Crowdsourcing Complex Language Resources: Playing to Annotate Dependency Syntax

Bruno Guillaume, Karën Fort, Nicolas Lefèbvre

Crowdsourcing Complex Language Resources: Playing to Annotate Dependency Syntax

Bruno Guillaume
Inria Nancy Grand-Est/LORIA
bruno.guillaume@inria.fr

Karën Fort
Université Paris-Sorbonne / STIH
karen.fort@paris-sorbonne.fr

Nicolas Lefebvre
Inria Nancy Grand-Est/LORIA
nicolas.lefebvre@inria.fr

Abstract

This article presents the results we obtained on a complex annotation task (that of dependency syntax) using a specifically designed Game with a Purpose, ZombiLingo.[1] We show that with suitable mechanisms (decomposition of the task, training of the players and regular control of the annotation quality during the game), it is possible to obtain annotations whose quality is significantly higher than that obtainable with a parser, provided that enough players participate. The source code of the game and the resulting annotated corpora (for French) are freely available.

1 Introduction

For better or worse,[2] the overwhelming domination of machine-learning systems in natural language processing (NLP) over the past 25 years and the increasing use of evaluation campaigns and shared tasks have put human-annotated corpora at the heart of the field. Manual annotation is now the place where linguistics hides.

The availability of manually annotated corpora of high quality (or, at least, reliability) is therefore key to the development of the field in any given language. However, the creation of such resources is notoriously costly, especially when complex annotations, e.g. for dependency syntax, are at issue. For example, the cost of the Prague Dependency Treebank was estimated at $600,000 in (Böhmová et al., 2001).

Over the years, many solutions have been investigated in the attempt to lower manual annotation costs. One obvious avenue is to use an appropriate annotation tool, as shown for example in (Dandapat et al., 2009). As a complementary aid, NLP systems can be used to reduce the annotation burden – either beforehand, for example with tag dictionaries (Carmen et al., 2010) and more generally with pre-annotation (Skjærholt, 2013), or more iteratively during the annotation process, for instance through active learning (Baldridge and Osborne, 2004). These solutions have proved efficient and have indeed helped reduce the annotation cost, but the creation of a large annotated corpus in the traditional manner remains very expensive.

Another way to address the issue is simply to limit the amount paid to the human annotators. This is the case with microworking crowdsourcing, especially through the use of platforms like Amazon Mechanical Turk, via which the workers are (micro)paid to perform simplified tasks (Snow et al., 2008). Apart from the ethical issues raised by these platforms (detailed in (Fort et al., 2011)), microworking platforms do not support true training of the workers: tests can be set up to select them, but they have to perform hidden work to train themselves (as shown in (Gupta et al., 2014)). Consequently, tasks must be simplified to be manageable by workers: for example, a task related to recognition of textual entailment across several sentences (which might actually reflect entailment, neutrality, or contradiction) could be simplified by presenting only one pair of sentences and a binary response to a single question, such as "Would most people say that if the first sentence is true, then the second sentence must be true?"

[1] See: http://zombilingo.org/.

[2] See (Church, 2011) for an in-depth reflection on the subject.

(Bowman et al., 2015). To paraphrase (Dandapat et al., 2009), it seems that there is no easy escape from the high cost of complex linguistic annotation.

We present here an on-line game that enables the production of quality annotations of complex phenomena (here, dependency syntax), tested on French. The produced corpus is constantly growing and is designed to be (i) completely free and available, and (ii) of sufficient quality. We first examine the existing treebanks for French and their limitations and detail some previous experiments with Games with a Purpose (GWAPs). Then we present the game we developed and the evaluation we performed on the produced annotations. We finally conclude by discussing the potentials and limits of GWAPs for the creation of complex language resources.

2 Previous Work

2.1 Treebanking for French

As surprising as it may seem, French has long been relatively low-resourced, primarily due to legal issues: lexicons and annotated corpora existed but could not be used or redistributed freely. This is in particular the case for the French Treebank (FTB or *corpus arboré de Paris 7*) (Abeillé et al., 2003), which is available for research purposes only and cannot be freely redistributed.[3] Several versions are reported in the literature, with the size varying from 12,351 sentences and 350,947 words (for the FTB-UC) to 18,535 sentences and 557,149 words[4] (for the FTB-SPRML).

To try and circumvent this restriction, Candito and Seddah created the Sequoia corpus (Candito and Seddah, 2012), which is freely available[5] under a LGPL-LR license, but is limited to 67,038 tokens. The same authors developed an additional question bank with 23,236 tokens (Seddah and Candito, 2016). Both corpora use the same annotation guide and set of relations as the FTB.

A Universal Dependency corpus (McDonald et al., 2013) was created for French and is freely available under a CC BY-NC-SA license. In version 1.3, released in May 2016, it contains 401,960 tokens, but it "has not been manually corrected systematically".[6] Moreover, the annotation format for Universal Dependencies suffers from certain drawbacks, for example, it does not distinguish between arguments and modifiers for nominal complements of verbs.

