4-Couv: A New Treebank for French

Philippe Blache, Grégoire Montcheuil, Laurent Prévot, Stéphane Rauzy
Laboratoire Parole et Langage - UMR 7309
5 avenue Pasteur, 13100 Aix-en-Provence, France
firstname.lastname@lpl-aix.fr

Abstract

The question of the type of text used as primary data in treebanks is of certain importance. First, it has an influence at the discourse level: an article is not organized in the same way as a novel or a technical document. Moreover, it also has consequences in terms of semantic interpretation: some types of texts can be easier to interpret than others. We present in this paper a new type of treebank which presents the particularity to answer to specific needs of experimental linguistic. It is made of short texts (book backcovers) that presents a strong coherence in their organization and can be rapidly interpreted. This type of text is adapted to short reading sessions, making it easy to acquire physiological data (e.g. eye movement, electroencepholagraphy). Such a resource offers reliable data when looking for correlations between computational models and human language processing.

Keywords: Treebank, French, Constituency, Treebanking Tools, Experimental Linguistic

1. Introduction

Several treebanks already exist for French. The most popular one is the *French Treebank* (Abeillé et al., 2003, *FTB*) and its different evolutions or enrichments (Schluter and van Genabith, 2007; Candito et al., 2010). Other French dependency treebanks are also available, for example through the *Sequoia project* (Candito and Seddah, 2012; Candito et al., 2014) or the *Universal Dependencies* project (Nivre et al., 2015). However, and this constitutes the first motivation for developing a new treebank, it still remains necessary to develop other resources in order to increase the size and the variety of available material. A second and even more important reason to build specific treebanks is the type of application it can be used for. More precisely, classical treebanks are usually made of newspaper articles, rather long, and which interest when reading can be poor. This can constitute a drawback (or even a bias) when using such texts in human experimentation (in particular reading). In such cases, short, semantically coherent and self-contained texts are preferable for the acquisition of physiological responses from reading subjects.

The solution we propose in this perspective consists in building a corpus of book backcovers[1] , that are short texts, easily available and frequently used. We present in this paper the corpus[2], its annotation and first results.

2. Corpus description

Backcovers are short texts of different genres: extract, synopsis, genesis of the book, comment about the work, or a combination of them. Each text is semantically self-contained, and interesting (minimizing attention and comprehension drops). A text contains between 4-10 sentences and 80-200 tokens (80% having at most 30 tokens).

We have build such a corpus, which is still under evolution. At this stage, it consists in 8,000 texts from different publishers (from Pocket and Gallimard publishers, with their agreement). These texts have been selected manually by three experts, according to a set of criteria among which interest when reading and semantic coherence. A subset of 500 texts, representing 3,500 sentences, has been annotated at morpho-syntactic, syntactic and to some extent discourse levels.

3. Annotations

4-Couv is a long-term project, that will be enriched progressively. The annotation process has then to be carefully documented. We present in this section the main features of the annotation guide and its application. We have decided to use a constituency-based representation first because we already have developed a parser in this format, second because we already did several experiment with the FTB, using different complexity indexes based on such format. An hybrid version of our treebank, integrating a dependency version, is planned.

3.1. Annotation guide

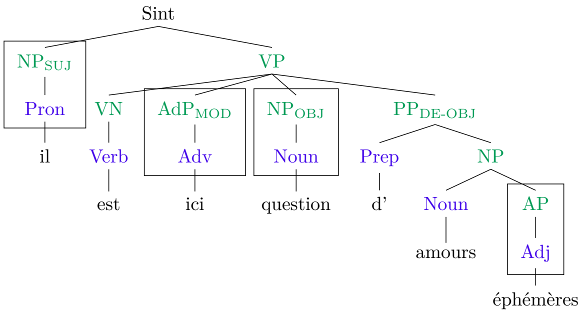
Most of the works in French parsing being done starting from the *FTB*, we decided to stay as close as possible with the *FTB* format. So the treebank is constituency-based and syntactic relations are represented by means of trees. However, the *FTB* contains several specific annotations, aiming at reducing the embedding of the trees (limitation of the projections, suppression of the VP level, affixation of clitics, etc.). In order to be compatible with more standard formats used in other treebanking projects, we slightly modified the initial *FTB* annotation guide. More generally, we apply the following formal constraints:

