Temporal@ODIL Project: Adapting ISO-TimeML to Syntactic Treebanks for the Temporal Annotation of Spoken Speech

Jean-Yves Antoine[1], Jakub Waszczuk[2], Anaïs Lefeuvre-Halftermeyer[3], Lotfi Abouda[4], Emmanuel Schang[5], and Agata Savary[6]

[1,2,6]LI, University François Rabelais of Tours
[2,3]LIFO, University of Orléans
[4,5]LLL, University of Orléans

Abstract

This paper presents Temporal@ODIL, a project that aims at building the largest corpus annotated with temporal information on spoken French. The annotation is based on an adaptation of the ISO-TimeML standard that consists in grounding the annotation on a treebank and not on raw text.

1 Introduction

The representation and the processing of temporal information is important for most understanding tasks on linguistic data. Temporal annotation has benefitted from the normalization efforts of the ISO TC37/SC4 committee which has led to the definition of the ISO-TimeML standard (ISO 24617-1:2012), following the seminal proposal of Pustejovsky et al. (2003). While originally developed for English, ISO-TimeML has been applied on a large variety of languages (Italian, Korean, Romanian, Chinese...) with only slight idiomatic adaptations. This is a clear indication of its relevance and genericity.

Only one French corpus (French Time Bank) has been annotated following the ISO-TimeML standard (Bittar et al., 2011). It was built on an extract of the French Tree Bank, FTB (Abeillé et al., 2003). Its size is reasonable (15,876 words) for pilot linguistics studies but is too restricted for computational purposes. In addition, the syntactic information that is present in the FTB was not considered during the annotation phase.

In this paper, we present Temporal@ODIL, a project which aims precisely at enlarging and deepening the seminal work conducted with the French Time Bank in two directions :

- The temporal annotation is not conducted on written text but on speech transcripts. Several language registers are considered, ranging from socio-linguistic interviews to highly interactive dialogues. The annotation is conducted on ANCOR (Muzerelle et al., 2014), the largest French coreference corpus and one of the largest ones of spoken language. Temporal@ODIL provides a complementary annotation layer, allowing studies combining coreference and temporal data. It will concern 20,000 words, which doubles the size of existing French TimeML-based resources.

- Temporal@ODIL proposes modifications to the ISO-TimeML standard and its main originality is to delimit temporal mentions not by their minimal chunk but by the range of the syntactic subtree that covers the temporal mention. In order to favor re-usability, we watch out carefully to maintain upward compatibility between the ISO standard and our annotation scheme.

The second section presents and substantiates the modifications we propose to the standard. The third section describes the semi-automatic treebank annotation procedure we are following. The discussion between dependency and constituency parsing is presented there. Finally, the fourth section introduces the annotation tool that has been developed for the project. This tool is not restricted to temporal annotation, it can also be used for treebank edition and correction.

2 Annotation scheme : modifications of the ISO-TimeML standard

The changes that are proposed in Temporal@ODIL involve some extensions that do not imply any modification of the structure of the XML TimeML documents. They are indeed limited to the definition of values that instantiate some attributes of the norm and preserve a structural compliance with the norm. These changes are largely detailed in (Lefeuvre-Halftermeyer et al., 2016). For the sake of clarity, we recall however briefly our main proposals[1].

- TIMEX: temporal functions. The attribute temporalFunction expresses whether the temporal reference of a time expression needs to be calculated considering its linguistic expression or an other reference. Instead of defining temporalFunction as binary value, three values based on the seminal work of Reichenbach (1947) are considered in the Temporal@ODIL project: Null (absolute references), S (enunciation-based relative references) and R (discourse-based relative ones). The details of the function (temporalFunctionID) themselves have never been described by the norm. Temporal@ODIL adopts a typology of function classes defined in Drat (2014).

- TLINK relation. TLINK types available in the norm are the ones identified by Allen (1983), in addition to a fourteenth one, IDENTITY. Like other authors of (MERLOT, 2016), we have decided to ignore this type which is a proper coreference relation, and to add a few more relations (Drat, 2014).

The last proposal of modification is deeper: it involves a move from a word-based to a tree-based annotation. ISO TimeML guidelines request <EVENT> tags to be delimited by their minimal event-denoting chunks. This restrictive delimitation was questioned by Pustejovsky et al. (2006). The Temporal@ODIL project follows a broader annotation that covers the whole event-denoting expression, in order to keep all the relevant information that is useful for temporal reasoning without asking the annotator to resolve the syntactic structure in addition to the semantic annotation.

