="t1"/>
</G>
```

A group may also hold optional attributes like syntactic tagsets of MSTAG type. The syntactic relations are represented with standoff annotations which refer to groups and word forms. A relation is defined by an identifier, a type, a source, and a target (see details in section 3.). All relation types, like "subject" or "direct object" are mappable onto the ISO registry. An unrestricted number of comments may be added to any element by means of the mark element (i.e. M). Finally, a "Sentence" element gathers tokens, word forms, groups, relations and marks and all sentences are included inside a "Document" element.

6. Parsing adaptation techniques for PASSAGE

When the PASSAGE project began, all the involved parsers were already running systems, and the annotation formalism was not designed to fit any particular system, in order to avoid any favoritism. Thus, the software developers had to take some decisions with respect to the way they produce PASSAGE conformant annotations. Among the participants, three different strategies were adopted.

Most participants decided to keep their internal structures and so, they had to write a post-processing mapping module to produce the expected PASSAGE format. This strategy was adopted, for instance by FRMG[24]. It should be noted that the original structures are usually richer, possibly keeping track of past hypothesis, etc. and finally that the task of writing such a mapping module is not a trivial one.

Other participants took the option that is to migrate their parser. In this case, instead of using an internal representation that is far away from the expected result, these structures are splitted into constituents and relations in a similar way to the PASSAGE format. Such systems can be considered as being "native" PASSAGE parsers. This strategy has been adopted by TagParser[25] (Francopoulo, 2008), for instance.

Finally, some participants decided to change neither their parsing strategies nor their output formats, but to modify the set of constraints that constitutes the grammar they use, in order to tend towards the PASSAGE expected results. This third technique implies that the parser's design allows to modify the parser's resources without changing the parser itself, and it can be the case mainly for symbolic approaches. For instance, the LPL's DeepParser and SeedParser proceeded this way (Balfourier et al., 2005).

7. Conclusion

We have described in details the PASSAGE annotation format that results from 9 years of evolution throughout a series of large scale generic black box quantitative objective evaluation campaigns for French parsing. Then using a set of linguistic common phenonema, we have looked at how it compares with other existing annotation schemes, particularly used in the context of evaluation. It appears that the PASSAGE format is not only a plausible common ground for representing French syntax in a theory-free way, but that it seems to be easily mappable to syntactic annotation schemes designed for other languages, with a minimum of modifications, at least to the extent of the proof elements that our initial tentative experiments with English and Italian provided. In the future, we will develop further our mapping experiments to other formalisms, starting with the small aligned corpus created when initiating the collabotration between PASSAGE and EVALITA.

8. References

J.-M. Balfourier, P. Blache, M.-L. Guénot, and T. Vanrullen. 2005. Comparaison de trois analyseurs symboliques dans une tâche d'annotation syntaxique. In *Actes de TALN 2005 - Workshop EASY*.

C. Bosco and V. Lombardo. 2006. Comparing linguistic information in treebank annotations. In *Proceedings of LREC 2006*. ELDA.

A. Cahill, M. Burke, R. O'Donovan, J. Van Genabith, and A. Way. 2004. Long-distance dependency resolution in automatically acquired wide-coverage pcfg-based lfg approximations. In *Proceedings of ACL 2004*.

J. Carroll, G. Minnen, and T. Briscoe. 1999. Proceedings of the eacl workshop "Linguistically Interpreted - Corpora (linc). In *Proceedings of EACL Workshop "Linguistically Interpreted - Corpora (LINC)*.

J. Carroll, D. Lin, D. Prescher, and H. Uszkoreit. 2002. Proceedings of the workshop "Beyond Parseval - Toward improved

[23] Data Category Registry, see `http://www.isocat.org`.

[24] `http://alpage.inria.fr/catalogue.en.html#frmg`

[25] `http://tagmatica.fr/produits/tagparser.htm`

evaluation measures for parsing systems". In *Proceedings of LREC 2002*.

E. de la Clergerie, C. Ayache, G. de Chalendar, G. Francopoulo, C. Gardent, and P. Paroubek. 2008a. Large scale production of syntactic annotations for French. In *Proceedings of the First Workshop on Automated Syntactic Annotations for Interoperable Language Resources at IGCL'08*.

E. de la Clergerie, O. Hamon, D. Mostefa, C. Ayache, P. Paroubek, and A. Vilnat. 2008b. PASSAGE: from french parser evaluation to large sized treebank. In *Proceedings of LREC 2008*.

M.-C. de Marneffe and C. D. Manning. 2008. The Stanford typed dependencies representation. In *Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation at COLING 2008*.

T. Declerck. 2006. SynAF: towards a standard for syntactic annotation. In *Proceedings of LREC 2006*.

G. Francopoulo, T. Declerck, V. Sornlertlamvanich, E. de la Clergerie, and M. Monachini. 2008. Data category registry: Morpho-syntactic and syntactic profiles. In *Proceedings of LREC 2008*.

G. Francopoulo. 2008. TagParser: Well on the way to ISO-TC37 conformance. In *Proceedings of the International Conference on Global Interoperability for Language Resources (ICGL)*.

V. Gendner, G. Illouz, M. Jardino, L. Monceaux, P. Paroubek, I. Robba, and A. Vilnat. 2003. PEAS the first instanciation of a comparative framework for evaluating parsers of French. In *Proceedings of EACL 2003*.

J. Hockenmaier and M. Steedman. 2007. Ccgbank: A corpus of ccg derivations and dependency structures extracted from the penn treebank. *Computational Linguistics*, 33(3).

N. Ide and L. Romary. 2002. Standards for language ressources. In *Proceedings of LREC 2002*.

T. King, R. Crouch, S. Riezler, M. Dalrymple, and R. Kaplan. 2003. The parc 700 dependency bank. In *Proceedings of the 4th International Workshop "*L*inguistically* I*nterpreted -* C*orpora (LINC)*.

Y. Miyao, T. Ninomiya, and J. Tsujii. 2004. Corpus-oriented grammar development for acquiring a head-driven phrase structure grammar from the penn treebank. In *Proceedings of the First International Joint Conference on Natural Language Processing (IJCNLP-04)*. Asia Federation of Natural Language Processing (AFNLP).

P. Paroubek, I. Robba, A. Vilnat, and C. Ayache. 2006. Data, annotations and measures in EASY - the evaluation campaign for parsers of French. In *Proceedings of LREC 2006*.

B. Sagot and P. Boullier. 2006. Efficient parsing of large corpora with a deep lfg parser. In *Proceedings of LREC 2006*. ELDA.

A. Vilnat, G. Francopoulo, O. Hamon, S. Loiseau, P. Paroubek, and E. Villemonte de la Clergerie. 2008. Large scale production of syntactic annotation to move forward. In *Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation (PE 2008) in conjuction with COLING*.