4-Couv, a backcover treebank

Philippe Blache¹, Grégoire Montcheuil², Stéphane Rauzy¹
and Marie-Laure Guénot²

(1) Aix Marseille Université, CNRS,
Laboratoire Parole et Langage UMR 7309
(2) Equipex ORTOLANG, CNRS
{firstname.lastname}@lpl-aix.fr

Abstract

We present in this paper 4-Couv, a treebanking project aiming at developing a multipurpose treebank for French. The main characteristic of this project is to provide adequate material for both linguistic and psycholinguistic research. The treebank is made of short and self-contained texts, selected from a corpus of backcovers coming from different editors. Such material makes possible classical linguistic research in syntax and discourse, but also offers new perspectives in experimental linguistics: the texts being short and semantically coherent, they perfectly fit with the requirements of eye-tracking or electro-encephalographic recordings. At this stage, 4-Couv contains 3,500 trees automatically tagged and parsed, and manually corrected. Its format is compatible with other French treebanks. This paper presents the corpus, its annotation and several treebanking tools that have been developed for the different stages of its elaboration: text selection, tagging, parsing and tree edition.

1 Introduction

Treebanks, that still constitute an essential resource in linguistic description as well as natural language processing, are now faced with new uses, in particular in the perspective of experimental linguistics and psycholinguistics. We present in this paper a new treebanking project (at this stage for French), 4-Couv, aiming at answering the needs of different possible perspectives. Before describing the project, let’s underline the fact that only a few truly available treebanks exist for French, mainly the French Treebank (FTB, Abeillé et al. [11]) and its derivatives, or the French part of the Universal Dependencies Treebank¹. However, only few experiments have been done using these resources in the perspective of studying human language processing. They consist in tracking eye-movement when reading texts in

¹https://code.google.com/p/uni-dep-tb/
the perspective of evaluating difficulty models (on the basis of the number and the length of the fixations). To this day, most of the studies only take into account the morphosyntactic level, such as the works done for English (Demberg and Keller [4]) or for French (Rauzy and Blache [11]), using extracts of FTB. In these experiments however, the nature of the texts could constitute an important bias: they are taken from the newspaper Le Monde, and consist in articles describing the economic situation 20 years ago. They are then of poor interest for a “normal” reader. This problem can induce an effect of “superficial” reading, leading to an important loss of attention as well as an understanding deficit.

In the perspective of developing such new uses of treebanks, as well as enriching the amount of available data for French, we have created a new treebank based on short texts, semantically consistent and self-contained, and arousing interest so as to maintain the attention during reading.

The treebank is built from a corpus of “backcovers” called 4-Couv, answering all these needs. This project is still under development, a first release will be done by the end of 2015. It consists in a set of texts from various publishers (Pocket, Gallimard) that gave their agreement. We collected first 8,000 texts, among which 500 have been selected, representing 3,500 sentences.

We present in this article the methodology and the tools that have been developed to create 4-Couv. The first section details the nature of the texts, the characteristics of the annotation scheme and the automatic parsing. The second section outlines the tools used for the selection of texts and the revision of annotations.

2 The Corpus, its annotations

2.1 The corpus

Backcovers are small texts, containing between 80-200 tokens for 4-10 sentences, generally short (80% of sentences having at most 30 tokens, and less than 10% are longer than 40 tokens). Texts are generally (a) an extract, (b) the synopsis of the story, (c) the genesis of the book, (d) a comment about the work, or (e) a combination of two or three of this elements. Each of these short texts are semantically autonomous and – a fundamental aspect for our purpose – are supposed to keep the reading interest alive, minimizing attention and comprehension drops.

2.2 Lexical annotations

The annotation of minimal syntactic units is based on the lexicon MarsaLex[^2] that associates each form with its part of speech and morpho-syntactic features. The segmentation into tokens is maximal in that highly constrained forms are split into distinct lexical units as long as they follow syntactic composition rules. For example, constituents of semi-fixed expressions such as “il était une fois” (once upon a

[^2]: MarsaLex, hdl:11041/sldr000850
time) or “mettre à nu” (lay bare) are split, while other multiword expressions such as “d’autant plus” (all the more) or “tant mieux” (even better) are not, as they do not follow any syntactic composition.

Each lexical category has a specific features set (see figure 1), 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)).

### 2.3 Syntactic annotation

In order to maintain interoperability with the FTB (even though it could be not direct and require some processing), the treebank is constituency-based and syntactic relations are represented by means of trees. 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.
- We distinguish between lexical and phrase level: we keep unary phrases, e.g. Simone is the unique constituent of a NP in (1).

(1) “Simone m’en donne trois.” (Simone gives me three.)

<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, modality, tense, person, gender, number, auxiliary, pronominal, (im)personal, direct object, indirect complement</td>
</tr>
<tr>
<td>Verb</td>
<td>nature, modality, tense, person, gender, number, auxiliary, pronominal, (im)personal, direct object, indirect complement</td>
</tr>
</tbody>
</table>

Figure 1: Lexical categories and features
No discontinuous constituent or unbounded dependencies directly encoded, such as in (1) or (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 2) is reduced to classical phrases, at the exclusion of other constructions such as coordination (at the difference with the FTB and its derivatives).

The same types of syntactic functions than those introduced for the FTB (see figure 2) are used. This annotation is less precise than other annotation frameworks (such as Gendner et al. [5]) where structural and functional informations were given independently.

