ZOMBILINGO : manger des têtes pour annoter en syntaxe de dépendances

Karën Fort[1] Bruno Guillaume [2] Valentin Stern [1]
(1) LORIA, Université de Lorraine
(2) LORIA, Inria Nancy Grand-Est
karen.fort@loria.fr, bruno.guillaume@loria.fr, valentin.stern@loria.fr

Résumé. Cet article présente ZOMBILINGO un jeu ayant un but (*Game with a purpose*) permettant d'annoter des corpus en syntaxe de dépendances. Les annotations créées sont librement disponibles sur le site du jeu.

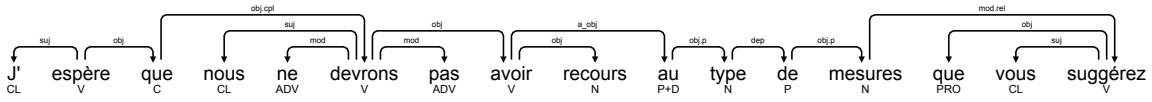
Abstract. This paper presents ZOMBILINGO, a Game With A Purpose (GWAP) that allows for the dependency syntax annotation of French corpora. The created resource is freely available on the game Web site.

Mots-clés : jeux ayant un but, complexité, annotation, syntaxe en dépendances.

Keywords: GWAP, complexity, annotation, dependency syntax.

La production de ressources linguistiques de grande taille est très coûteuse, en particulier en main d'œuvre. Ainsi, le coût d'annotation du Prague Dependency Treebank a été estimé à 600 000 dollars (Böhmová *et al.*, 2001). Une alternative pour produire des ressources est l'utilisation de la myriadisation (*crowdsourcing*), c'est-à-dire le recours à la « foule pour réaliser une tâche. Les jeux ayant un but, par exemple, ont été utilisés pour différentes tâches en TAL : JEUXDEMOTS [1] (Lafourcade, 2007) a pour but de créer un réseau lexical ; PHRASE DETECTIVES [2] (Chamberlain *et al.*, 2008) fait annoter un corpus en anaphores. Ces deux jeux ont eu un succès considérable et ont permis de créer des ressources de qualité raisonnable pour un coût réduit. Le premier fait appel au sens commun et le deuxième à des connaissances scolaires. Dans d'autres domaines, il a été possible d'utiliser un jeu pour des tâches nettement plus complexes et qui nécessitent une formation des personnes qui participent. Ainsi, dans FOLDIT (Cooper *et al.*, 2010) les joueurs doivent manipuler des représentations 3D de protéines pour étudier la façon dont elle peuvent interagir. ZOMBILINGO est inspiré de ces succès et a pour but de faire réaliser à des joueurs une tâche de TAL réputée complexe : annoter des dépendances syntaxiques.

Les données que nous souhaitons produire sont des analyses en dépendances syntaxiques compatibles avec celles utilisées pour le corpus Sequoia (Candito & Seddah, 2012). Elles sont illustrées par l'exemple ci-dessous.



Ce choix nous permet d'utiliser le corpus Sequoia comme amorce pour ZOMBILINGO, notamment pour la phase de formation des joueurs. Le système sera ensuite alimenté par des phrases issues de textes libres de droits, qui seront pré-annotés à l'aide d'analyseurs syntaxiques. Quand une nouvelle phrase est ajoutée dans la base de données, sa pré-annotation est considérée comme correcte ; dans la suite du jeu, si suffisamment de joueurs donnent un avis contraire à la pré-annotation, l'annotation de la phrase considérée est modifiée pour en tenir compte. Il est donc possible à tout moment de faire une extraction de la ressource annotée en syntaxe, qui tient compte de ce que tous les joueurs ont fait précédemment.