Other treebanks exist for French, but they either concern spoken language, as with Rhapsodie (Lacheret et al., 2014) or the oral Treebank (Abeillé and Crabbé, 2013), or specific language types, like the Social Media Treebank (Seddah et al., 2012).

This situation (in which references exist, but a large, fully available, manually annotated corpus is lacking) makes dependency syntax for French an appropriate candidate for testing a new paradigm for complex linguistic resource development: the use of on-line Games with a Purpose.

2.2 Playing to Create Language Resources

Games with a Purpose are games in which participants, knowingly or not, create data by playing. They are not serious games as such, as their main purpose is not to train people, but to produce data (such as annotations, lexicon entries, image labels, etc). GWAPs for NLP are broadly surveyed in (Lafourcade et al., 2015), and a detailed analysis of their performance in comparison with other means of language resource production is provided in (Chamberlain et al., 2013). Thus, we will focus here on the complexity of the resource to be produced, rather than on giving an exhaustive list of games.

Most GWAPs rely on the players' knowledge of the world and innate capabilities to enable creation of data, in our case language resources. This reliance was seen for example in the very well-known ESP Game (von Ahn and Dabbish, 2004), in which participants played by labeling images. Another early GWAP is JeuxDeMots[7] (Lafourcade, 2007), in which players created a lexical network of more than 47 million relations between more than 900,000 entries (terms and named entities), simply by entering

[3] See `http://www.llf.cnrs.fr/fr/Gens/Abeille/French-Treebank-fr.php`.

[4] These numbers correspond to the version found on: `http://gforge.inria.fr/projects/fdtb-v1`.

[5] See here: `http://deep-sequoia.inria.fr/`.

[6] See `http://universaldependencies.org/fr/overview/introduction.html`.

[7] The game is still running online: `http://www.jeuxdemots.org`.

terms and named entities associated in a more or less specified way with another entry. The video games presented in (Jurgens and Navigli, 2014) also seem[8] to target the users' intuition and knowledge of the world to perform word sense disambiguation and a mapping from WordNet senses to images. A more limited gamified interface is used in the Language Quiz[9], which asks the participants to perform sentence or tweet sentiment analysis, without any training.[10]

Other GWAPs benefit from the players' school knowledge. One such game is Phrase Detectives[11] (Poesio et al., 2013), in which participants are asked to identify the antecedent of a noun phrase in a text. The task is more complex than labeling images or associating ideas, so the game includes a mandatory short training phase, allowing the players to become familiar with it. Almost 8,000 players participated in the game, which enabled the annotation of anaphora relations in a 162,000 word corpus. The agreement between players and experts was evaluated overall at 84%.

The wordrobe website[12] presents a family of eight gamified tasks[13] whose results help to improve the Groningen Meaning Bank (Venhuizen et al., 2013). The proposed tasks all relate to semantic disambiguation (noun *vs* verb, co-reference identification, named entity annotation, etc) and while some are relatively easy – like Play Twins (noun *vs* verb) or Play Names (named entity annotation) – most require some more advanced (at least school-level) knowledge. However, the interface does not provide a training phase and the only help available is a short guide to the task.

In a very different domain (biochemistry), the creators of FoldIt[14] demonstrated that people with no specific knowledge of the subject could be trained to perform complex protein folding tasks with impressive results: players helped find the solution to a long-standing problem concerning protein crystal structure (Khatib et al., 2011). To achieve these results, players are trained within "introductory levels", solving puzzles of varying complexity "[...] introducing the player to new problem related concepts as though they are the rules of a game." (Cooper et al., 2010).

This approach seemed particularly suited to our task (dependency annotation with nearly twenty relations to annotate), so we decided to build upon it, asking the player to annotate one relation at a time, with a specific training process for each. We report here the results we obtained, which show that a GWAP can be used for a complex linguistic annotation task that is not directly linked to world knowledge or to human intuition. To our knowledge, no other such experiment has been carried out to date. The model of dependency syntax that we use is well-known in linguistics and NLP but not in the French educational system, so we can assume that people without a linguistic or NLP background have no knowledge of it.

It should be noted that an un-gamified crowdsourcing-based method has been proposed in (Hana and Hladká, 2012; Hladká et al., 2014) for the dependency syntax annotation of Czech. (In the Czech Republic, the presentation of syntax in school is very close to dependency syntax.) However, the authors report in (Hana and Hladká, 2012) that the accuracy of the annotations they obtained is significantly lower than that of their parser. Moreover, they evaluate the results on a small set of 100 sentences selected or created by the authors. In their more recent publication, they report only the tree editing distance between the produced annotation and the reference annotation; hence, we are unable to compare their results to those presented here.

3 Designing a Game for a Complex Task

We describe below the mechanisms used in the game to take into account the complexity of the task and the game's avoidance of direct reliance on intuition. We have chosen to rely on the Sequoia annotation guidelines (which largely follow the FTB guidelines). This is a natural choice if we want to use available resources for training players and for evaluation. Moreover, existing parsers for French have been

[8] The games do not seem to be running as of mid-July 2016, so we could not test them.

[9] See: `http://quiz.ucomp.eu`

[10] Two of the authors tested it in September 2016 and were at that time the only participants.

