Combiner parseur automatique et révision manuelle pour la constitution d'un corpus arboré de parole spontanée : retour d'expérience sur le corpus ODIL_syntaxe

Ilaine Wang, Aurore Pelletier, Jakub Waszczuk, Anais Anais Lefeuvre-Halftermeyer, Jean-Yves Antoine, Lotfi Abouda, Emmanuel Schang, Agata Savary

Combiner parseur automatique et révision manuelle pour la constitution d'un corpus arboré de parole spontanée : retour d'expérience sur le corpus ODIL_syntaxe

Ilaine Wang[1, 2] Aurore Pelletier[2] Jakub Waszczuk[3] Anaïs Halftermeyer[1]
Jean-Yves Antoine[2] Lotfi Abouda[4] Emmanuel Schang[4] Agata Savary[2]
(1) Laboratoire LIFO, U. Orléans, 45000 Orléans, France
(2) Laboratoire LIFAT, U. Tours, 37000 Blois, France
(3) ISI, Heinrich-Heine-Universität Düsseldorf, 40225 Deutschland
(4) LLL, CNRS U. Orléans, 45000 Orléans, France
ilaine.wang,anais.halftermeyer@univ-orleans.fr,
jakub.waszczuk@phil.uni-duesseldorf.de,jean-yves.antoine@univ-tours.fr

RÉSUMÉ

Cet article présente l'utilisation d'une plateforme d'annotation syntaxique (*Contemplata*) qui intègre un parseur pour annoter automatiquement des corpus écrits ou oraux puis permettre leur révision manuelle par un·e expert·e, afin de limiter son travail d'annotation. Dans le cadre du projet ODIL, cet outil a permis de réaliser un corpus de français parlé spontané annoté en arbres de constituants, ceci dans la perspective d'une annotation en temporalité. Nous présentons ici la démarche mise en œuvre pour l'annotation ainsi que les conventions, et proposerons une démonstration de l'outil.

ABSTRACT

Combining Automatic Parsing and Manual Revision for the Constitution of a Spontaneous Speech Treebank : Experience Feedback on the ODIL_Syntaxe Corpus.

This paper describes a syntactic annotation platform (*Contemplata*) that integrates a parser (*Stanford Parser* precisely) to automatically annotate written text or oral transcriptions and then allows their manual revision by an expert, in order to ease the annotation task. This tool was used in the ODIL Project to produce a phrase-structure treebank based on a corpus of spontaneous speech. In this paper, we present the annotation process that has been implemented as well as our annotation guidelines and plan to provide a demonstration of the annotation tool during the presentation.

MOTS-CLÉS : annotation syntaxique semi-automatique, corpus arboré, arbres de constituants, outil d'annotation, parseur, français parlé spontané.

KEYWORDS: syntactic annotation, treebank, phrase-structure representation, annotation tool, parser, spontaneous spoken French.

1 Projet ODIL : annotation en temporalité

Le travail qui sera présenté avec ce poster a été réalisé dans le cadre du lot « annotation temporelle » du projet ODIL, financé par la région Centre Val de Loire. Le sous-projet Temporal@ODIL vise la réalisation d'un corpus oral annoté en temporalité (identification des éventualités et caractérisation

des relations temporelles entre éventualités) qui n'a d'équivalent pour le français qu'une seule ressource : le *French TimeBank* (Bittar et al., 2011). L'originalité de nos travaux est précisément de se focaliser, contrairement à ce dernier, sur l'oral spontané. Une première partie du projet a consisté à étudier l'adaptation au langage oral de la norme d'annotation de la temporalité ISO TimeML (Pustejovsky et al., 2003). Cette extension (Antoine et al., 2017) de la norme consiste en particulier à décrire la temporalité non plus au niveau de la tête lexicale des éventualités, mais au niveau des nœuds d'une représentation arborée des énoncés. La considération de la structure syntaxique des énoncés autorise en effet une représentation parfois nécessaire de l'empan des éventualités, tout en facilitant la tâche de l'annotateur. Par ailleurs, une ambiguité syntaxique peut nuire à l'annotation sémantique, une fois séparés les niveaux d'annotation, la charge cognitive de l'annotateur s'en trouve considérablement réduite. Les éventualités correspondent en effet toujours à un nœud de l'arbre syntaxique, dès lors que celui-ci relève d'une représentation en constituants et non en dépendances.

Ce choix nous a donc conduit à élaborer au préalable un corpus arboré en constituants. Cet article présente précisément notre démarche d'annotation syntaxique. Un des intérêts de notre démarche est de conduire une annotation semi-automatique : nous utilisons en effet un analyseur syntaxique reconnu, le *Stanford Parser* (Green et al., 2011), pour proposer une première annotation qui sera ensuite révisée manuellement. Pour profiter au mieux de cette stratégie, nous avons apporté un soin très particulier à l'utilisabilité du nouvel outil d'annotation que nous avons développé pour le projet : *Contemplata*. Il s'agit d'un outil générique qui permet d'annoter tout corpus en arbres de constituants mais aussi de décorer les nœuds de cet arbre et d'y ajouter des relations de natures variées. Cet article présente un retour d'expérience sur l'utilisation de l'outil dans le cadre d'ODIL.