- no empty category is inserted in the trees (e.g. in the case of an elliptical construction), each node is instantiated by a lexical or a phrase-level unit.
- distinction between lexical and phrase level: we keep unary phrases, e.g. in (1) *il* or *question* are the unique constituent of a *NP*, as *ici* is of a *AdP* and *éphémère* is of a *AP*.

[1] The French appellation of a backcover is "*4ème de couverture*", leading to the treebank acronym: *4-Couv*.

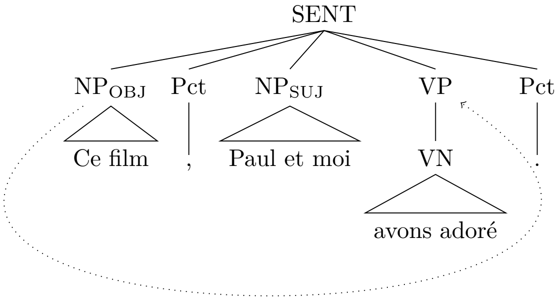
[2] Available on the ORTOLANG platform, hdl:11403/4-couv

(1) *"il est ici question d'amours éphémères"*
(*it is here an issue of ephemeral loves*)



- no discontinuous constituent or unbounded dependencies directly encoded, such as in (2).

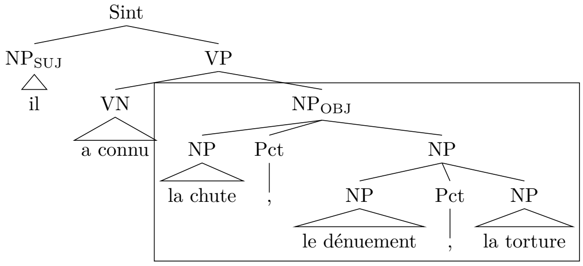
(2) *"Ce film, Paul et moi on a adoré."*
(*This movie, Paul and I we really do like.*)



- the phrase-level tagset (see figure 1) is reduced to classical phrases, at the exclusion of other constructions such as coordination (at the difference with the *FTB* and its derivatives).

- coordinations are represented by a succession of binary trees, as in (3).

(3) *"il a connu la chute, le dénuement, la torture"*
(*he known the fall, the deprivation, the torture*)



- the same types of syntactic functions than those introduced for the *FTB* (see figure 1) are used. This annotation is less precise then other annotation frameworks, such as (Gendner et al., 2009) where structural and functional informations were given independently.

Phrase-level constructions			
AdP	adverbial phrase	VN	verbal nucleus
AP	adjectivial phrase	VNinf	infinitive VN
NP	noun phrase	VNpart	participial VN
PP	prepositional phrase	SENT	sentence
VP	verbal phrase	Srel	relative clause
VPinf	infinitive clause	Ssub	subordinate clause
VPpart	participial clause	Sint	other clause
Syntactic functions			
SUJ	subject	indirect complement	
OBJ	direct object	A-OBJ	- introduced by *à*
MOD	modifier or adjunct	DE-OBJ	- introduced by *de*
predicative complement		P-OBJ	- other preposition
ATS	- of a subject		
ATO	- of a direct object		

Figure 1: Syntactic tagset

3.2. Morphosyntax

At the lexical level, the first step has concerned tokenization. It is based on the *MarsaLex* French lexicon[3], containing 595.000 entries with their frequencies. Tokenization is maximal in that even highly constrained forms are split into distinct lexical units provided they follow syntactic composition rules. For example, constituents of semi-fixed expressions such as *"il était une fois"* (*once upon a time*) or *"mettre à nu"* (*lay bare*) are decomposed, at the difference with other multiwords expressions as *"d'autant plus"* (*all the more*) or *"tant mieux"* (*great*) as they do not follow any syntactic composition.