[11] See: `https://anawiki.essex.ac.uk/phrasedetectives/`.

[12] See: `http://wordrobe.housing.rug.nl`.

[13] The game features are reduced to a leader board and a betting option, we are therefore reluctant to call it a game, although there is obviously a continuum here.

[14] See: `http://fold.it`.

developed with the same schema.

3.1 Decomposing the Complexity of the Task

Since the analysis of a whole sentence is complex,[15] we decided to decompose it: the full annotation of a sentence is split into atomic tasks, where each task is linked to one type of dependency relation. Thus, the player is trained on one type of relation at a time, instead of twenty, and must focus on only one type of dependency relation across several sentences. S/he therefore has fewer elements of information to remember at a time. Moreover, each relation was awarded a level which reflects its difficulty (from the point of view of the game developers) and corresponds to a level in the game. The player can therefore choose amongst more and more relations as s/he progresses in the game. We decided not to include certain relations in the current version of the game. This concerns **punct**, which is not consistently annotated in reference, reldep, which is an underspecified relation used in different types of contexts, and relations which are not processed by Talismane. Note that the results given in Section 4 is based on all the relations except **punct**.

For each relation, explanations and examples are provided with the different contexts where it may appear.[16]

In addition to explanations, it is also important to give players examples of real utterances taken from a corpus. Reference examples are taken from the Sequoia corpus and are used in the two game mechanisms, TRAINING and CONTROL, described below.

We used only a part of the corpus in order to keep sentences aside for the evaluation of the produced annotations (see Section 4). In Table 1, we show how the reference corpus is split for different uses.

$\text{REF}_{Train\&Control}$	REF_{Eval}	*Unused*
50%	25%	25%
1,549 sentences	776 sentences	774 sentences

Table 1: Uses of the 3,099 sentences of the Sequoia reference corpus.

3.2 Playing the Game

The organization of ZombiLingo is illustrated in Figure 1, which presents how the TRAINING and play phases are articulated with both the sub-corpora described in Table 1 and the raw corpora to be annotated (usually extracted from Wikipedia). The reference corpus is used in three different phases, TRAINING, CONTROL and EVAL, as explained below.

3.2.1 TRAINING phase

Following the decomposition of the task, for each dependency relation, a specific TRAINING phase is required before entering the game. During this phase, sentences from the $\text{REF}_{Train\&Control}$ corpus are presented to the player and feedback is given in case of error.

Figure 2 illustrates the feedback given to the players in the TRAINING phase: the player was asked to find the second conjunct of a coordination *et* (and) and s/he wrongly answered *Europe*. A skull and crossbones flashes and a message revealing the right answer *parvient* (reaches) are displayed.

An advantage of offering a separate TRAINING for each relation is that the player does not have to wait long before starting the game. Once s/he is connected to the game, only a few minutes are required before starting actual play and production of annotations.

3.2.2 Play phase

In the general mode, the player chooses a relation from those available and a sequence of ten questions is proposed. The questions vary as follows:

[15] Insight concerning the complexity of syntactic annotation for the Penn Treebank is provided in (Marcus et al., 1993), where the learning curve for syntax was estimated to be twice that for part-of-speech (two months *vs* one).

[16] This is much like in an annotation guide but with simpler vocabulary and fewer details.

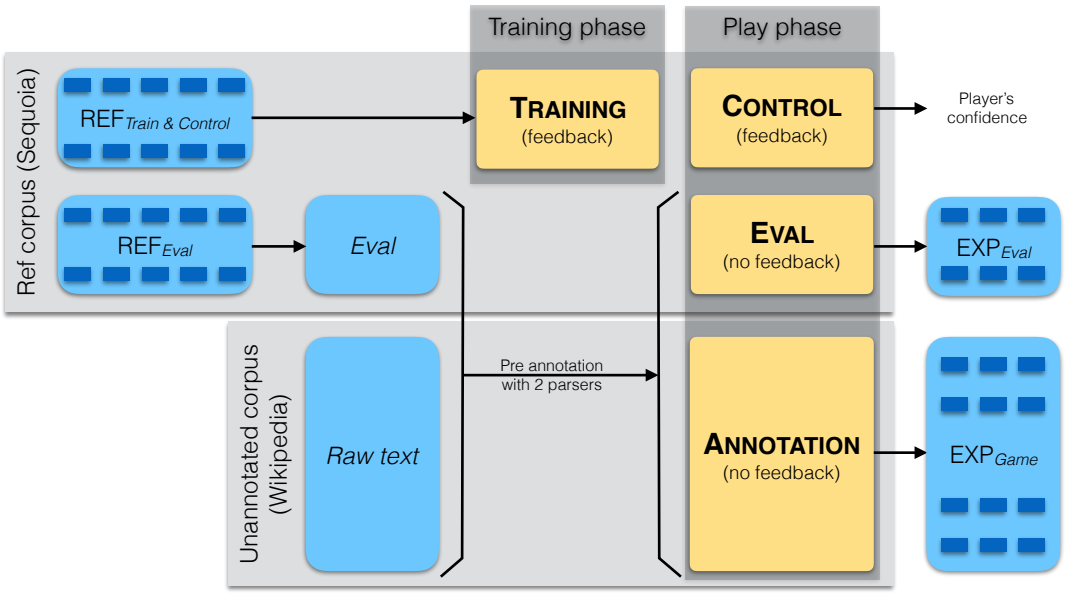

Figure 1: Organization of the different mechanisms and corpora

- For relations where the dependent element tends to be unique for a given governor (e.g. a verb has one subject, a noun has one determiner, and so on), the governor is given and the players must find the dependent;
- For relations involving several dependent elements (e.g. a verb has several prepositional phrases modifying it, a noun may be modified by several adjectives, and so on), the dependent element is given and the players must find the governor.

