Creating Zombilingo, a Game With A Purpose for dependency syntax annotation

Karen Fort, Bruno Guillaume, Hadrien Chastant

Creating *Zombilingo*, a Game With A Purpose for dependency syntax annotation

Karën Fort
Université de Lorraine / LORIA
Campus scientifique, BP 239
54506 Vandœuvre-lès-Nancy, France
karen.fort@loria.fr

Bruno Guillaume
Inria NGE / LORIA
Campus scientifique, BP 239
54506 Vandœuvre-lès-Nancy, France
bruno.guillaume@loria.fr

Hadrien Chastant
Université de Lorraine
Campus Artem, CS 14 234
54042 Nancy, France
hadrien.chastant@gmail.com

ABSTRACT

This paper presents the design of ZOMBILINGO, a Game With A Purpose (GWAP) that allows for the dependency syntax annotation of French corpora. The development will start mid-2014 and the game is to be made available by the end of the year. The created language resource will be freely and continuously available on the game Web site.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*

Keywords

GWAP, complexity, annotation, dependency syntax

1. INTRODUCTION

The availability of syntactic resources is known to be important to develop and to evaluate statistically based parsers; and [14] shows that using a sentence dependency structure can improve the information retrieval task, as compared to the basic "bag of words" techniques. Although French is far from being an under-resourced language, it lacks *freely available* language resources for natural language processing, in particular large syntactically annotated corpora. Crowdsourcing, in particular using a Game With A Purpose (GWAP), is a promising and ethically sound[1] alternative to the traditional, high-cost human annotation.[2]

Indeed, GWAPs like ESP GAME [19], which consisted in tagging images according to their content, or JEUXDEMOTS[3] [13], which relies on the players associating ideas according to a simple instruction, allowed to create large language resources of reasonable quality at a relatively low cost [4]. However, the complexity of the tasks to be performed in these games is limited, and they basically rely on the general knowledge of the world that the players have.

In PHRASE DETECTIVES[4] [6], the task – identifying the referent in an anaphora relation – is less simple and the creators of the game had to find ways to help players perform correctly: they have to go through a training period before they can really play, and the corpus is pre-annotated to facilitate the task.[5] The quality obtained for this task is satisfactory [5]. However, another part of the game, which concerned the identification of properties, proved to be too complex and gave unsatisfactory results [4].

To our knowledge, only one GWAP has been developed, which addressed the syntactic annotation task, using a Tetris-like game, PHRATRIS[6], but, although it won the INSEMTIVES Game Challenge[7], the project seems to be abandoned and we could not find any published results. Another related project – though not a game – was designed to target Czech school children [10] and have them annotate sentences with morphology and dependency-based syntax using an appropriate editor, without any specific training. The authors obtained mixed results, with relatively low agreement between the teachers, and high agreement between the children. Without any further publication or information on the project, one cannot drive any significant conclusion.

Despite these somewhat discouraging results, we decided to develop a GWAP to perform the annotation of corpora with dependency syntax. Our role models in this challenge are PHRASE DETECTIVES and FOLDIT [8, 11], another GWAP, which success was largely advertised. In this molecular biology game, the players are asked to fold proteins according to – sometimes complex – instructions.

We first present, in section 2, the resource we want to build and the annotation reference we created. In section 3, we give details about the solutions we propose to address the issue of complexity. Finally, we describe in section 4 some of the mechanisms that will be used to motivate the players.

[1] As opposed to Amazon Mechanical Turk, for example [9].
[2] For example, the Prague Treebank annotation cost $600,000 [1] for 1.8 million tokens.
[3] See: `http://www.jeuxdemots.org`.

GamifIR'14, April 13 2014, Amsterdam, The Netherlands.

[4] See: `http://anawiki.essex.ac.uk/phrasedetectives`.
[5] There are two types of games in PHRASE DETECTIVES, one consists in (dis)confirming a previous game, the other is about selecting the appropriate referent.
[6] The following Web page provides some details about the project, as well as the source code: `http://galoap.codeplex.com/`.
[7] See: `http://challenge.insemtives.eu/`.

2. RESOURCE TO CREATE

2.1 Dependency syntax

Traditionally, syntactic structures of natural language sentences are described either as dependency trees or as phrase structure trees. We chose here to rely on dependency syntax because it is often used in statistically based parsers [12] and it has proven to be easier to map to semantic relations, for example to produce semantic role labeling. Modern presentation of dependency syntax is due to Tesnière [18] and is based on the concept of labeled binary relations bonding lexical units of the sentence.