Each lexical category has a specific features set (see figure 2), although many features are common to different categories (typically the gender, number, person). The part-of-speech and feature sets are relatively standard and compatible with most of automatically tagged corpus, and enable to indicate a combination of lexical, morphologic, syntactic and occasionally semantic informations that will have effect on the syntactic construction of upper levels, e.g. the number of a determiner, the subcategorization or the case of a clitic pronoun. We do not have discontinuous lexical constituent, and the tagging is disambiguated (i.e. each element have one part-of-speech, whose sub-categories features could be underspecified when necessary). We do not modify the category of units that change their paradigm (*"une tarte maison"* (*an home[made] pie*), *"il est très zen"* (*he is very zen*)).

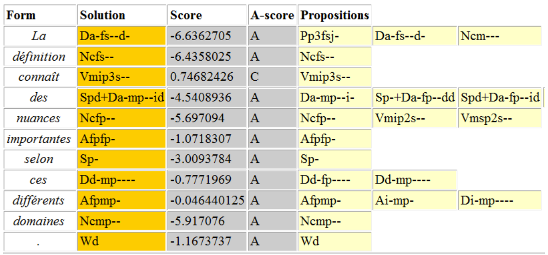
The second step, POS-tagging, is done with the *MarsaTag*[4] tagger (0.975 F-Measure on written texts), trained on the LPL-Grace corpus (700,000 tag manually corrected). It associates each token with a list of possible tags with their probability, as described in figure 3.

The automatic POS tagging has been manually corrected thanks to a specific editing tool, making it possible to correct labels as well as features (see section 4.2.). Among recurrent errors, several concern the agreement features, in particular determiner and adjective genders. Only few errors concern categorization itself, which confirms the F-

[3] hdl:11041/sldr000850
[4] hdl:11041/sldr000841

Category	features
Adjective	nature, type, gender, number, position
Adverb	nature, type
Connector	nature
Determiner	nature, type, person, gender, number
Interjection	
Noun	nature, type, gender, number, referent type
Punctuation	nature
Preposition	type
Pronoun	nature, type, person, gender, number, case, reflective, postposed
Verb	nature, modality, tense, person, gender, number, auxiliary, pronominal, (im)personal, direct object, indirect complement

Figure 2: Lexical categories and features



Form	Solution	Score	A-score	Propositions		
La	Da-fs--d-	-6.6362705	A	Pp3fsj-	Da-fs--d-	Ncm---
définition	Ncfs--	-6.4358025	A	Ncfs--		
connaît	Vmip3s--	0.74682426	C	Vmip3s--		
des	Spd+Da-mp--id	-4.5408936	A	Da-mp--i-	Sp-+Da-fp--dd	Spd+Da-fp--id
nuances	Ncfp--	-5.697094	A	Ncfp--	Vmip2s--	Vmsp2s--
importantes	Afpfp-	-1.0718307	A	Afpfp-		
selon	Sp-	-3.0093784	A	Sp-		
ces	Dd-mp----	-0.7771969	A	Dd-fp----	Dd-mp----	
différents	Afpmp-	-0.046440125	A	Afpmp-	Ai-mp-	Di-mp----
domaines	Ncmp--	-5.917076	A	Ncmp--		
.	Wd	-1.1673737	A	Wd		

Figure 3: POS tagging

score level of the tagger. Finally, many errors concern words with a high amount of possible tags (category and feature ambiguity) such as "*comme*" or "*que*".

3.3. Syntax

The generation of the treebank has been done automatically thanks to the *MarsaTag* parser, trained on a subset of 100,000 words of the FTB, manually corrected and labelled following the *4-Couv* annotation format (as shown in figure 4).