The player's answer is registered in the database and points are awarded if the same answer was previously given by a parser or by another player.[17] The number of points (in the game, they are represented as brains) that can be won on a certain relation depends on a global measure of the sentence complexity (measured as the maximum number of nested dependency relations) and on the level of the relation in the game.

3.2.3 CONTROL mechanism

TRAINING is not sufficient to ensure that players annotate data consistently: if they did the TRAINING on a certain relation a long time ago or if they play a lot, their ability to annotate this relation may decrease. To deal with this problem, we added a CONTROL mechanism during the game. The player is sometimes asked to annotate a relation in a sentence taken from the $\text{REF}_{Train\&Control}$ corpus. If the player fails to find the right answer, feedback including the solution is displayed (as in Figure 2). After a given number of failures on the same relation, the player has to redo the corresponding TRAINING. Using this CONTROL mechanism, we can also estimate our degree of confidence in a specific player on a specific relation (see Section 4) and take this into account to weigh his or her answer.

3.3 Behind the Curtain

3.3.1 Preprocessing Data

In the first version of the game, when new sentences were added, a parser was used to pre-annotate the sentences. Then, the players were asked to confirm or correct the parser's predictions. The main drawback of this arrangement is that the player would have a large number of very easy annotations to decide on and would usually agree with the parser.

[17] Unfortunately, if a player is the first to give a right answer, which is not given by the pre-annotation, s/he receives no points. A mechanism to give points off-line in this situation is planned but not yet implemented.

Figure 2: The main interface of the game during the TRAINING phase.

Modifying this arrangement, we now use two parsers to pre-annotate the corpora and ask the participants to play items for which the parsers give different annotations. The two parsers used are Talismane (Urieli, 2013) and FRDEP-PARSE (Guillaume and Perrier, 2015).

Talismane is a statistical parser trained on the training part of the FTB. It was evaluated on the dev and test part of the FTB with LAS scores ranging from 86.8% to 88.5%. As for FRDEP-PARSE, it is a parser which combines statistical methods for POS-tagging (using MElt (Denis and Sagot, 2012)) and symbolic methods for dependency parsing (with graph rewriting). FRDEP-PARSE was evaluated on Sequoia with a LAS score of 76.04% and a precision of 85.96%. (The system returns partial dependency syntax structures.)

It is important that the two parsers are based on different paradigms (statistical and symbolic): we can hope that the two tools are complementary and will produce different types of errors.

The next section will provide the two parsers' detailed results for the REF_{Eval} corpus.

3.3.2 Exporting Data

The players' annotations are stored in a database and each annotation receives a score. When a player is asked a question, we consider the set of possible answers in the database and adjust the score as follows:

- If the player's answer belongs to the set, the score of the answer is increased and the scores of its competing annotations (the rest of the set) are decreased.
- If the player gives an answer not in the set, a new annotation is created in the database with a default score and the scores of the answers in the set are decreased.

The positive or negative score adjustments are weighted by the level of the player, we thus award higher confidence to heavy players (who have usually reached higher levels) than to beginners. When a corpus is exported, for each token (lexical unit), we consider all the annotations in the database for which it is a dependent element and select the one with the highest score. Thus, each token receives exactly one governor with one relation and we can ensure that the exported corpus contains well-formed dependency trees.

4 Quantitative and Qualitative Evaluations

4.1 Participation and Production

If a crowdsourcing approach to annotation is to succeed, enough participants must be induced to participate and create data. As of 2016, July 10, there were 647 players registered for our game. They have

collectively produced 107,719 annotations.

As a reminder, Table 2 shows a quantitative comparison of the existing corpora for French. What this table does not show is that the corpus produced through the game is still growing, and will grow as long as we support and advertise the application.

	Sequoia 7.0	UD-French 1.3	FTB-UC	FTB-SPMRL	ZombiLingo
Sentences	3,099	16,448	12,351	18,535	5,221
Tokens	67,038	401,960	350,947	557,149	128,046
Tokens/sentence	21.6	24.4	28.4	30.1	24.5

Table 2: ZombiLingo corpus size, as compared to other existing French corpora annotated with dependency syntax.

On the ZombiLingo corpus the average number of players' annotations by tokens (called the *density*) is 0.84 (107,719 / 128,046). Figure 3 shows the density of annotations per relation. The coverage is clearly not homogeneous: the density on the relation **aff** (affix) is greater than 6, whereas for some relations it remains below 1.