In figure 1, a dependency tree for the French sentence *"J'espère que nous ne devrons pas avoir recours au type de mesures que vous proposez."* (*I hope that we will not have to resort to the type of measures you propose*) is given; each arc describes a relation, the source of the arc is called the *governor* or the *head*, and the target is called the *dependent*. One lexical unit is the root of the structure (*"espère"*, in the example); each lexical unit (except for the root) is the dependent of exactly one dependency relation.

2.2 Corpus

A freely available syntactically annotated corpus named SEQUOIA exists for French. It was first annotated with phrase structure trees, then converted to dependency structures [3]. The main problem with the corpus is its size: it only contains 4,000 sentences and we need a much larger corpus. SEQUOIA will be used as a the seed and reference corpus for the game.

We are working on a new version of the corpus, including deep syntax annotation [2]. The annotation guide for this campaign comprises about 50 pages. The full corpus was corrected and annotated by two different research teams. In each team, several annotators (students in linguistics) worked on the corpus and a local expert adjudicated the differences. A second collaborative adjudication by members of the two teams is ongoing. This multi-layer correction of SEQUOIA will allow to produce very high quality annotations, that will be used as reference in ZOMBILINGO.

2.3 Pre-annotation

The goal of this project is to build a corpus annotated following SEQUOIA's guidelines, without the deep syntax layer, with a much larger set of sentences. We want to make the annotated corpus freely available, so we have to select freely available texts to feed the system, for instance WIKIPEDIA articles, public domain novels, etc. These new sentences will be pre-annotated using available parsers for French (either symbolic ones like LEOPAR [16] or statistical ones). This pre-annotation will be done in an asynchronous way, each time a corpus is added to the game. We will also use the confidence measures given by parsers as a score to be used later in the comparison with solutions provided by players. The pre-annotation will not be shown to the players and will be used, instead, as a temporary reference annotation.

3. PERFORMING A COMPLEX TASK USING A GWAP

3.1 Previous experiments

In both PHRASE DETECTIVES and FOLDIT, the players have to go through a training period before they can fully participate in the game, in order to ensure that they have read and understood the instructions. During this phase, the players are given feedback about their answers through more or less interactive corrections and hints.

In PHRASE DETECTIVES this training phase is mandatory and ends when the player reaches a certain level (50% of right answers), whereas in FOLDIT, the training is organized into introductory levels, each one introducing the player to new concepts, with only beginners puzzles being available to players with less than 150 points. The somewhat more complex and more flexible training mechanism of FOLDIT reflects the complexity of the task at hand.

As for dependency syntax annotation, in the campaign we led with students in linguistics, we decomposed the task into phenomena, and trained the annotators on one phenomenon just before they annotated it. This method proved to be quite efficient, as it prevented the annotators from having to read and memorize a complex annotation guide. Annotators learn details about each phenomenon through examples and after a while, they are able to deal with all phenomena in complex sentence. We hope to apply the same idea to players even if we know that we cannot except the same skills from them.

3.2 Decomposing the complexity

In ZOMBILINGO, we intend to combine the methods presented in the previous section: the task will be decomposed into relation types (SUBJ, OBJ, A_OBJ, ...) or phenomena (relative clause, coordination, ...), and the player will have to go through a training phase for each phenomenon before s/he can play the real game. This focused example-based training should help dealing with the large amount of details in the annotation guide, reducing the tediousness of the task.

Moreover, the players will be spared the annotation of obvious relation types like determiner to noun governor. This kind of relations will be dealt with by the automatic pre-annotation process.

In order to make sure the results are as accurate as possible, we need to assess the difficulty of the sentence (D), to match it to a player skilled enough to analyze it. While the latter is simple, using solely his/her level (L), we need a more complex, flexible indicator for the former. First, a computation will give an initial rating for the difficulty D of the sentence, based on the number of phenomena and their nature, as well as the corpus they come from. Second, as we cannot rely only on this computation, we will make D evolve depending on the results of the player on the sentence and his/her level, L. For example, if a player analyzes correctly a phenomenon with a difficulty rated above L, the system will lower D. On the contrary, if a player analyzes incorrectly a phenomenon with a difficulty rated below L, the system will increase D. The score and reward offered after a correct analysis will depend on the difficulty of the analyzed phenomenon. It should also give us insights on the types of sentences that are considered complex for players as compared to our system.

