The PASSAGE Syntactic Representation

P. Paroubek[+] E. de la Clergerie[*] S. Loiseau[+] A. Vilnat[+] G. Francopoulo[$]
[+]LIMSI-CNRS [*]ALPAGE-INRIA-U. Paris 7 [$]TAGMATICA
E-mail: [+]{pap,sloiseau,anne.vilnat}@limsi.fr
[*]Eric.De_La_Clergerie@inria.fr
[$]gil.francopoulo@tagmatica.com

Abstract

We present the PASSAGE syntactic representation based on syntactic relations, initially developed for French in the scope of national evaluation campaigns. After a brief presentation of the non-nested chunks and syntactic relations of PASSAGE, we reuse the comparison elements that Marneffe and Manning have selected to compare the Standford typed dependencies (SD) against the GR and PARC representations, and show that PASSAGE is for a large part compatible compatible with these representation, standing closer to GR than to SD. After a presentation of the collaborative software support for PASSAGE representation, we conclude on some essential characteristics that pivot representation for syntax should exhibit.

1 Introduction

The work presented in the paper takes place in the context of PASSAGE [1][10][5] , a 3-year French action with the following main tasks:

- automatically annotating a French corpus of about 100 million words using 10 parsers;
- merging the resulting annotations using a combination algorithm in order to improve annotation quality ;
- manually building a reference annotated subcorpus (around 400,000 words),
- performing knowledge acquisition experiments from combined annotations,

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

- running two parsing evaluation campaigns on the model of the EASy French evaluation campaign [7]. The first campaign was run during October 2007, with 10 parsers. From the data collected on this occasion, we extracted parameters for the the combination algorithm. The second campaign, at the end of PASSAGE (2009), should provide information about the evolutions of the parsers during the project.

The representation used in PASSAGE [2] is based on the EASy representation whose first version was crafted in an experimental project PEAS [4], with inspiration taken from the propositions of [2]. The representation has been completed with the input of all the actors involved in the EASy evaluation campaign (both parsers' developers and corpus providers) and refined with the input of PASSAGE participants. This formalism aims at making it possible to compare all kinds of syntactic annotation (shallow or deep parsing, complete or partial analysis), without giving any advantage to any particular approach. It has six kinds of non-nested chunks [3] and 14 kinds of relations Some of them are illustrated in Figure 1. Like [1], the annotation formalism allows the annotation of minimal, continuous and non-nested chunks, as well as the encoding of relations wich represent syntactic functions. These relations (all of them being binary, except for the ternary coordination) have sources and targets which may be either word forms or chunks.and either extra-chunks or intra-chunk. The direction between source and target has been arbitrarily defined according to custom, this constitutes a minor point since the essential information lies in the label of the relation because we do not require the annotations to build trees but content with graphs. Note that the PASSAGE annotation formalism does not postulate any explicit *head*, see section 4.

2 Chunk annotation

For the PASSAGE campaigns, 6 kinds of chunks have been considered as illustrated below and in the table 2. These chunks are minimal and not embedded. The reason of this choice is to allow the evaluation of different kinds of parsers, as explained previously.

[2] http://www.limsi.fr/Recherche/CORVAL/PASSAGE/eval_1/2007_10_05 PEAS_reference_annotations_v11.12.html

[3] Defined in the Data Category Registry of ISO 12620 as a flat sequence of words typically containing more than one word, see http://syntax.inist.fr.

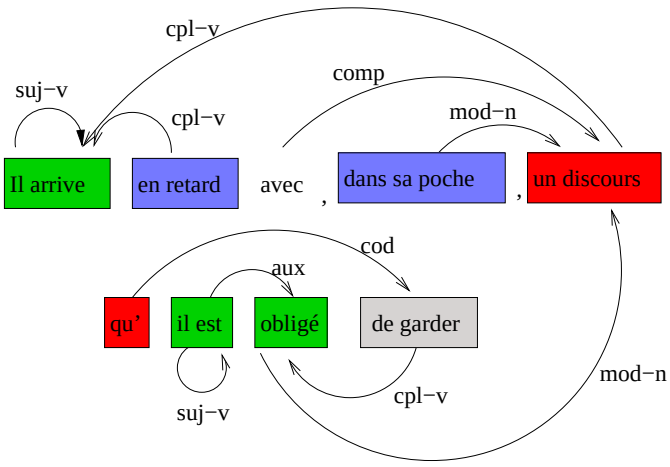

Figure 1: Example of a sentence annotated in chunks and relations. *He arrives late with, in his pocket, a discourse he is prevented to pronounce*