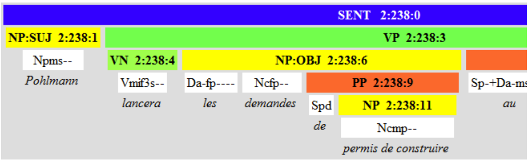


Figure 4: Output of the parser

Among the initial corpus of 500 texts, 200 have been manually corrected, representing 1,500 trees. The correction has been done thanks to a specific editor (see section 4.3.). Several recurrent types of errors has been identified and fixed, such as the coordination between conjuncts of different types, as well as enumerations. Other kinds of errors come, classically, from ambiguous attachments. Some of the errors concern the labelling. For example `VPinf` introduced by a `Prep` should be encoded as a `PP` plus a `VPinf`. Finally, several errors concern function labels (clitics and subordinates, in particular).

3.4. Discourse

Back covers naturally host interesting discourse structures. We consider them as a good starting point for applying eye-tracking methodology to discourse structure studies.
As a pilot study (Prévot et al., 2015), we annotated 7 back covers selected for their interest following *Annodis* guidelines (Muller et al., 2012). These guidelines provide instruction to segment a text into *discourse units* as well as to relate the units through discourse relations such as *Narration, Explanation, etc.*. The annotation framework is grounded in *Segmented Discourse Representation Theory* (Asher and Lascarides, 2003). From the structures annotated we are able to extract a set of predictors for reading time corresponding to the main characteristics of the discourse structures. The predictors extracted were, for each discourse unit: *# relations involving the unit, # potential attachment points, distance to attachment point, length of furthest anaphoric, New topic , Position in the discourse.*
The pilot study involving only 16 subjects reading these 7 texts have not provided enough data to establish clear discourse structure impact on reading time. The results however validated our overall set-up since known effects about proper names, numbers and word size (at token level) taking longer to read were replicated. The amount of data needed to establish subtle discourse constraints is too large[5] given the costly discourse annotation process. However a solution of manipulating control 'original' texts annotated to form hypotheses-driven modified text from an intermediate size corpora seems to be a very promising option. It will allow to preserve all the good properties of back covers dataset while allow discourse experiements by direct comparison of parallel texts.

4. Treebanking tools

The *4-Couv* treebank required the development of two different types of tools, adapted to this project: corpus building (selection, checking and ranking of the texts) and tree edition for manual correction.

Annotation layer	automatic tool	manual (post-)edition
Interestingness	-	*Text selector*
POS	*MarsaTag*	*POS-tagging corrector*
Syntax	*MarsaTag*	*Tree editor*
Discourse	?	???

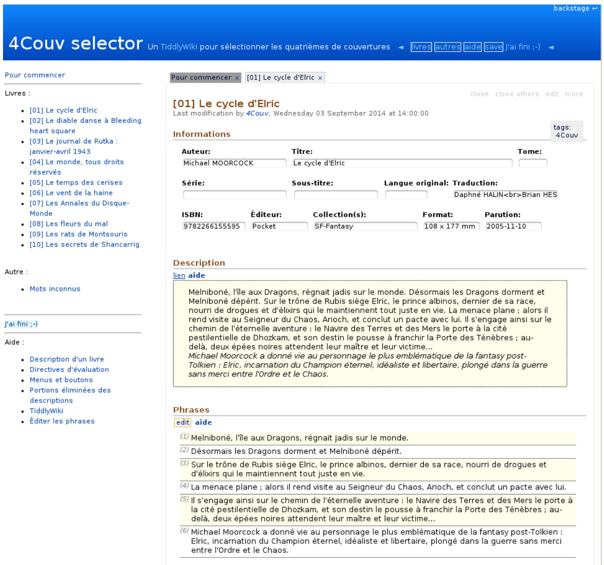
Figure 5: Annotation layers and tools

4.1. Text Selector

The *Text Selector* is a tool helping in the selection of the texts, on the form of HTML files each containing 10 texts to evaluate. Each unit presents the book description and the text, segmented into sentences. It also proposes an evaluation form (containing check boxes and drop-down lists),

