4-Couv: A New Treebank for French
Philippe Blache, Grégoire Montcheuil, Laurent Prevot, Stéphane Rauzy

To cite this version:

HAL Id: hal-01462178
https://hal.archives-ouvertes.fr/hal-01462178
Submitted on 16 Feb 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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, electroencephalography). 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 backcover that are short texts, easily available and frequently used. We present in this paper the corpus, 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 built 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.

1The French appellation of a backcover is “4ème de couverture”, leading to the treebank acronym: 4-Couv.

2Available on the ORTOLANG platform, hdl:11403/4-couv
The second step, POS-tagging, is done with the MarsaTag 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-Measure.
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 (Prevot 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 has 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 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 experiments 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.

#### 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),

<table>
<thead>
<tr>
<th>Annotation layer</th>
<th>automatic tool</th>
<th>manual (post-)edition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interestingness</td>
<td>-</td>
<td>Text selector</td>
</tr>
<tr>
<td>POS</td>
<td>MarsaTag</td>
<td>POS-tagging corrector</td>
</tr>
<tr>
<td>Syntax</td>
<td>MarsaTag</td>
<td>Tree editor</td>
</tr>
<tr>
<td>Discourse</td>
<td>?</td>
<td>??</td>
</tr>
</tbody>
</table>

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

# Figure 2: Lexical categories and features

<table>
<thead>
<tr>
<th>Category</th>
<th>features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjective</td>
<td>nature, type, gender, number, position</td>
</tr>
<tr>
<td>Adverb</td>
<td>nature, type</td>
</tr>
<tr>
<td>Connector</td>
<td>nature</td>
</tr>
<tr>
<td>Determiner</td>
<td>nature, type, person, gender, number</td>
</tr>
<tr>
<td>Interjection</td>
<td></td>
</tr>
<tr>
<td>Noun</td>
<td>nature, type, gender, number, referent type</td>
</tr>
<tr>
<td>Punctuation</td>
<td>nature</td>
</tr>
<tr>
<td>Preposition</td>
<td>type</td>
</tr>
<tr>
<td>Pronoun</td>
<td>nature, type, person, gender, number, case,</td>
</tr>
<tr>
<td></td>
<td>reflective, postposed</td>
</tr>
<tr>
<td>Verb</td>
<td>nature, modality, tense, person, gender,</td>
</tr>
<tr>
<td></td>
<td>number, auxiliary, pronominal, (im)personal,</td>
</tr>
<tr>
<td></td>
<td>direct object, indirect complement</td>
</tr>
</tbody>
</table>

# 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 2).

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 have 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.

# Figure 4: Output of the parser

# Figure 5: Annotation layers and tools
and the list of unknown words, to be manually tagged. This interface (see figure [5]) 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 enriched with scripts for adding extra information to the texts.

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 choosing a new one in the list or manually editing it.

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]).

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. 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

8. Language Resource References