- the noun phrase (**GN** for *Groupe Nominal*): a noun preceded by a determiner and/or by an adjective with its own modifiers, a proper noun or a pronoun;
- the prepositional phrase (**GP**, for *groupe prépositionnel*): a preposition and the GN it introduces, a contracted determiner and preposition, followed by the introduced GN, a preposition followed by an adverb or a relative pronoun replacing a GP; in some constructions, the preposition is separated from the GN (as in the example of Figure 1), the corresponding annotation will be explained below;
- the verb kernel (**NV** for *noyau verbal*) includes a verb, the clitic pronouns[4] (always closed to the verb) and possible particles attached to it. Verb kernels may have different forms: conjugated tense, present or past participle, or infinitive. When the conjugation produces compound forms, distinct NVs are identified for each part of the compound;
- the adjective phrase (**GA** for *groupe adjectival*) contains an adjective when it is not placed before the noun, or past or present participles when they are used as adjectives;
- the adverb phrase (**GR** for *groupe adverbial*) contains an adverb;
- the verb phrase introduced by a preposition (**PV** *noyau verbal à préposition*): a verb kernel with a non-inflected verb introduced by a preposition. Some modifiers or adverbs may also be included in a PV.

[4]The French personal pronouns except disjunctive ones are all clitics.

GN	- [la très grande porte] (*the very big door*), [Rouletabille] - [eux] (*they*), [qui] (*who*)
GP	- [de la chambre] (*from the bedroom*), [du pavillon] (*from the lodge*) - [de là] (*from there*), [dont] (*whose*)
NV	- [j'entends] (*I hear*), [on ne l'entend] plus (*we hear her no more*) - [désobéissant] à leurs parents (*disobeying their parents*) - Il [ne veut] pas [venir] (*He doesn't want to come*) - [ils etaient] [fermés] (*they were closed*)
GA	- les barreaux [intacts] (*the intact bars*) - la solution [retenue] fut... (*the chosen solution was...*) - les enfants [désobéissants] (*the disobeying children*)
GR	- [aussi] (*also*), vous n'auriez [pas] (*you would not*)
PV	- [pour aller] à Paris (*to go to Paris*), [de vraiment bouger] (*to really move*)

Table 1: [Chunk examples], (with their English translation).

3 Syntactic relation annotation

The dependencies establish all the links between the chunks described above. All participants, corpus providers and campaign organizers agreed on a list of 14 kinds of dependencies listed below:

- subject-verb (**SUJ-V**): may be inside the same NV as between *elle* and *était* in *elle était* (*she was*), or between a GN and a NV: *Mademoiselle appelait* (*Miss was calling*);
- auxiliary-verb (**AUX-V**), between two NVs: *on a construit une maison* (*we have built a house*);
- attribute-subject/object (**ATB-SO**): between the attribute and the verb kernel, and precising that the attribute is relative to (a) the subject: *il est grand* (*he is tall*), or (b) the object: *il trouve cette explication étrange* (he finds this explanation strange);
- 3 kinds of dependencies between the verb and complements or modifiers
  - direct object-verb (**COD-V**): *on a construit la première automobile* (*we have built the first car*);
  - complement-verb (**CPL-V**): in case of adjuncts or indirect objects: *en quelle année a-t on construit* la première automobile (*In which year did we build the first car*);

  - modifier-verb (**MOD-V**):for not mandatory modifiers, as adverbs or adjunct clauses:
    Jean dort quand la nuit tombe (*Jean sleeps when the night falls*);