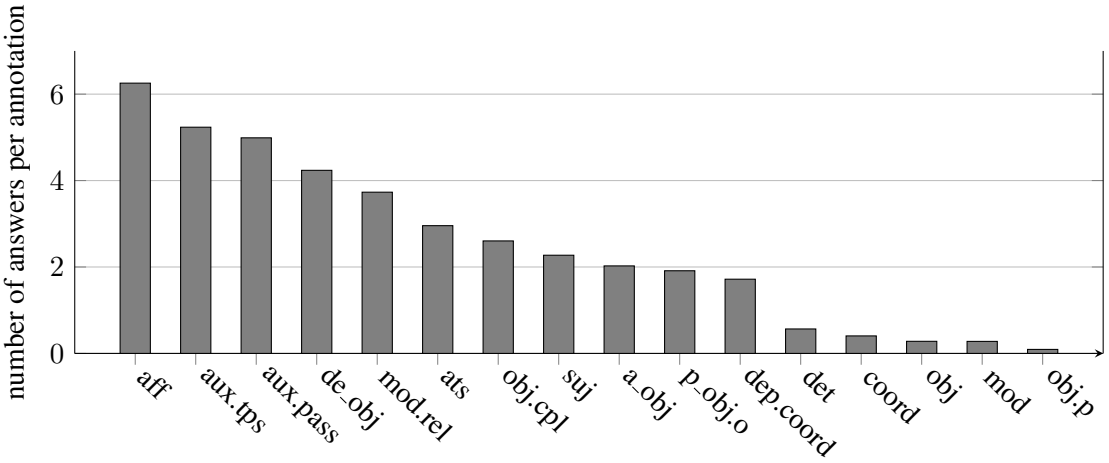

Figure 3: Density of answers for each relation.

This means that we need to attract at least some players to play a wider variety of relations types or to direct them towards some under-played relations, such as **mod** (modifier) or **coord** (coordination).

4.2 Qualitative Evaluation Methodology

In the first version of the game, we used the CONTROL mechanism to evaluate the quality of annotations produced by participants. We observed that 85.4% of CONTROL items were correctly annotated.

While this evaluation showed that it is possible to train players for a difficult task, it remains a partial evaluation, with severe drawbacks. It measures the performance of the participants on reference sentences but it gives no information about the annotations produced on pre-annotated sentences: if such a sentence contains a dependency relation not predicted by the parser, the player will never play this relation in this sentence. We call this omission *silence* in the produced data. On the other hand, if a relation is predicted by the parser but cannot be found in the sentence, the player will be asked to give an opinion on a question for which there is no sensible answer: we call such nonsensical situation *noise* in the data.

During the game, the user is supposed to click on a specific image (crossed bones) if a question has no answer but, although we train the participants on this alternative, we know that it is underused. For instance, when a player is tasked with finding the object of a verb, s/he tends to search for a word which looks like an object without realizing that this verb may actually have no object at all.

In order to obtain a precise evaluation of the produced annotations, we entered a part of our reference

corpus (named REF$_{Eval}$) into the game (see Figure 1), as if it were a raw corpus. We used it in the following way:

- the raw text corresponding to the REF$_{Eval}$ corpus is put into the general pipeline (i.e., it is parsed with the two parsers, the differences in the annotation being proposed to the players; this is the EVAL mechanism in the Figure 1);
- we report the scores obtained by each parser (i.e., recall/precision/F-measure) and the score of the corpus EXP$_{Eval}$, as exported by the game.

4.3 Results

The REF$_{Eval}$corpus contains 12,660 relations which can be played. On this subset of relations, the two parsers give the same output in 10,134 of the cases. Their analysis is correct in 96.84% of the cases (i.e. in 9,814 cases out of 10,134) and it is wrong in only 320 cases, which represents 3.16% of the output (i.e. 320 cases out of 10,134). However, even in cases where the two parsers provide two different wrong answers, the players will be asked to give their opinion on these cases and, hopefully, will produce a correct final annotation. In the end, the number of wrong annotations that will never be proposed to the players represents only 2.53% (i.e. 320 cases out of 12,660) of the relations which can be played.

In Table 3, we report the scores in three settings: the Talismane parser, the FRDEP-PARSE parser, and data exported by the game (EXP$_{Eval}$). In all of these setting, we may have partial dependency structures, so we report recall, precision and F-measure. We consider all relations (except punctuation which is not consistently annotated in the Sequoia corpus).

	Talismane	FRDEP-PARSE	Game
LAS / Recall	0.745	0.743	0.674
Precision	0.759	0.852	0.932
F-measure	0.752	0.794	0.782

Table 3: Evaluation of the data produced by the game on the REF$_{Eval}$ corpus (except punctuation).