[5] Our previous study on morpho-syntax effects (Blache and Rauzy, 2012) included 10000 tokens).

and the list of unknown words, to be manually tagged. This interface (see figure 6) makes it possible to easily correct different types of errors, including sentence segmentation as well as metadata.
Using autonomous HTML files makes easier the distribution of the revision work between several annotators. It does not require any particular environment (files being edited directly in a browser), neither a connection to a server. The revision tool relies on an adaptation of a *TiddlyWiki*[6] enriched with scripts for adding extra information to the texts.



backstage ↩

4Couv selector Un TiddlyWiki pour sélectionner les quatrièmes de couvertures ◂ livres autres aide save J'ai fini ;-) ◂

Pour commencer

Livres :

- [01] Le cycle d'Elric
- [02] Le diable danse à Bleeding heart square
- [03] Le journal de Rutka : janvier-avril 1943
- [04] Le monde, tous droits réservés
- [05] Le temps des cerises
- [06] Le vent de la haine
- [07] Les Annales du Disque-Monde
- [08] Les fleurs du mal
- [09] Les rats de Montsouris
- [10] Les secrets de Shancarrig

Autre :

- Mots inconnus

J'ai fini ;-)

Aide :

- Description d'un livre
- Directives d'évaluation
- Menus et boutons
- Portions éliminées des descriptions
- TiddlyWiki
- Éditer les phrases

Pour commencer × [01] Le cycle d'Elric ×

close close others edit more

[01] Le cycle d'Elric
Last modification by *4Couv*, Wednesday 03 September 2014 at 14:00:00

tags: 4Couv

Informations

Auteur: Michael MOORCOCK
Titre: Le cycle d'Elric
Tome:
Série:
Sous-titre:
Langue original:
Traduction: Daphné HALIN
Brian HES
ISBN: 9782266155595
Éditeur: Pocket
Collection(s): SF-Fantasy
Format: 108 x 177 mm
Parution: 2005-11-10

Description

lien **aide**

Melniboné, l'île aux Dragons, régnait jadis sur le monde. Désormais les Dragons dorment et Melniboné dépérit. Sur le trône de Rubis siège Elric, le prince albinos, dernier de sa race, nourri de drogues et d'élixirs qui le maintiennent tout juste en vie. La menace plane ; alors il rend visite au Seigneur du Chaos, Arioch, et conclut un pacte avec lui. Il s'engage ainsi sur le chemin de l'éternelle aventure : le Navire des Terres et des Mers le porte à la cité pestilentielle de Dhozkam, et son destin le pousse à franchir la Porte des Ténèbres ; au-delà, deux épées noires attendent leur maître et leur victime...
Michael Moorcock a donné vie au personnage le plus emblématique de la fantasy post-Tolkien : Elric, incarnation du Champion éternel, idéaliste et libertaire, plongé dans la guerre sans merci entre l'0rdre et le Chaos.