- complementor (**COMP**): to link the introducer and the verb kernel of a subordinate clause: *Je pense qu' il viendra* (*I think that he will come*); it is also used to link a preposition and a noun phrase when they are not contiguous, preventing us from annotating them as a GP as in:
  avec dans sa poche un discours (see Figure1);
- different modifiers to relate to the noun (resp. adjective, adverb or preposition) all the chunks which modify it:
  - modifier-noun (**MOD-N**): for the adjective, the genitive, the relative clause: *l'unique fenêtre* (*the unique window*); *la porte de la chambre* (*the bedroom door*);
  - modifier-adjective (**MOD-A**): *la très belle collection* (*the very impressive collection*) or *elle est fière de son fils* (*she is proud of her son*);
  - modifier-adverb (**MOD-R**): *elle vient très gentiment* (*she comes very kindly*);
  - modifier-preposition (**MOD-P**): *elle vient peu avant lui* (*she comes little before him*);
- coordination (**COORD**): to relate the coordination and the coordinated elements, as between *Pierre*, *Paul* and *et* in *Pierre et Paul arrivent* (*Pierre and Paul are arriving*);
- apposition (**APP**): to link the elements which are placed side by side, when they refer to the same object: *Le député Yves Tavernier...* (*the MP Yves Tavernier...*);
- juxtaposition (**JUXT**): to link chunks which are neither coordinated nor in an apposition relation, as in an enumeration. It also links clauses as in:
  on ne l' entendait plus ... elle était peut-être morte (*we did not hear her any more... perhaps she was dead*).

Some examples are provided in Figure 1 or in section 4.

4 Some elements of comparison with SD, GR and PARC

In this section, we consider how the PASSAGE annotation scheme addresses the comparison elements that [3] used to ascertain the position of the Standford typed

dependencies (SD) against the GR and PARC representations. Since SD was designed with task based evaluation as opposed to intrinsic evaluation, we found these elements of comparison to be more likely to enhance the contrasts between the two representations. We briefly address: argument/adjunct distinction, NP-internal relations, noun-modifier dependencies, head identification, the SD dependency collapsing mechanism, preposition modifiers, arity of the syntactic relations and the choice of having a representation more oriented towards syntax or semantics.

The SD scheme is not concerned with the argument/adjunct distinction but in contrast it includes many NP-internal relations such as *appos* (appositive modifier), *nn* (noun compound), *num* (numeric modifier), number (element of compound number) and *abbrev* (abbreviation). The following example, taken from [3] *"I feel like a litle 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)* yields the following relations which we have completed with the PASSAGE ones in table 4. It shows that PASSAGE includes dependencies similar to the other schemes, but of a coarser grain for what concerns the dependencies source/target text extents. PASSAGE was designed with the aim of addressing only the essential level of syntactic functions, leaving aside finer grain relations like the determiner one 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 was not designed either to address semantics since the conditions of its creation representation were intrinsic parser evaluation.

Note that with PASSAGE, intra-chunk relations such as the MOD-N relation between *gleeful* and *Alex* 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 consituent holding the apposed adjective and an intra-chunk MOD-N relation instead of an adjectival and nominal chunk linked by a MOD-N relation because adjectives occuring before the noun are much less frequent in French than those occuring after and the corresponding syntactic structure is generally straightforward.

The last remarkable point about Table 4 lies in the label variability among the representation schemes for the *to slip - stopped* dependency. PASSAGE does not have the notion of head, instead it uses its six basic chunks presented in section 1 to restrict the portion of text where a head can possibly occur. Since the notion of head is controversial, see for instance [9], we did not want to have in the PASSAGE formalism any explicit reference to this notion, not even in the documentation describing how to annotate chunks. Our initial design choice was motivated by the wish to have a syntactic representation as simple as possible to ease up the annotation task and to be able to compare parsing schemes which have heads against

ones which do not. Note that PASSAGE does not forbid to address single words as target dependency, so a representation scheme with heads can be mapped directly onto PASSAGE, at the price that if a comparison is done with another annotation which has chunks, the precise location of the head will be lost in the process, since it will be assimilated with the scope of the enclosing chunk. So instead of identifying *effective* as head of the quoted phrase like SD in the example used by [3] *Considered as a whole, Mr Lane said, the filings required under the proposed rules "will be at leas as* **effective***, if not more so, for investors following transactions"*