L'un des enjeux essentiels de ce jeu est d'être capable de gérer la complexité de la tâche. Il n'est bien entendu pas possible de demander à un joueur de produire l'annotation d'une phrase complète ; il faut décomposer la tâche globale en une série de tâches plus élémentaires qui peuvent être confiées à des joueurs sans les décourager. Dans ZOMBILINGO, cette gestion

1. Voir : `http://www.jeuxdemots.org`.
2. Voir : `http://anawiki.essex.ac.uk/phrasedetectives`.

de la complexité s'appuie sur le découpage de la tâche suivant les différents phénomènes linguistiques présents dans la phrase. Ce découpage permet également de mettre en place des séances de formations pour chacun des phénomènes et donc de ne pas surcharger les joueurs d'informations : le joueur choisit un phénomène, suit la formation correspondante, et peut ensuite commencer à jouer avec ce phénomène.

Un autre élément essentiel à la réussite de ZOMBILINGO est la motivation des joueurs. En effet, la production d'une ressource de grande ampleur de qualité n'est possible que si beaucoup de joueurs utilisent le jeu et si une proportion raisonnable d'entre eux restent longtemps et reviennent régulièrement jouer. Pour attirer les joueurs, le design est un élément essentiel. Nous avons choisi le thème des zombies parce qu'il est fédérateur dans le monde du jeu et par clin d'œil à la notion de tête d'une dépendance linguistique : annoter c'est « manger des têtes », c'est donc une tâche pour les zombies ! La capture d'écran ci-dessous présente l'interface du jeu.

1. profil du joueur
2. progression de la partie
3. aide interactive
4a. mot joué
4b. relation ou phénomène à annoter
4c. « main » pour le choix de la réponse
5. accès aux objets du jeu

Les mécanismes qui encouragent les joueurs à jouer suffisamment longtemps et à revenir régulièrement sont aussi un élément clé de la réussite du jeu. En se basant sur les notions souvent utilisées pour les jeux (sérieux ou non), nous avons prévu différents mécanismes qui correspondent aux différents types de joueurs existants. Ainsi, les mécanismes que nous avons mis en place ont pour but de répondre aux attentes des quatre types de joueurs identifiés par Bartle (1996) : *killers*, *achievers*, *explorers* et *socializers*.

Les données produites par les joueurs permettront de produire un corpus annoté en dépendances de surface qui sera mis à jour en continu en fonction des actions des joueurs. Ce corpus sera mis à disposition librement.

Les auteurs tiennent à remercier Hadrien Chastant pour la première maquette, Charles Ancé pour ses magnifique dessins, Alice Guyot pour les éléments de design et Mathieu Lafourcade pour son aide dans la conception du jeu.

Références

BARTLE R. (1996). Hearts, clubs, diamonds, spades : Players who suit MUDs. *The Journal of Virtual Environments*.

BÖHMOVÁ A., HAJIČ J., HAJIČOVÁ E. & HLADKÁ B. (2001). The prague dependency treebank : Three-level annotation scenario. In A. ABEILLÉ, Ed., *Treebanks : Building and Using Syntactically Annotated Corpora*. Kluwer Academic Publishers.

CANDITO M. & SEDDAH D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Traitement Automatique des Langues Naturelles (TALN)*, Grenoble, France.

CHAMBERLAIN J., POESIO M. & KRUSCHWITZ U. (2008). Phrase Detectives : a web-based collaborative annotation game. In *Proceedings of the International Conference on Semantic Systems (I-Semantics'08)*.

COOPER S., TREUILLE A., BARBERO J., LEAVER-FAY A., TUITE K., KHATIB F., SNYDER A. C., BEENEN M., SALESIN D., BAKER D. & POPOVIĆ Z. (2010). The challenge of designing scientific discovery games. In *Proceedings of the Fifth International Conference on the Foundations of Digital Games*, FDG '10, p. 40–47.

LAFOURCADE M. (2007). Making people play for lexical acquisition. In *Proceedings of the 7th Symposium on Natural Language Processing (SNLP 2007)*.