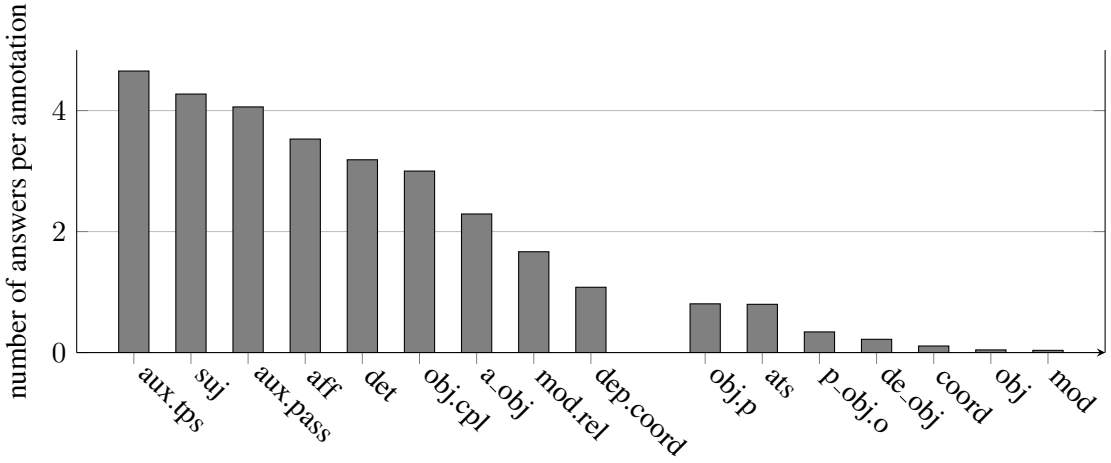

Figure 4: Density of answers for each relation on the REF$_{Eval}$ corpus (a whitespace separates relations where density is greater than 1 from the others).

Note that the score of the annotations produced by the game is lower than that of the second parser. This difference comes about because some relations were not played by enough players, as already observed in Subsection 4.1. To explore in more detail how the players influence the corpus score, we must look at these values relation by relation.

First, 3,575 game items correspond to the REF$_{Eval}$ corpus and only 2,644 game actions were performed by the player on these items (so the average density is 0.74). Moreover, as already observed in

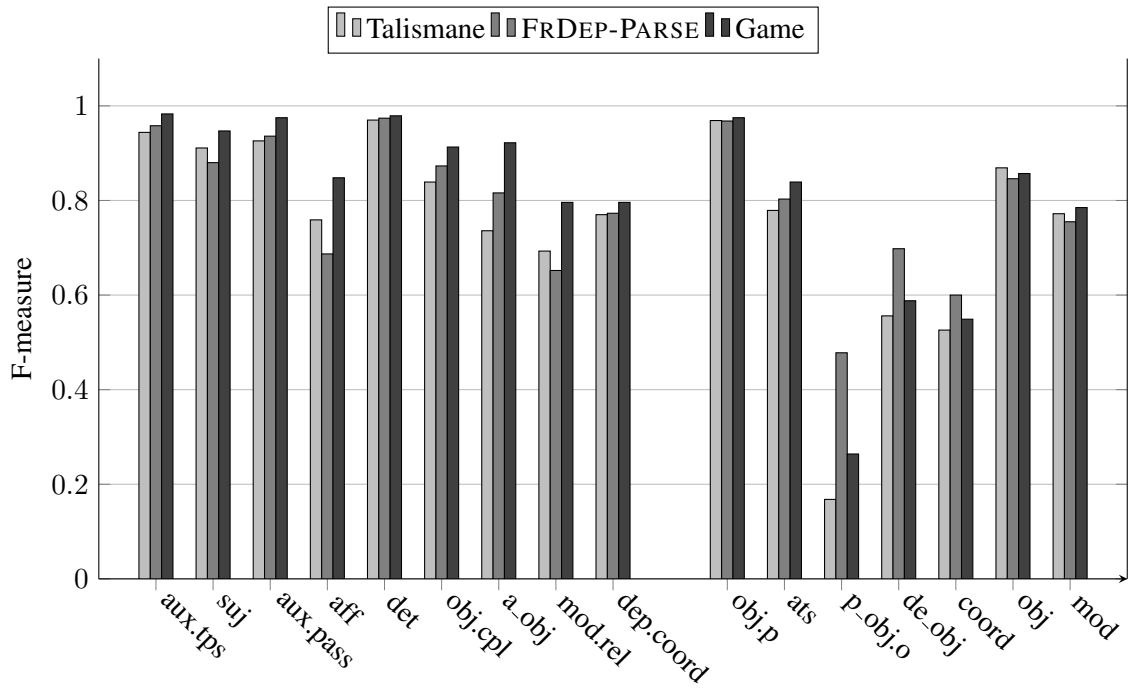

Figure 5: F-measures for the two parsers and the game, on each relation. (As in Figure 4, relations are split into two groups, with density greater than 1 on the left and density lower than 1 on the right.)

Figure 3, the distribution of the game actions is not at all homogeneous. If we discard relations for which the number of occurrences in the REF_{Eval} corpus is too low[18] to obtain significant results, the average density of game actions on game items ranges from 0.04 on **mod** (modifier) to 4.65 on **aux.tps** (tense auxiliary).