We are aware that this decomposition of the full annotation into separate phenomenon might generate consistency issues in the sentence. For instance, if many players decide that a word w_1 is the object of a verb whereas the automatic pre-annoation has chosen another word w_2 to be the

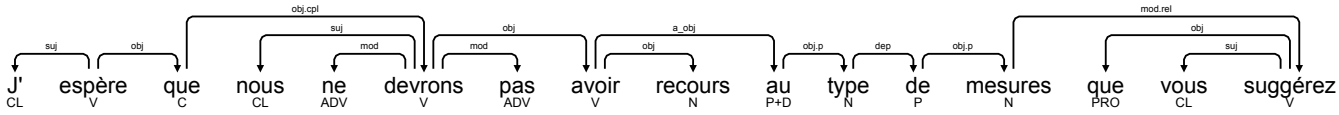

Figure 1: Example of dependency structure

object of the same verb, this will lead to a conflict. Most of these inconsistencies can be detected statically and will be used to increase the probability of annotations in conflict to be submitted to players: in the example above, players will be asked for the governor and the function of w_2. If the inconsistency remains, an expert will decide which is the correct annotation. However, some inconsistencies are unpredictable and will be corrected by an expert as soon as they are detected in the output corpus.

Another issue to tackle is the sentence length. If a sentence is too long to fit the screen, we can imagine to put the focus on the lexical unit concerned by the relation. However, this may not be enough, for example for long distance dependencies. For this last case, we will explore different solutions like a magnifying glass to search for the governor or a way to fold or unfold subparts of the sentence.

We will follow the examples of FoldIt and JeuxDeMots and will finalize the design of the game iteratively, using the feedback of the players to improve the decomposition of the complexity and the gameplay.

4. MOTIVATING PLAYERS

There are several motivation theories, including the *Game-Flow* theory [17], derived from the more general *Flow* theory, dating from 1975 and detailed in [15]. However, these theories are game-centric. The researcher who created JeuxDeMots, M. Lafourcade, uses a more people-centric approach, the CIA motivational factors[8], referred to with the MICE acronym: *Money* (in our case, reward), *Ideology* (here, interest), *Constraint* (obviously light), *Ego* (related to the position of the player in the community of players).

4.1 Motivational factors using the MICE method

4.1.1 Money and reward

We inserted in Zombilingo reward mechanisms that are popular in video games: achievements and in-game items.

As in Foldit, we plan to give achievements when players complete a challenging task (finishing a tutorial, a level, ...) to motivate them to progress further. We will also grant achievements for less glorious tasks (buying items, reading the FAQ, ...) to push players to discover unknown mechanisms or information. The majority of these achievements will have no real value (just a "completed" mark), existing only to satisfy the player's ego, but the most challenging achievements will offer in-game rewards when completed. As opposed to achievements, in-game items are active: players can use them to play. We imagined two ways to gain items: by buying them thanks to the in-game currency earned by players or by looting them randomly, exploiting the eagerness of players to get lucky. In the future, we plan to roll out a full market where players can exchange items, enhancing the social aspect of the game. Once the items are acquired, they can be used for customizing the player's profile thanks to cosmetic elements, for relieving the efforts of the players during the game (for example, increasing the timer for timed games) or for getting a competitive advantage in the future challenge mode. By making items a strategic element, it should motivate players to acquire them by playing more to earn more in-game currency and later, trade them with other players and develop a self-sufficient ecosystem.

[8]Personal communication, November 6th, 2013.

4.1.2 Ideology and interest

Previous GWAP experiments, like JeuxDeMots, demonstrated that most players do not play because they want to help science, but because it is fun. Subsequently, we decided not to brand Zombilingo with its research-helping objective and instead, to focus on an attractive theme related to the task to perform. We chose the Zombie one for two reasons: first, it has become very popular and second, we can relate dependency syntax annotation to Zombies looking for brains (heads) to devour, as shown in figure 2. The success of video game like Plants vs. Zombies demonstrated that all kind of players can be attracted to this universe with the right touch of humor.

As it was pointed out to be effective in Phrase Detectives [7], we also plan to let the player choose among corpora according to his/her personal interests (history, geography, medicine, etc).