Phrases

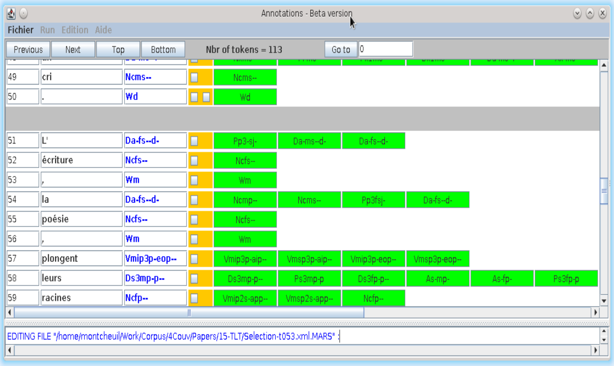
edit **aide**

(1) Melniboné, l'île aux Dragons, régnait jadis sur le monde.
(2) Désormais les Dragons dorment et Melniboné dépérit.
(3) Sur le trône de Rubis siège Elric, le prince albinos, dernier de sa race, nourri de drogues et d'élixirs qui le maintiennent tout juste en vie.
(4) La menace plane ; alors il rend visite au Seigneur du Chaos, Arioch, et conclut un pacte avec lui.
(5) Il s'engage ainsi sur le chemin de l'éternelle aventure : le Navire des Terres et des Mers le porte à la cité pestilentielle de Dhozkam, et son destin le pousse à franchir la Porte des Ténèbres ; au-delà, deux épées noires attendent leur maître et leur victime...
(6) Michael Moorcock a donné vie au personnage le plus emblématique de la fantasy post-Tolkien : Elric, incarnation du Champion éternel, idéaliste et libertaire, plongé dans la guerre sans merci entre l'Ordre et le Chaos.

Figure 6: Selection tool: global view

4.2. POS-tagging corrector

The morphosyntactic correction tool (see figure 7) presents one token per line, each line containing the form and the list of possible tags associated to the form, starting with the chosen one. All possible labels anf features are then proposed. Correcting a tag simply consists in chosing a new one in the list or manually editing it.



Nbr of tokens = 113

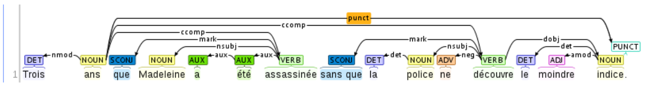
49	cri	Ncms--	□	Ncms--					
50	.	Wd	□ □	Wd					
51	L'	Da-fs--d-	□	Pp3-sj-	Da-ms--d-	Da-fs--d-			
52	écriture	Ncfs--	□	Ncfs--					
53	,	Wm	□	Wm					
54	la	Da-fs--d-	□	Ncmp--	Ncms--	Pp3fsj-	Da-fs--d-		
55	poésie	Ncfs--	□	Ncfs--					
56	,	Wm	□	Wm					
57	plongent	Vmip3p-eop--	□	Vmip3p-aip--	Vmsp3p-aip--	Vmip3p-eop--	Vmsp3p-eop--		
58	leurs	Ds3mp-p--	□	Ds3mp-p--	Ps3mp-p	Ds3fp-p--	As-mp-	As-fp-	Ps3fp-p
59	racines	Ncfp--	□	Vmip2s-app--	Vmsp2s-app--	Ncfp--			

EDITING FILE "/home/montcheuil/Work/Corpus/4Couv/Papers/15-TLT/Selection-t053.xml.MARS" :

Figure 7: POS-tagging corrector

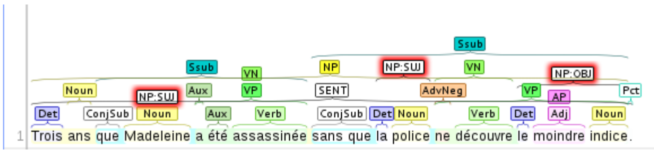
4.3. Tree editor

The syntactic correction tool is a tree editor. Only few of them already exist such as *WordFreak* (Morton and LaCivita, 2003) or *TrED 2.0* (Pajas and Štěpánek, 2008). More recently, some web-based annotation platforms have also been developed, offering an intuitive and fast annotation (*brat* (Stenetorp et al., 2012), as well as project management facilities for example by specifying the roles such as annotator, curator or project manager (*GATE Teamware* (Bontcheva et al., 2013), *WebAnno* (Yimam et al., 2013)). However, if some of these platforms have been developed or adapted for dependency-based treebanks (see figure 8), none is suited for constituency-based treebanks (requiring therefore to deal with a potentially large number of levels, see figure 9).



Trois ans que Madeleine a été assassinée sans que la police ne découvre le moindre indice.

Figure 8: a dependency tree with *brat*



Trois ans que Madeleine a été assassinée sans que la police ne découvre le moindre indice.