Scheme	Relation
SD	appos(Castro, salesman)
PARC	appos(Alex de Castro, salesman)
GR	adjunct(Castro, salesman)
PASSAGE	appos(FIRST:[a gleeful Alex de Castro]^{GN} APPOSED:[a car salesman]^{GN})
SD, PARC, GR	num(card, six)
SD	nn(cards, Campaneris)
PASSAGE	mod-n(MODIFIER:six, NOUN:cards)
SD	amod(Castro, gleeful)
PARC	adjunct(Alex de Castro, gleeful)
PASSAGE	mod-n(MODIFIER:gleeful, NOUN:Alex)
SD	amod(kid, little)
PARC	adjunct(kid, little)
PASSAGE	mod-n(MODIFIER:little, NOUN:kid)
SD	xcomp(stop, slip)
PARC	adjunct(stop, slip)
PASSAGE	mod-v(MODIFIER:[to slip]^{NV}, VERB:[stopped]^{NV})
SD	prep_of(workout, Suns)
PARC	adjunct(workout, of)
PASSAGE	MOD-N(MODIFIER:[of the Suns]^{GP}, NOUN:[by a workout]^{GP})

Table 2: Comparing SD, PARC, GR and PASSAGE, an example.

(WSJ-R) or identifying *be* as head like GR, PASSAGE will identify *will* as a target of a subject-verb relation originating at *filings*. Here we see that PASSAGE is nearer GR, we could say in a way more oriented toward the coding of explicit syntax, while SD tends to address more semantic aspects. Lastly, it is not because in some cases PASSAGE chunks identify to heads when they have a size 1 (e.g. with some verbal chunks) that one could systematize such correspondence, for instance identifying a head in a verbal chunk become problematic when it contains also the clitic pronoun. In the same idea of annotating an explicit link between content words, SD proposes a collapsing mechanism for dependencies involving prepositions, in the previous example, instead of having two relations, as will be the case with GR and PARC, SD may have only one given in the table 4. Here PASSAGE

Scheme	Relation
PARC	adjunct(workout, of), obj(of, Suns)
GR	ncmod(workout, of), dobj(of, Suns)
SD	prep_of(workout, Suns)
PASSAGE	MOD-N(MODIFIER:[of the Suns]^{GP},NOUN:[by a workout]^{GP})

Table 3: Collapsing dependencies with SD

is closer to the SD representation, since it will also have a single MOD-N relation between the two chunks identified as prepositional chunks (GP tags above).

The fact that SD leans toward semantics and PASSAGE does toward syntax can also be seen in the example [3]: *A similar technique is almost impossible to apply to other crops, such as cotton, soybean and rice (WSJ-R)* for which SD gives direct *prep_such_as* links between *crops* and all the coordination elements while this information can only be accessed through indirect links with PASSAGE as shown in figure 2.

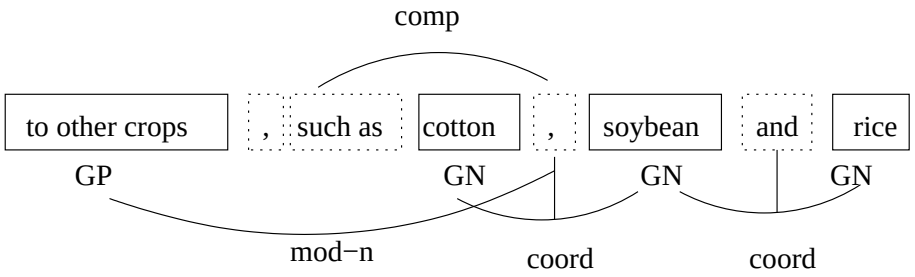

Figure 2: PASSAGE annotation of the “crops” example

One point where PASSAGE is more faithful to linguistic interpretation that SD concerns preposition modifiers; while PASSAGE has a *MOD-P* relation for the purpose (see 4), SD binds the preposition modifier to the head of the clause in which they appear and not on the preposition itself [3].