2.4 Parser

The treebank is generated with the LPL stochastic parser\textsuperscript{3} (Rauzy and Blache [10]). The processing flow follows a classical scheme. After tokenization, POS-tagging is done by means of a stochastic HMM tagger using Rabiner [9]. Finally, the stochastic parser generates the possible tree structures and selects the most probable one.

The probabilistic model for the POS tagger was trained with the GraceLPL corpus, a version of the Grace/Multi-tag corpus (Paroubek and Rajman [8]) that contains 700,000 tokens and which we correct and enrich regularly. In the model the morphosyntactic information is organized into 48 distinct tags (version 2013). On this tagset, the score (F-measure) of the tagger is 0.974.

On its side, the parser has been trained with FTLPL treebank (Blache and Rauzy [2]), a version of the MFT (Schluter and van Genabith [12]) extracted from

\textsuperscript{3}MarsaTag, hdl:11041/sldr000841
the FTB that contains at the moment 1,500 validated sentences with both constituent structure and syntactic functions (around 26,000 tokens).

3 The 4-Couv treebanking tools

3.1 Text selector

We have developed a tool helping in the texts selection, in the form of HTML files that comes to genuine autonomous wiki. This strategy to use autonomous HTML files allow to easily distribute the revision work between different experts, without needing to install any particular software (files are working directly in most of web browsers), neither to connect with a central server (that allows off-line revision). Each file containing 10 texts to evaluate, presenting the book description, the text segmented into sentences, and an evaluation form (containing check boxes and drop-down lists, see figure 3). The wiki syntax renders easy to correct errors in the sentence division (each sentence is a row in a one-column table) or separate the different parts of the text (inserting a blank line). Furthermore, it also proposes to associate information to unknown words and edit the metadata fields.

3.2 Revision tools

The correction of the automatic annotations is done in two steps. The first concerns the morphosyntactic tags and the second consists in the revision of the constituents trees produced by the parser.

The morphosyntactic correction tool (see figure 5) presents one token per line, each line containing the form, and a list of possible tags associated to the form, starting with the proposed one. Selecting a new tag consists in clicking another one from the suggested list.

The syntactic correction tool is a tree editor. Only a few of them already exist such as WordFreak (Morton and LaCivita) or TrED 2.0 (Pajas and Štěpánek). More recently, some “web-based” annotation platforms have also been created, offering an intuitive and fast annotation (brat (Stenetorp et al.) and sometimes project management facilities (for example by specifying the roles such as annotator, curator or project manager (GATE Teamware (Bontcheva et al.) or WebAnno (Yimam et al.))). However, most of these tools have been developed for dependency-based treebanks. As our approach is constituency based (requiring therefore to deal with a potentially large number of levels), we had to develop a specific editor, that could run in a single HTML (see figure 6) or be integrated into an annotation platform such as brat or WebAnno.

4 We customize a TiddlyWiki (http://classic.tiddlywiki.com/, version 2.8.1) that supply the autonomous wiki, and use a Perl script to “fill” each file with the information.

5 Only a small plugin could be required to save the modified files.
4 Conclusion

The purpose of this paper is twofold. First it aims to present a new treebank, not only proposing the classical information of this kind of resource in terms of linguistic annotation, but also answering the specific needs of experimental linguistic, in the perspective of acquiring neuro-physiological data on the basis of short and self-contained text. Secondly it also presents new treebanking tools, helping at the different stages of the process: corpus creation, pre-edition, and manual correction of the automatically generated parses. A first resource of 500 texts (3,500 trees) has been created to be distributed, together with the tools, by the end of 2015.

Acknowledgments

This work has benefit from the support of ORTOLANG (ANR-11-EQPX-0032), BLRI (ANR-11-LABX-0036) and A*MIDEX (ANR-11-IDEX-0001-02).

References


Description

«Nous croyons nos vies constituées d'événements, quand ce sont les instants d'absence, les fragments oubliés, qui les forment et les nomment. Par exemple un orgie rongé, le souvenir d'un chien, la cendre d'un regard, une odeur, un cri. L'écriture, la poésie, plongent leurs racines dans ces failles, dans les instants proscrits, ceux que la mémoire réfute.»

Quarante «vies brèves», illustres ou anonymes, de Jules César à Emily Dickinson, d’Agrippa d’Aubigné à Marina Tsvetaïeva, de Marie-Madeleine à Catulle, du Caravage à Guilhaum de Cabestanh.

Phrases

1. Nous croyons nos vies constituées d'événements, quand ce sont les instants d'absence, les fragments oubliés, qui les forment et les nomment.

2. Par exemple un orgie rongé, le souvenir d'un chien, la cendre d'un regard, une odeur, un cri.

3. L'écriture, la poésie, plongent leurs racines dans ces failles, dans les instants proscrits, ceux que la mémoire réfute.

4. Quarante «vies brèves», illustres ou anonymes, de Jules César à Emily Dickinson, d’Agrippa d’Aubigné à Marina Tsvetaïeva, de Marie-Madeleine à Catulle, du Caravage à Guilhaum de Cabestanh.

Sentences:

- non vérifié
- ok
- à revoir
- corrigé

Commentaires:

Evaluation

directives

Figure 3: Text selection
(description of Vidas/Vies volées, Christian Garcin, edited by Gallimard)
Figure 4: Editing sentences split and sections

Figure 5: Morphosyntactic tags correction

Figure 6: Syntactic tree editor