4.1.3 Constraint and retention

The idea for constraints is to find a balance between pushing people to play the game as much as possible without putting them off. The gameplay is organized around game sessions of five items to analyze before really earning points and rewards. The scoreboard displays the scores for the day, week, month and all-time, motivating people to play often if they want to stay on the lists. Moreover, thanks to weekly or daily objectives, working like achievements but in a short, limited timespan, players should come back regularly to earn bonuses.

4.1.4 Ego and community

In the same line as achievement, i.e. easy to implement rewarding mechanisms, we will also add titles and cosmetic elements. [9] Titles are earned through achievements and will be displayed under the name of the player, to show his/her prowess and enhance his ego as other players will be able to see them. Cosmetic elements are for the moment limited to the avatar, which will change over time. It will undergo a process of decomposition as players gain new level, exciting their curiosity about what the new avatar will look like[10].

[9]Popular games like Team Fortress 2 demonstrated that such cosmetic items is very successful in motivating players.
[10]Much like in Phrase Detectives, but much more gory.

1. Player profile
2. Game progression
3. Interactive help
4a. Word played
4b. Relation type or phenomena to find
4c.''Hand'' to drag-and-drop to link the word played to the chosen answer
5. Access to in-game objects

Figure 2: Example of ongoing game in Zombilingo (mockup): "She had her assistant sent up".

We also aim at creating and maintaining an active social community so that players can feel welcome and share their experience. In the future, a challenge mode, inspired from the one in JeuxDeMots, where skilled players will be able to post their own sentence for other players to analyze, will surely help in strengthening the social aspect of the game.

4.2 Quality of the data

4.2.1 Player level as a trustworthiness indicator

Level progression is the core of the game, as it is directly linked to the accuracy of the performed analysis. Higher level should always correspond to experienced, reliable players creating good quality resources. They gain a score per phenomenon, which increases each time they earn points while playing on this phenomenon. Once they reach a predefined number of points, it will increase their level, and thus the difficulty of the game for this phenomenon.

Following the model in Phrase Detectives, we find important to give players a goal through level progression, so they know what they are aiming at and do not feel lost or without clear purpose. Gaining new level will be increasingly difficult, in order to avoid frustration at the beginning and to push experienced players to play more and more as they progress, allowing us to capitalize on their skills.

4.2.2 Asserting the correctness of the analyses

In order to reward a player and to guarantee the quality of the data, we need to make sure his/her analyses are correct. GWAPs like JeuxDeMots and Phrases Detective compare the results of the players, and consider that the common answers are the correct ones. We have decided to complement this method, taking into account the pre-annotation scores.

The sentences uploaded into the game will have been pre-annotated, with a confidence score associated to each annotation (C). This score will evolve each time a player analyses the sentence. In the case of a player giving the same analysis as the computer, C will increase. If a player gives a different analysis, the system will register a new entry for the phenomenon (see figure 3) with a new C. C will increase or decrease depending on the number of players playing the analysis, as well as their levels.

ID sentence	ID phen.	ID player	word from	word to	C
11	24	48	monter	assistant	0.88
11	24	82	monter	elle	0.20
11	24	53	monter	son	0.02

Figure 3: How the confidence scores (C) for one phenomenon are registered.

The correct analysis for the phenomenon will be the one with the highest C. In order to keep the players motivated, each time an analysis is registered as the new correct one, all the players having given this analysis (thus earning no point at the time because it was considered false) will earn an attractive retroactive bonus.

5. CONCLUSION AND FUTURE WORK

Detailed specifications of ZOMBILINGO are already available and the development itself should start in April. We first aim at creating a prototype that will be rapidly tested by our colleagues at LORIA, in order to validate our design choices and check that the good player/good resource goal is met. We will then correct and improve what will have to be and open the game to everyone on the Web.

We plan to advertise ZOMBILINGO using our various personal and professional networks, be they social networks like FACEBOOK or LINKEDIN, or researchers networks like conferences. We will also benefit from the help of the LORIA communication department and from the University of Lorraine students networks.

The annotated corpus will be made freely and continuously available, following the example of JEUXDEMOTS[11]. If the game is successful in attracting players and producing quality annotation, we will consider adapting it to other languages lacking syntactically annotated data.