Although PASSAGE is very close to GR, it does not annotate explicitly passive constructions (while GR, PARC and SD do, for a comparison see [8]) or deep subjects: a deliberate choice to stick to explicit first level interpretation for the initial trial that was the first PASSAGE evaluation campaign. But extra annotations are already in discussion for the version that will be used for semantic extraction at the end of the PASSAGE project.

[3] argue that to have only binary relations (all dependencies are a triples, a grammatical relation, the head and the dependent), makes the representation more readable, easier to encode in software and to map to semantic WEB representation such as OWL and RDF triples. If we agree with the later points, we chose in PASSAGE to have a surface form for the annotations which have some ternary relations because they are closer to the intuitive syntactic representation (in general a coordination involves three elements, a conjunction and two conjuncts). Nevertheless, all these ternary relations, except the coordination, have a third argument which is used only for specifying a subtype of the relation, like for instance the distinction between subject-attribute versus object-attribute for the verb-attribute relation and could be automatically transformed into a canonical binary representation by defining new relations for the subtypes. PASSAGE coordination relation could also be transformed automatically into two coordination relations, one linking the left part and the conjunction and the other for the right part. Here we address "syntactic sugar" issues, since we allow instances of the coordination relation where the left part is empty, for instance when a coordination conjunction is used to start a sentence, a phenomena often encountered in speech transcriptions.

Another characteristic of PASSAGE that we see as an advantage over SD for what concerns the annotation task, comes from having a representation as intuitive as possible in terms of syntax; we are not faced with the dilemma of either altering the semantics of a sentence or duplicating verbs when dealing with preposition conjunction and preposition collapsing. For the sentence *Bill went over the river and right through the woods* from [3], SD will duplicate the verb *went* while PASSAGE will straightforwardly represent the initial syntactic information preserving uniformity of handling of the two GPs at the price of an indirection throughout the coordination relation to represent the link between *went* and its two adjuncts, see table 4. We see here that PASSAGE adopts the "Prague style" for annotating coordination, see [6] for a more detailed discussion. 4.

Scheme	Relation
SD	a prep_over(wen-2, river-5) prep_through(went-2', woods-10) conj_and(went-2, went-2')
PASSAGE	MOD-V(MODIFIER: and, VERB:[went]^{NV}) COORD([over the river]^{GP}, [through the woods]^{GP}) MOD-P(MODIFIER:right, PREPOSITION:through)

Table 4: Preposition collapsing with SD

5 EASYREF the collaborative software support for PASSAGE

EASYREF is a collaborative WEB browsing/editing/versioning software developed by INRIA for corpus annotated with the PASSAGE representation. The display uses a linear representation of the sentences with color-coded chunks above the forms. The idea was extended to dependencies, represented on several lines below the forms using color codes related to their type and span given by their anchors. One may select which kinds of dependencies are to be displayed and when moving the mouse over a dependency, its anchors are highlighted and a tooltip box is displayed providing more detailed information. For a given sentence, it is possible to show or hide additional pieces of information, such as the list of its associated bug reports, the history of its revisions and a list of potential annotation errors automatically detected by EASYREF.

Sentences may be searched using various administrative and linguistic criteria; e.g. one may search for all the sentences with potential errors but no bug reports, or for sentences with reports but no corrections. A more linguistic query such as "\verb+évaluer@NV les@GN+" would return all the sentences where the word "évaluer" (*evaluate*) in a NV chunk is followed by "les" (*the*) in a GN chunk. Linguistic queries are applied as regular expressions on a linear representation of both text and chunk annotations. Querying with syntactic relations is under development along with an extension of the PASSAGE annotation scheme which will have lemma, morphosyntactic tags, a fine grained representation of token/word form segmentation and nested chunks.