A first interest of a broader annotation is obvious with temporal abstract anaphora, whose resolution often needs the consideration of a whole clause (Zinsmeister and Dipper, 2011). This large-span annotation questions the ability for the annotators to delimit the eventualities with a satisfactory reliability. To ease this delimitation, we adopt a solution that was investigated for multi-word expressions in the Prague Dependency Treebank (Bejček and Straňák, 2010): eventualities are defined on the syntactic structures of a treebank. Then, the delimitation task boils down to the selection of one specific node. Data reliability is favored by a reduction of the annotator's cognitive load. For this, we propose to only annotate the two syntactic nodes implied in a supposed SLINK (in a constituency paradigm) and to resolve the SLINK automatically by overloading the syntactic link. We conducted pilot experiments that showed that a phrase-structure treebank is required for temporal annotation: the annotators indeed encountered difficulties to characterize the span of time-denoting items on dependency trees. This tree-based annotation does not violate the XML structure of the ISO-TimeML annotation files: the span of the eventualities in the standoff annotation is simply based on tree nodes identifiers rather than on token identifiers.

3 From a syntactic annotation towards a semantic one

Temporal@ODIL adopts an incremental process of annotation: the annotation is divided in several successive stages that combine automatic and manual procedures. The annotation is composed of 5 phases:

- automatic pre-processing of the corpus: the first stage consists in a preprocessing of the speech transcripts to ease the parsing. We proceed in the sidelining of noises interjections, and phatic expressions not carrying patent temporal information: "*oui*" for '*yes*', "*bonjour*" for '*good morning*', etc., whereas verbal expressions like "*excusez-moi*" for '*excuse-me*', "*s'il vous plait*" for '*please*' are kept[2].

- automatic syntactic annotation: this phase consists of using a parser to build the constituency-based treebank on which the temporal annotation will be conducted. Two strategies have been considered: using a robust dependency parser and building afterwards the constituency treebank from the

[1] Our proposals have been updated recently on some points: these changes are integrated in this paper.

[2] We expect the studies lead during this project to help us confirm or disprove this choice.

parsing dependency trees; using directly a constituency-based parser, provided it presents a satisfactory behaviour. Temporal@ODIL focuses on spontaneous speech transcripts, which challenge the robustness of the syntactic analysis. We conducted pilot experiments that show dependency parsers do not outperform noticeably constituency ones on spoken French. We thereby decided to use the Stanford parser trained on the FTB. All the analyses are lead on speech turns separately, further boostrapping of the parser on speech data will be also considered.

Figure 1 illustrates this parsing output for the sentence :

(1) *oui bonjour madame j'aurais voulu parler à madame Nom mais je crois que sa ligne directe ne répond pas*
'yes good morning Madam I would like to talk to madam Name but her direct phone line does not answer'

One should note that the phatic expression "*oui bonjour madame*" ('*yes good morning Madam*') has been automatically put aside in the pre-processing phase: it does not appear in the syntactic structure whereas the whole dialogue (before pre-processing) is available on the right side of the interface.

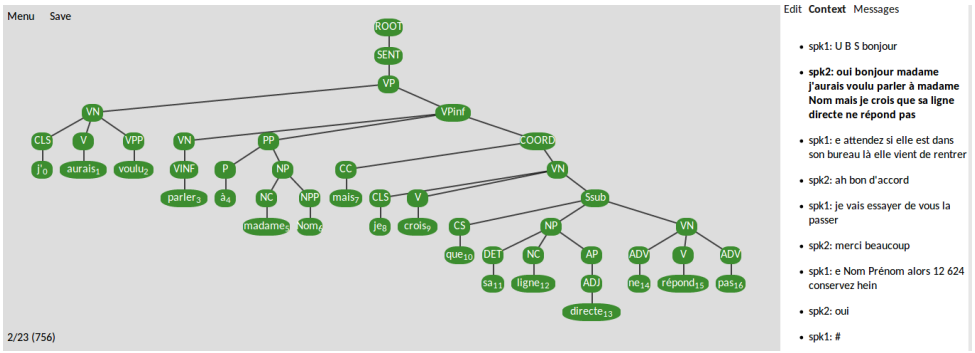

Figure 1: Output of the second phase: Standford parse tree on the pre-processed sentence (1).

- revision of the syntactic annotations: this phase is required to correct parsing errors but also to reach additional purposes. The parsing trees have to be adapted in order to allow the representation of speech disfluences and to obtain deeper constituency trees.

First, annotators are asked to put aside all the interjections and noises not dealt with in the first phase[3], and speech repairs are annotated following the Rhapsodie project guidelines (Lacheret et al., 2014). Annotators are asked to readdress the right POS if wrongly labeled, then a new analysis can be invoked respecting the new declared POS if needed.