Figure 9: a constituent tree with *brat*

To take advantage of web-based platforms, we developed a specific editor javascript library, that could run in a single HTML or to be integrated into an annotation platform such as *brat* (see figure 10) or *WebAnno*. The library shows each tree in a resizable and zoomable vectorial image (SVG), whose colors are customizable with CSS style sheets. Subtrees could be moved by a simple drag and drop. A double-click on nodes allows to edit its tag, and buttons or contextual menu offers other edition functionalities (insertion, deletion, etc., see figure 11).

5. Perspectives

The *4-Couv* treebank constitutes a new kind of resources, answering at the same time to the classical needs of linguistic description as well as experimental linguistics. This treebank, because being made of short texts, can propose a complete annotation at both syntactic and discourse levels. Moreover, this characteristics also makes it possible to acquire physiological data such as eye movement or brain activity by controlling easily different parameters. A first eye-tracking experiment have been done, studying different effects of word categories.
4-Couv is an ongoing long term project. The first release (200 texts, 1,500 trees) will be distributed through the SLDR data warehouse (`http://www.sldr.org`).
4-Couv is also becoming multilingual: a comparable treebank is under construction for Mandarin Chinese thanks to a collaboration with Hong Kong Polytechnic University.

[6] `http://classic.tiddlywiki.com/` (version 2.8.1)

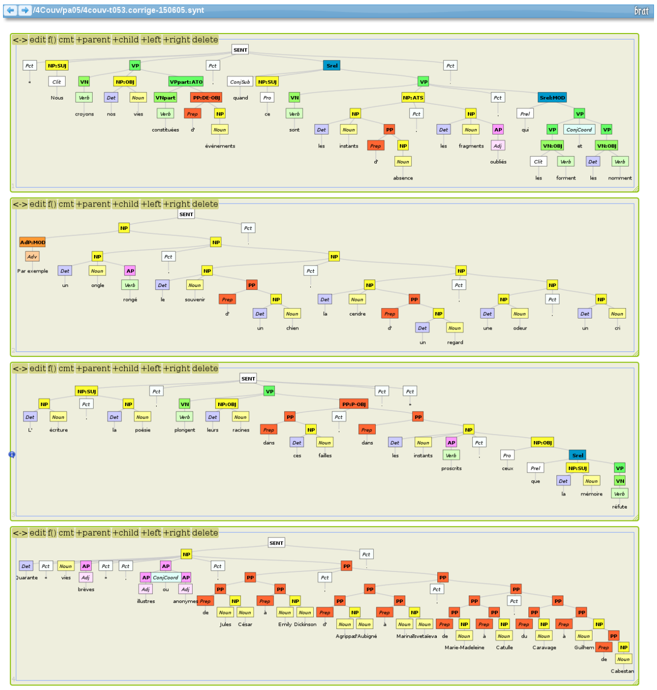

/4Couv/pa05/4couv-t053.corrige-150605.synt
brat
<-> edit f() cmt +parent +child +left +right delete
SENT

Figure 10: the syntactic tree editor in *brat*

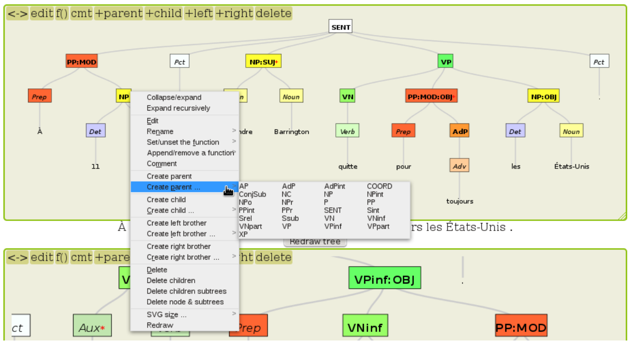


Figure 11: editing a syntactic tree

6. Acknowledgments

This work has benefited from the French State support via its “Investissements d’avenir” programs ORTOLANG (ANR-11-EQPX-0032), BLRI (ANR-11-LABX-0036) and A*MIDEX (ANR-11-IDEX-0001-02).