In Figure 4, we report the density of annotations on the corpus REF_{Eval} and Figure 5 gives, for the same relations in the same order, the values of the F-measure in the three settings (with Talismane in light gray, FRDEP-PARSE in a darker gray and finally the game export in dark gray).

For the relations where the density is higher than 1 (seen in the left hand part of the figures), we observe that the F-measure computed on the corpus from the game is always higher than that of the two parsers. From these experiments, we can conclude that we manage to obtain annotations with a quality significantly higher than that obtainable with a parser, provided that enough players participate. It is worth noting that amongst the relations which are densely played, some are considered as complex, such as **dep.coord** (the player has to find the head of the second conjunct of a coordination). Figure 2 shows an example of such a difficult case: the player has to read and to understand the whole sentence to give the right answer (it is a long distance dependency).

The next challenge is to induce players to annotate a wider variety of relations so as to increase the quality of the whole corpus. Relations with a very low density are either relations with a high number of occurrences (the REF_{Eval}corpus contains 1,874 **mod** relations) or relations which are less intuitive for the players. We will take this last factor into account and improve the documentation on these relations.

5 Conclusion

We have presented ZombiLingo, a game designed for a complex linguistic annotation task, namely dependency syntax. A first prototype of the game was released in July 2014 and an engineer started working on a production version in October 2015, completing the first version by the end of that year. As of July 2016, the game had enabled the production of more than 100,000 annotations for French, with a precision of 0.93. These results are very promising, especially for low-resourced languages. We still need to

[18]We have discarded relations with less than 25 occurrences in REF_{Eval}, namely **aux.caus** (causative auxiliary, 4 occurrences), **arg** (specific relation linking two parts of a range, 2 occurrences) and **dis** (dislocation, 1 occurrence).

fill some annotation gaps (e.g. for relations insufficiently played) and have developed for this purpose a new duel mode, which will allow senior players to compete with each other in the annotation of whole sentences.

The most difficult factor when using GWAPs is to find ways of attracting and keeping participants (Poesio et al., 2013), which requires a continuing communication effort. We have experienced "waves" of players following specific events we participated in or challenges we organized. We also attracted new players when advertising the game on social networks, but this effort must be regularly maintained and renewed.

The game source code is freely available on GitHub[19] under an CeCILL open-source license[20]. The code is designed to be easily adaptable to any human language: all messages are isolated from the code, which is Unicode compliant. We plan to adapt it to English and to a less-resourced language in the coming months.

The resource created for French is directly and freely available under a CC BY-NC-SA license from the game website[21] and is updated every night.

Acknowledgements

The development of ZombiLingo is funded by Inria, through an ADT grant. The French Ministry of culture has also helped us to communicate and publish results concerning ZombiLingo through two grants.

Above all, we want to thank all of the ZombiLingo players, without whom the resource would simply not exist! Special thanks to players JYA, Chouchou and Xohwohxo, who gave us valuable feedback on the game. We also wish to thank Djamé Seddah, for his help in building a solid state of the art and Guy Perrier, who, as Professor Frankenperrier, helped us to write the guidelines and to check errors in the reference. Finally, we want to thank the reviewers, who gave us interesting suggestions for improving this article.

References

[Abeillé and Crabbé2013] Anne Abeillé and Benoît Crabbé. 2013. Vers un treebank du français parlé. In *Proceedings of TALN 2013 - 20ème conférence du Traitement Automatique du Langage Naturel*, Sables d'Olonne, France, June.

[Abeillé et al.2003] Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a treebank for French. In Anne Abeillé, editor, *Treebanks*, pages 165 –187. Kluwer, Dordrecht.

[Baldridge and Osborne2004] Jason Baldridge and Miles Osborne. 2004. Active learning and the total cost of annotation. In *Proceedings of Empirical Methods in Natural Language Processing*, volume 15, pages 9–16.

[Böhmová et al.2001] Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2001. The prague dependency treebank: Three-level annotation scenario. In Anne Abeillé, editor, *Treebanks: Building and Using Syntactically Annotated Corpora*. Kluwer Academic Publishers.

[Bowman et al.2015] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

[Candito and Seddah2012] Marie Candito and Djamé Seddah. 2012. Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Proceedings of Traitement Automatique des Langues Naturelles (TALN)*, Grenoble, France, June.

[Carmen et al.2010] Marc Carmen, Paul Felt, Robbie Haertel, Deryle Lonsdale, Peter McClanahan, Owen Merkling, Eric Ringger, and Kevin Seppi. 2010. Tag dictionaries accelerate manual annotation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta, May.

[19] `https://github.com/zombilingo`

[20] `http://www.cecill.info/licences/Licence_CeCILL_V2.1-en.html`

[21] See the bottom of the following page: `http://zombilingo.org/informations`.

[Chamberlain et al.2013] Jon Chamberlain, Karën Fort, Udo Kruschwitz, Mathieu Lafourcade, and Massimo Poesio. 2013. Using games to create language resources: Successes and limitations of the approach. In Iryna Gurevych and Jungi Kim, editors, *The People's Web Meets NLP*, Theory and Applications of Natural Language Processing, pages 3–44. Springer Berlin Heidelberg.