FTB annotation scheme leads to rather flat syntactic trees that do not fulfill all the need of our temporal annotation. This phase aims at correcting this problem: it combines the contextual activation of automatic deepening rules and a fully manual revision. These deepening rules consist of:

- a VN followed by a VPInf under a node SENT, VPInf or SRel are grouped into a single VP in the same context, as one can see in figure 2 for "*j'aurais voulu parler à madame Nom*".
- a VN followed by a NP under a node SENT, SSub, COORD or Sint are grouped into a VP.
- a VN followed by a PP under a node VPinf or SSub are grouped into a VP.
- a VN followed by an ADV under a node SENT are grouped into a VP.

- manual eventualities annotation: the detection of the linguistic items that denote eventualities or signals is achieved completely through a manual procedure. This annotation is conducted directly on the treebank built during the two previous phases.

- manual temporal relations annotation: the relations between eventualities are also defined manually. Every relation is characterized manually by several temporal features[4].

[3]The annotator visualizes the whole dialogue containing the disfluences as one can see in the right part of the figure 1. Noises and interjections are kept in the context even if put aside for syntactical purposes.

[4]We hope to detect automatically some semantic relations carried by syntactic links like SLINKs. In the figure 3, the two red subgraphs are good candidates for carrying a SLINK: VP-VPINF-VP and VP-SSUB-VP added to the annotation of the first

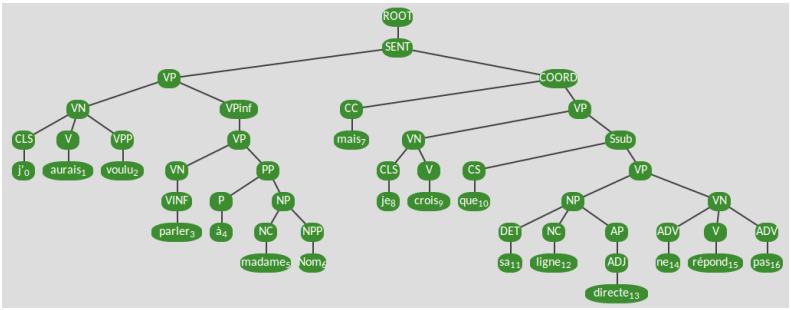

Figure 2: Output of the third phase: correction process and application of the automated rules.

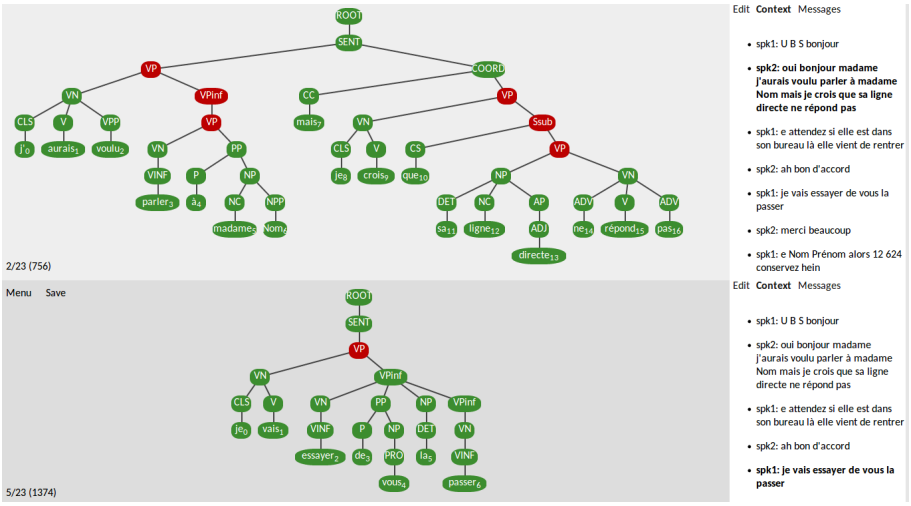

Figure 3: Full interface of the tool and first hint for locating interesting SLINK schemes.

4 Annotation tool

All of the manual annotation phases are conducted using a single annotation tool, which has been specifically implemented for the purpose of the project.

The tool's workspace consists of two vertically arranged annotation windows, showing two syntactic trees assigned to two (typically different) speech turns in a given file. As mentioned in the previous section, one of the primary functionalities provided by the tool is to allow correction of syntactic trees. To this end, the tool allows annotators to perform several structure-modifying operations: adding and deleting nodes, changing the parent of a node, changing the position of the node w.r.t. its parent, etc. Only the operations which preserve the well-formedeness of syntactic structures are allowed.