7. Bibliographical References

Asher, N. and Lascarides, A. (2003). *Logics of conversation*. Cambridge University Press.

Blache, P. and Rauzy, S. (2012). Robustness and processing difficulty models. a pilot study for eye-tracking data on the french treebank. In *24th International Conference on Computational Linguistics*, page 21.

Bontcheva, K., Cunningham, H., Roberts, I., Roberts, A., Tablan, V., Aswani, N., and Gorrell, G. (2013). Gate teamware: a web-based, collaborative text annotation framework. *Language Resources and Evaluation*, 47(4):1007–1029.

Gendner, V., Vilnat, A., Monceaux, L., Paroubek, P., Robba, I., Francopoulo, G., and Guénot, M.-L. (2009). Les annotation syntaxiques de référence PEAS. Technical report, version 2.2.

Morton, T. and LaCivita, J. (2003). Wordfreak: An open tool for linguistic annotation. In *Proceedings of NAACL-Demonstrations ’03*, pages 17–18, Stroudsburg, PA, USA. Association for Computational Linguistics.

Muller, P., Vergez-Couret, M., Prévot, L., Asher, N., Benamara, F., Bras, M., Le Draoulec, A., Vieu, L., et al. (2012). Manuel d’annotation en relations de discours du projet annodis. Technical report, Toulouse University.

Pajas, P. and Štěpánek, J. (2008). Recent advances in a feature-rich framework for treebank annotation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 673–680, Manchester, UK, August. Coling 2008 Organizing Committee.

Prévot, L., Pénault, A., Rauzy, S., de Montcheuil, G., and Blache, P. (2015). Discourse structure of back covers: A pilot study. In *In Proceedings of TextLink First Action Conference*, page 54.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France, April. Association for Computational Linguistics.

Yimam, S. M., Gurevych, I., de Castilho, R. E., and Biemann, C. (2013). Webanno: A flexible,web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (System Demonstrations) (ACL 2013)*, pages 1–6, Stroudsburg, PA, USA, August. Association for Computational Linguistics.

8. Language Resource References

Abeillé, A., Clément, L., and Toussenel, F. (2003). *Corpus arboré pour le français / French Treebank*. Laboratoire de Linguistique Formelle. URL: `http://www.llf.cnrs.fr/fr/Gens/Abeille/French-Treebank-fr.php`.

Candito, M. and Seddah, D. (2012). *Sequoia Treebank*. INRIA - Alpage, Sequoia Treebank, 6.0. URL: `https://www.rocq.inria.fr/alpage-wiki/tiki-index.php?page=CorpusSequoia`.

Candito, M., Crabbé, B., and Denis, P. (2010). *French Converted Dependency Treebank*. INRIA - Alpage. URL: `http://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html`.

Candito, M., Perrier, G., Guillaume, B., Ribeyre, C., Fort, K., Seddah, D., and Clergerie, E. D. L. (2014). *Deep-Sequoia*. INRIA - Alpage - LORIA, Sequoia Treebank, 7.0. URL: `https://deep-sequoia.inria.fr`.

Nivre, J., Bosco, C., Choi, J., de Marneffe, M.-C., Dozat, T., Farkas, R., Foster, J., Ginter, F., Goldberg, Y., Hajič, J., Kanerva, J., Laippala, V., Lenci, A., Lynn, T., Manning, C., McDonald, R., Missilä, A., Montemagni, S., Petrov, S., Pyysalo, S., Silveira, N., Simi, M., Smith, A., Tsarfaty, R., Vincze, V., and Zeman, D. (2015). *Universal Dependencies 1.0*. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague, Universal Dependencies, 1.0. *Handle*: hdl:11234/1-1464.

Schluter, N. and van Genabith, J. (2007). *Modified French Treebank*. National Centre for Language Technology. URL: `http://doras.dcu.ie/15265/`.