2 Corpus ODIL_Syntaxe

Le projet ODIL fournira, à sa clôture, un corpus de français parlé de 12355 mots qui comportera (1) une annotation en arbres de constituants et (2) une annotation en relations temporelles. Le corpus arboré, *ODIL_Syntaxe*, est désormais finalisé et sera distribué sous licence libre d'ici à fin 2019. Il repose sur l'enrichissement de trois corpus bruts de transcriptions de l'oral correspondant à trois registres différents : les corpus ESLO (entretiens sociolinguistiques issus de Eshkol Taravella et al. (2011)), OTG (dialogue oral en présentiel) et Accueil_UBS (dialogue oral par téléphone). Ces deux dernières ressources ont déjà servi à la réalisation d'ANCOR_Centre, un très grand corpus de français parlé annoté en coréférence (Muzerelle et al., 2014).

La conception de corpus annotés en arbres syntaxiques a pris un nouvel essor depuis le début du millénaire. De telles ressources sont cruciales pour l'apprentissage de systèmes de traitement automatique des langues dans des applications nécessitant la prise en compte de la structure syntaxique des énoncés. Elles constituent également un préalable pour les études linguistiques en corpus opérant à ce niveau de description linguistique (coréférence, saillance discursive, relations temporelles…). Le français dispose déjà de ressources importantes avec le *French Treebank* (Abeillé et al., 2003) et *SEQUOIA* (Candito and Seddah, 2012) mais leur confection s'est faite à partir de textes écrits et les résultats ne sont pas transposables sur les réalisations orales. *SEQUOIA* propose par ailleurs une annotation en dépendances et non pas en constituants. A un moment où l'ingénierie des langues et la linguistique de corpus sont confrontées de manière croissante à des contenus de parole,

ce corpus arboré se propose de remédier à cette lacune. Un seul corpus peut être comparé au nôtre : *Rhapsodie* (Lacheret et al., 2014), dont l'accès pour la version en constituance est toutefois limité.

Les conventions d'annotation du corpus reposent principalement sur celles du *French TreeBank* telles qu'utilisées pour l'apprentissage du *Stanford Parser*. Nous avons toutefois dû ajouter certaines règles pour les adapter à l'oral spontané. Ces adaptations ont essentiellement concerné la représentation des disfluences orales, en particulier les inachèvements et toutes les formes d'entassement paradigmatique (reprises, répétitions, autocorrections). Pour ces adaptations, nous avons cherché à rester aussi proches que possible des choix d'annotation de Rhapsodie, qui nous sont apparues des plus pertinentes.

3 Contemplata : outil d'annotation syntaxique semi-manuelle

Contemplata se présente sous la forme d'une application Web dont la partie client est développée en langage Elm puis recompilé en JavaScript et utilisable dans tout navigateur Web sans aucune installation spécifique. Le client communique avec un serveur développé en Haskell, qui permet en particulier l'interrogation du *Stanford Parser* ou tout autre outil équivalent respectant les formats d'entrée et de sortie des données manipulées. Au final, l'application Web permet de gérer l'intégralité de la réalisation d'un corpus arboré, en passant de la supervision du projet d'annotation (rôle administrateur) à l'annotation en elle-même (rôle d'annotateur ou d'adjudicateur). En pratique, l'annotation suit une succession d'étapes que l'annotateur doit respecter, avant tout pour lui simplifier la tâche. Cette décomposition en sous-tâches réduit en effet la charge cognitive de l'expert·e, et favorise la qualité de l'annotation. L'annotation respecte la succession suivante :

- Analyse syntaxique automatique par le *Stanford Parser*, auquel nous avons ajouté une étape de pré-traitement qui met à l'écart (sans les éliminer) les formules d'introduction et les phatiques, qui sont très présents en oral spontané tout en ne portant pas de contenu propositionnel.
- Correction manuelle de l'annotation automatique en parties du discours (POS) et élimination des phatiques résiduels encore présents.
- Relance de l'analyse syntaxique automatique après correction des POS. Cette première correction permet en effet souvent au parseur de corriger l'ensemble de son analyse.
- Segmentation manuelle des tours de paroles en plusieurs énoncés. Les longs tours de parole peuvent induire en erreur le parseur. C'est notamment le cas des registres de dialogue du type « entretien » (ESLO). On découpe donc les tours de parole en énoncés successifs cohérents.
- Relance de l'analyse syntaxique automatique, le découpage en énoncé pouvant, une fois encore, améliorer la qualité des arbres obtenus automatiquement.
- Correction manuelle des arbres résultants. Il s'agit de l'opération la plus coûteuse humainement, mais les étapes précédentes ont amélioré la qualité de l'annotation automatique, ce qui réduit la charge de l'expert·e. *Contemplata* permet toutes les modifications possibles des arbres syntaxiques : ajout, déplacement et suppression de nœuds, déplacement de sous-arbres, modification de l'étiquette d'un nœud etc. (Figure 1). Cette étape gère également l'annotation manuelle des disfluences orales, qui ne sont pas prises en compte par le parseur, puisque celui-ci a été entraîné sur un corpus d'écrit.