[Church2011] Kenneth Church. 2011. A pendulum swung too far. *Linguistic Issues in Language Technology - LiLT*, 6.

[Cooper et al.2010] Seth Cooper, Adrien Treuille, Janos Barbero, Andrew Leaver-Fay, Kathleen Tuite, Firas Khatib, Alex Cho Snyder, Michael Beenen, David Salesin, David Baker, and Zoran Popović. 2010. 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. ACM.

[Dandapat et al.2009] Sandipan Dandapat, Priyanka Biswas, Monojit Choudhury, and Kalika Bali. 2009. Complex linguistic annotation - no easy way out! a case from bangla and hindi POS labeling tasks. In *Proceedings of the third ACL Linguistic Annotation Workshop*, Singapore.

[Denis and Sagot2012] Pascal Denis and Benoît Sagot. 2012. Coupling an annotated corpus and a lexicon for state-of-the-art pos tagging. *Language Resources and Evaluation*, 46(4):721–736.

[Fort et al.2011] Karën Fort, Gilles Adda, and Kevin Bretonnel Cohen. 2011. Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics (editorial)*, 37(2):413–420.

[Guillaume and Perrier2015] Bruno Guillaume and Guy Perrier. 2015. Dependency Parsing with Graph Rewriting. In *Proceedings of IWPT 2015, 14th International Conference on Parsing Technologies*, pages 30–39, Bilbao, Spain.

[Gupta et al.2014] Neha Gupta, David Martin, Benjamin V. Hanrahan, and Jacki O'Neill. 2014. Turk-life in india. In *Proceedings of the 18th International Conference on Supporting Group Work*, GROUP '14, pages 1–11, New York, NY, USA. ACM.

[Hana and Hladká2012] Jirka Hana and Barbora Hladká. 2012. Getting more data - schoolkids as annotators. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 4049–4054, Istanbul, Turkey, May.

[Hladká et al.2014] Barbora Hladká, Jirka Hana, and Ivana Luksová. 2014. Crowdsourcing in language classes can help natural language processing. In *Proceedings of the Second AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2014, November 2-4, 2014, Pittsburgh, Pennsylvania, USA*.

[Jurgens and Navigli2014] David Jurgens and Roberto Navigli. 2014. It's All Fun and Games until Someone Annotates: Video Games with a Purpose for Linguistic Annotation. *Transactions of the Association for Computational Linguistics (TACL)*, 2:449–464.

[Khatib et al.2011] Firas Khatib, Frank DiMaio, Seth Cooper, Maciej Kazmierczyk, Miroslaw Gilski, Szymon Krzywda, Helena Zabranska, Iva Pichova, James Thompson, Zoran Popović, et al. 2011. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature structural & molecular biology*, 18(10):1175–1177.

[Lacheret et al.2014] Anne Lacheret, Sylvain Kahane, Julie Beliao, Anne Dister, Kim Gerdes, Jean-Philippe Goldman, Nicolas Obin, Paola Pietrandrea, and Atanas Tchobanov. 2014. Rhapsodie: a prosodic-syntactic treebank for spoken french. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May.

[Lafourcade et al.2015] Mathieu Lafourcade, Nathalie Le Brun, and Alain Joubert. 2015. *Games with a Purpose (GWAPS)*. Wiley-ISTE, July.

[Lafourcade2007] Mathieu Lafourcade. 2007. Making people play for lexical acquisition. In *Proceedings of the 7th Symposium on Natural Language Processing (SNLP 2007)*, Pattaya, Thailand, December.

[Marcus et al.1993] Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English : The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

[McDonald et al.2013] Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu, and Castelló Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of ACL 13*, Sofia, Bulgaria, August.

[Poesio et al.2013] Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi. 2013. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Trans. Interact. Intell. Syst.*, 3(1):3:1–3:44, April.

[Seddah and Candito2016] Djamé Seddah and Marie Candito. 2016. Hard time parsing questions: Building a questionbank for french. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portoroz, Slovenia, May.

[Seddah et al.2012] Djamé Seddah, Benoît Sagot, Marie Candito, Virginie Mouilleron, and Vanessa Combet. 2012. The French Social Media Bank: a Treebank of Noisy User Generated Content. In *Proceedings of COLING 2012 - 24th International Conference on Computational Linguistics*, Mumbai, India, December.

[Skjærholt2013] Arne Skjærholt. 2013. Influence of preprocessing on dependency syntax annotation: speed and agreement. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 28–32, Sofia, Bulgaria, August.

[Snow et al.2008] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP 2008*, pages 254–263.

[Urieli2013] Assaf Urieli. 2013. *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Ph.D. thesis, Université de Toulouse II le Mirail, France.

[Venhuizen et al.2013] Noortje Joost Venhuizen, Valerio Basile, Kilian Evang, and Johan Bos. 2013. Gamification for word sense labeling. In Katrin Erk and Alexander Koller, editors, *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Short Papers*, pages 397–403, Potsdam, Germany, March.

[von Ahn and Dabbish2004] Luis von Ahn and Laura Dabbish. 2004. 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. ACM.