6. ACKNOWLEDGMENTS

The authors would like to thank Charles Ancé for his wonderfully funny drawings, which will undoubtedly help attracting and motivating players. We also want to thank M. Lafourcade for his inputs and feedback on the project.

7. REFERENCES

[1] A. Böhmová, J. Hajič, E. Hajičová, and B. Hladká. The prague dependency treebank: Three-level annotation scenario. In A. Abeillé, editor, *Treebanks: Building and Using Syntactically Annotated Corpora*. Kluwer Academic Publishers, 2001.

[2] M. Candito, G. Perrier, B. Guillaume, C. Ribeyre, K. Fort, D. Seddah, and E. de la Clergerie. Deep Syntax Annotation of the Sequoia French Treebank. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Reykjavík, May 2014.

[3] M. Candito and D. Seddah. Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Proceedings of the Traitement Automatique des Langues Naturelles (TALN)*, Grenoble, France, June 2012.

[4] J. Chamberlain, K. Fort, U. Kruschwitz, M. Lafourcade, and M. Poesio. Using games to create language resources: Successes and limitations of the approach. In I. Gurevych and J. Kim, editors, *The People's Web Meets NLP*, Theory and Applications of Natural Language Processing, pages 3–44. Springer Berlin Heidelberg, 2013.

[5] J. Chamberlain, U. Kruschwitz, and M. Poesio. Constructing an anaphorically annotated corpus with non-experts: assessing the quality of collaborative annotations. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, People's Web '09, pages 57–62, Stroudsburg, PA, USA, 2009.

[6] J. Chamberlain, M. Poesio, and U. Kruschwitz. Phrase Detectives: a web-based collaborative annotation game. In *Proceedings of the International Conference on Semantic Systems (I-Semantics'08)*, Graz, Austria, 2008.

[7] J. Chamberlain, M. Poesio, and U. Kruschwitz. A new life for a dead parrot: Incentive structures in the phrase detectives game. In *Proceedings of WWW 2009*, Madrid, Spain, 2009.

[8] S. Cooper, A. Treuille, J. Barbero, A. Leaver-Fay, K. Tuite, F. Khatib, A. C. Snyder, M. Beenen, D. Salesin, D. Baker, and Z. Popović. The challenge of designing scientific discovery games. In *Proceedings of the Fifth International Conference on the Foundations of Digital Games*, FDG '10, pages 40–47, New York, NY, USA, 2010. ACM.

[9] K. Fort, G. Adda, and K. B. Cohen. Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics (editorial)*, 37(2):413–420, June 2011.

[10] J. Hana and B. Hladká. Getting more data - schoolkids as annotators. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 4049–4054, Istanbul, Turkey, May 2012.

[11] F. Khatib, F. DiMaio, S. Cooper, M. Kazmierczyk, M. Gilski, S. Krzywda, H. Zabranska, I. Pichova, J. Thompson, Z. Popović, et al. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature structural & molecular biology*, 18(10):1175–1177, 2011.

[12] S. Kübler, R. T. McDonald, and J. Nivre. *Dependency Parsing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2009.

[13] M. Lafourcade. Making people play for lexical acquisition. In *Proceedings of the 7th Symposium on Natural Language Processing (SNLP 2007)*, Pattaya, Thailand, December 2007.

[14] C. Liu, H. Wang, S. McClean, E. Kapetanios, and D. Carroll. Weighting Common Syntactic Structures for Natural Language Based Information Retrieval. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1485–1488, New York, NY, USA, 2010. ACM.

[15] J. Nakamura and M. Csikszentmihalyi. The concept of flow. *Handbook of positive psychology*, pages 89–105, 2002.

[16] G. Perrier and B. Guillaume. Leopar: an Interaction Grammar Parser. In *Proceedings of the Workshop on High-level Methodologies for Grammar Engineering, ESSLLI*, pages 121–122, Dusseldorf, 2013.

[17] P. Sweetser and P. Wyeth. Gameflow: a model for evaluating player enjoyment in games. *Computers in Entertainment (CIE)*, 3(3):3–3, 2005.

[18] L. Tesnière. *Éléments de syntaxe structurale*. Klincksieck, Paris, 1959.

[19] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '04, pages 319–326, New York, NY, USA, 2004. ACM.

[11] The created lexical network is available at: `http://www2.lirmm.fr/~lafourcade/JDM-LEXICALNET-FR/?C=M;O=D`.