La réalisation du corpus *ODIL_Syntaxe* a montré que l'annotation est grandement facilitée par *Contemplata*, ce qui nous a permis d'annoter un corpus de taille déjà raisonnable (comparable au *French TimeBank*) à un coût humain relativement limité. Notons enfin que *Contemplata* permet une

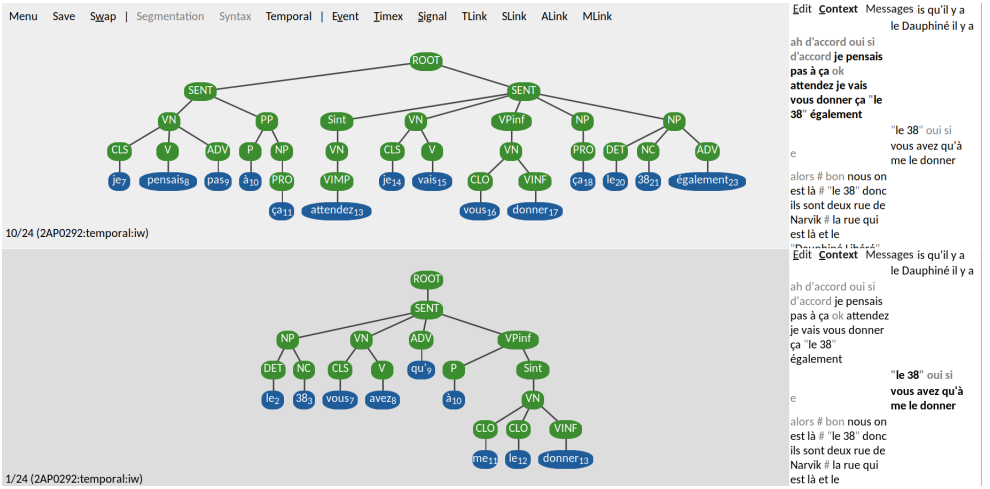

FIGURE 1 – Interface de *Contemplata* affichant en simultané la structure syntaxique en constituant de deux tours de parole (ROOT) comportant ici un ou deux énoncés (SENT)

révision experte de l'annotation obtenue par le biais d'une aide à l'adjudication d'annotations (comparaison graphique de 2 annotations concurrentes sur un même extrait de corpus).

Références

Abeillé, A., Clément, L., and Toussenel, F. (2003). Building a treebank for French. In *Treebanks*.

Antoine, J.-Y., Waszczuk, J., Lefeuvre Halftermeyer, A., Abouda, L., Schang, E., and Savary, A. (2017). Temporal@ ODIL Project : Adapting ISO-TimeML to Syntactic Treebanks for the Temporal Annotation of Spoken Speech. In *Proceedings isa-13 (IWCS'2017)*.

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

Candito, M. and Seddah, D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Actes de TALN'2011*, pages 321–334.

Eshkol Taravella, I., Baude, O., Maurel, D., Hriba, L., Dugua, C., and Tellier, I. (2011). Un grand corpus oral « disponible » : le corpus d'Orléans 1 1968-2012. *TAL*, 52(3) :17–46.

Green, S., De Marneffe, M.-C., Bauer, J., and Manning, C. D. (2011). Multiword expression identification with tree substitution grammars : A parsing tour de force with french. In *Proceedings EMNLP'2011*.

Lacheret, A., Kahane, S., Beliao, J., Dister, A., Gerdes, K., Goldman, J.-P., Obin, N., Pietrandrea, P., and Tchobanov, A. (2014). Rhapsodie : un Treebank annoté pour l'étude de l'interface syntaxe-prosodie en français parlé. In *Actes de CMLF'2014*, pages 2675–2689.

Muzerelle, J., Lefeuvre, A., Schang, E., Antoine, J.-Y., Pelletier, A., Maurel, D., Eshkol, I., and Villaneau, J. (2014). ANCOR_Centre, a large free spoken French coreference corpus : description of the resource and reliability measures. In *Proceedings LREC'2014*.

Pustejovsky, J., Castano, J. M., Ingria, R., Sauri, R., Gaizauskas, R. J., Setzer, A., Katz, G., and Radev, D. R. (2003). TimeML : Robust specification of event and temporal expressions in text. *New directions in question answering*, 3 :28–34.