Duplication of the annotation workspace facilitates, among others, viewing and editing the temporal relations occurring between different trees. Such relations can be created by selecting the corresponding nodes and using an appropriate keyboard command. The newly created events and temporal relations are supplied with default ISO-TimeML-related attribute values, which can be subsequently changed manually in the side windows. We plan to later experiment with a semi-automatic annotation of the attribute values, where the corresponding machine-learning annotation model is being gradually bootstrapped from the already annotated part of the corpus.

The tool is implemented in a client/server architecture. The frontend annotation tool is written in Elm (`http://elm-lang.org/`), which compiles to JavaScript, thus the tool can be used in any modern internet browser. The client annotation tool communicates with a Haskell server via websockets, with

VNs under the first VPs as I_STATEs should raise SLINKs between "*j'aurais voulu*" ('*I would like*') and "*parler à madame Nom*" ('*to talk to Madam Name*') and between "*je crois*" ('*I think*') and "*sa ligne directe ne répond plus*" ('*her direct phone line does not answer*') .

the annotated files serialized to JSON before being sent. On both sides, annotation data is represented with appropriate data types, which guarantees, among others, that malformed data is never sent to the server to be stored in the database.

Such an architecture has a couple of advantages. The annotator does not have to install anything locally, and the server can provide the user with more advanced functionality. The server can be requested to syntactically re-analyze a given sentence in a way which takes the constraints specified directly by the annotator (e.g. a particular tokenization) into account. In the long run, the client/server architecture should also allow a more collaborative annotation style.

The Temporal@ODIL project will end in spring 2018. The resulting corpus, providing a 100,000 words syntactic annotation layer, and a 20,000 words temporal annotation layer, will be freely available from June 2018 under Creative Commons CC-BY-SA license[5].

References

Abeillé, A., L. Clément, and F. Toussenel (2003). Building a treebank for french. *Treebanks*, 165–187.

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM 26*.

Bejček, E. and P. Straňák (2010, Apr). Annotation of multiword expressions in the prague dependency treebank. *Language Resources and Evaluation 44*(1), 7–21.

Bittar, A., P. Amsili, and P. Denis (2011). French TimeBank : un corpus de référence sur la temporalité en français. In *TALN 2011*, Montpellier, France, pp. 259–270.

Drat, L. (2014). Projet TourInFlux. Annotation des expressions temporelles. Master's thesis.

ISO (2012). Language resource management - Semantic annotation framework (SemAF) Part 1: Time and events. ISO 24617-1:2012, International Organization for Standardization, Geneva, Switzerland.

Lacheret, A., S. Kahane, J. Beliao, A. Dister, K. Gerdes, J.-P. Goldman, N. Obin, P. Pietrandrea, and A. Tchobanov (2014, July). Rhapsodie: un Treebank annoté pour l'étude de l'interface syntaxe-prosodie en français parlé. In *4e Congrès Mondial de Linguistique Française*, Volume 8, Berlin, Germany, pp. 2675–2689.

Lefeuvre-Halftermeyer, A., J.-Y. Antoine, A. Couillault, E. Schang, L. Abouda, A. Savary, D. Maurel, I. Eshkol-Taravella, and D. Battistelli (2016). Covering various Needs in Temporal Annotation: a Proposal of Extension of ISO TimeML that Preserves Upward Compatibility. In *LREC 2016*, Portorož, Slovenia.

MERLOT (2016). Annotation scheme for the merlot french clinical corpus. Technical report.

Muzerelle, J., A. Lefeuvre, E. Schang, J.-Y. Antoine, A. Pelletier, D. Maurel, I. Eshkol, and J. Villaneau (2014). ANCOR_Centre, a Large Free Spoken French Coreference Corpus: description of the Resource and Reliability Measures. In *LREC 2014*, Reyjavik, Iceland.

Pustejovsky, J., J. M. Castao, R. Ingria, R. Sauri, R. J. Gaizauskas, A. Setzer, G. Katz, and D. R. Radev (2003). Timeml: Robust specification of event and temporal expressions in text. In M. T. Maybury (Ed.), *New Directions in Question Answering*, pp. 28–34. AAAI Press.

Pustejovsky, J., J. Littman, and R. Saurí (2006). Argument structure in timeml. In *Dagstuhl Seminar Proceedings*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Reichenbach, H. (1947). *Elements of Symbolic Logic*. New York, Macmillan.

Zinsmeister, H. and S. Dipper (2011). *Towards a Standard for Annotating Abstract Anaphora*.

[5]This licence is inherited from the original oral corpus.