It is also possible to compare two sets of annotations coming from two different parsers, for instance, as illustrated in Figure 3. Color codes provide an easy identification of mismatching chunks. Comparing dependencies is more complex:

both sets of dependencies are actually mixed in the display, here the text color and weight indicate the status of the dependencies: shared or belonging only to one set.

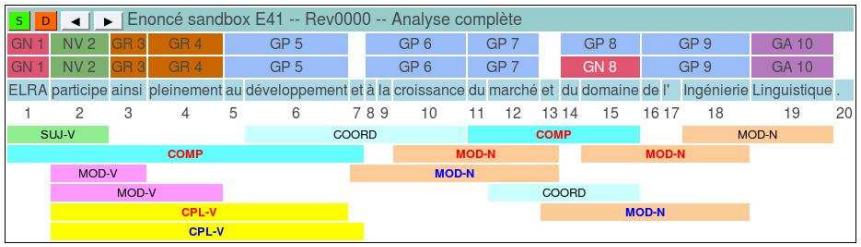

Figure 3: Comparing two annotation sets.

6 Conclusion

We have presented the PASSAGE syntactic representation based on syntactic relations, initially developed for French in the scope of national evaluation campaigns and shown that it stands closer to GR than SD or PARC. Software support for the representation is provided by EASYREF, a collaborative WEB browser/editor that we have presented. PASSAGE representation contributes to the ongoing debate on a common pivot formalism for syntactic information by proposing to replace the requirement of precise head localization with one level of chunking in complement of its functional dependencies.

7 Acknowledgements

We wish to thank the organizers and participants to the workshop on Cross-Framework and Cross-Domain Parser Evaluation at COLING 2008 for their warm welcome and useful suggestions.

References

[1] S. Buchholz, J. Veenstra, and W. Daelemans. Cascaded grammatical relation assignment. In *In proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 239–246, 1999.

[2] J. Carroll, D. Lin, D. Prescher, and H. Uszkoreit. Proceedings of the workshop "Beyond Parseval - Toward improved evaluation measures for parsing systems". In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Spain, 2002.

[3] M.-C. de Marneffe and C. D. Manning. The Stanford typed dependencies representation. In *Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation at COLING 2008*, pages 1–8, Manchester, August 2008. Association for Computational Linguistics.

[4] V. Gendner, G. Illouz, M. Jardino, L. Monceaux, P. Paroubek, I. Robba, and A. Vilnat. PEAS the first instanciation of a comparative framework for evaluating parsers of French. In *Proceedings of the 10^{th} Conference of the European Chapter fo the Association for Computational Linguistics*, pages 95–98, Budapest, Hungary, April 2003. ACL. Companion Volume.

[5] E. de la Clergerie, O. Hamon, D. Mostefa, C. Ayache, P. Paroubek, and A. Vilnat. Passage: from French parser evaluation to large sized treebank. In ELRA, editor, *In proceedings of the sixth international conference on Language Resources and Evaluation (LREC)*, Marrakech, Morroco, May 2008.

[6] J. Nilsson, J. Nivre, and J. Hall. Graph transformations in data-driven dependency parsing. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 257–264, Sydney, Australia, July 2006. Association for Computational Linguistics.

[7] A. Vilnat C. Ayache P. Paroubek, I. Robba. Data, annotations and measures in EASY - the evaluation campaign for parsers of French. In ELRA, editor, *In proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006)*, pages 315–320, Genoa, Italy, May 2006.

[8] L. Rimell and S. Clark. Constructing a parser evaluation scheme. In *Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation at COLING 2008*, pages 44–50, Manchester, August 2008. Association for Computational Linguistics.

[9] Yoav Seginer. *Learning Syntactic Structure*. PhD thesis, Institute for Logic, Language and Computation, University of Amsterdam, October 2007. ILLC Dissertation Series DS-2007-05.

[10] A. Vilnat, G. Francopoulo, O. Hamon, S. Loiseau, P. Paroubek, and E. Villemonte de la Clergerie. 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*, pages pp 36–43, Manchester, August 2008.