Proceedings of the LREC 2020 Workshop Games and Natural Language Processing, pages 39–43
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

A 3D Role-Playing Game for Abusive Language Annotation

Federico Bonetti[a,b], Sara Tonelli[b],
[a]Dept. of Psychology and Cognitive Science, University of Trento
[b]Fondazione Bruno Kessler, Trento
{fbonetti,satonelli}@fbk.eu

Abstract

Gamification has been applied to many linguistic annotation tasks, as an alternative to crowdsourcing platforms to collect annotated data in an inexpensive way. However, we think that still much has to be explored. Games with a Purpose (GWAPs) tend to lack important elements that we commonly see in commercial games, such as 2D and 3D worlds or a story. Making GWAPs more similar to full-fledged video games in order to involve users more easily and increase dissemination is a demanding yet interesting ground to explore. In this paper we present a 3D role-playing game for abusive language annotation that is currently under development.

Keywords: games with a purpose, game design, linguistic annotation, abusive language

1. Introduction

Games with a Purpose (GWAPs) have been exploited for many linguistic annotation tasks to enrich data with different information layers, ranging from word senses to anaphora. Gathering annotations from experts hired for the tasks can be expensive and time-consuming. Using gamification to collect annotations from players, instead, allows to combine the stronger motivation and the will to play again that games foster (Ryan et al., 2006) with lower average costs (Poesio et al., 2013; Vannella et al., 2014; Jurgens and Navigli, 2014). One of the main problems with GWAPs, however, is the low resemblance with commercial games, which are devised specifically for entertainment purposes (Jamieson et al., 2012). This is the case especially for existing games aimed at linguistic annotation.

Another common criticism to GWAPs is that they tend to exploit ephemeral extrinsic rewards (like collecting points, achieving high places in leaderboards, obtaining badges and so on) which might even harm motivation in the long run (Seaborn and Fels, 2015) and that do not represent the real essence of video games, as game designer Margaret Robertson claims (Robertson, 2010).

In the light of these criticisms, we show in this paper that it is possible to create a 3D video game for linguistic annotation using simple models created in Blender[1] even without domain-specific (i.e. 3D modelling) professional skills. In particular, we are presenting a role-playing game (RPG) rendered in 3D cel shading graphics (which means the style is cartoonish) with the purpose of collecting abusive language annotations. The goal is to create sentences that can be used to train a hate speech detection systems. The game is being developed with multiple target devices in mind so the ergonomics will fit both keyboard and touchscreen setups.

2. Related work

To date, there have been many attempts in the direction of gamifying a wide range of linguistic annotation tasks. These include *Phrase Detectives* for anaphora resolution (Poesio et al., 2013), *The Knowledge Towers* (Vannella et al., 2014) and *Puzzle Racer* (Jurgens and Navigli, 2014) for concept-image linking, *Infection* (Vannella et al., 2014), *OnToGalaxy* (Krause et al., 2010) and *JeuxDeMots* (Joubert et al., 2018) for semantic linking, *Argotario* (Habernal et al., 2017) for fallacious argumentation identification, *Zombilingo* (Fort et al., 2014) for dependecy syntax annotation, *Sentimentator* (Öhman and Kajava, 2018) for sentiment annotation, *Wordrobe* (Venhuizen et al., 2013) and *Ka-Boom!* (Jurgens and Navigli, 2014) for sense annotation. Researchers stress the fact that GWAPs should be designed in such a way that they integrate the task without sacrificing their 'gamefulness', otherwise the tasks may be perceived as work (Vannella et al., 2014). Some of these games try to exploit *disjoint design* (Krause et al., 2010), i.e. a technique by which the goal of the player and the goal of the task are kept separate. In particular, in *OnToGalaxy* players control a spaceship and have to shoot other spaceships with a certain label that does not satisfy the condition expressed in the instructions. We take into account that this separation, or task abstraction, could potentially harm the quality of the outcome, so tasks have to be thought very carefully. A goal that is phrased as *shoot the spaceships with a name that does not satisfy this condition* may very well drive the player's actions differently than a task that says *click on the label that satisfies the following condition*, if only because of the sense of challenge or excitement that arises. On the other hand, challenge and a gameful environment might be exactly what drives the players' actions in the right direction, to the point of improving the annotation quality over standard crowdsourcing methods (Vannella et al., 2014).

This separation is useful for hiding the task and making the whole experience feel less like work and more like play. However, hiding a task does not necessarily mean that the users must not be made aware of its presence. In fact, saying clearly that a game is useful for research purposes can be a motivator for players (Tuite, 2014).

Among the contributions we have analysed, some try to exploit this technique and we noticed that although two text-based annotation games take advantage of it – *Infec-*

[1] The Blender Foundation, https://www.blender.org/.

tion (Vannella et al., 2014) and *OnToGalaxy* (Krause et al., 2010) – they focus on word-level annotation tasks, while to our knowledge no existing GWAP with disjoint design performs a task at sentence level. Probably the games that push the most their looks and feel towards commercial games are *Infection* and *The Knowledge Towers*, where the player actually controls a character and is rather free to explore the virtual environment. However, as mentioned before, these games focus on word-level annotation and are in 2D, while we are experimenting with sentence-level annotation in a 3D scenery.

In Table 1 we summarise the main games developed for linguistic annotation, specifying which ones rely on disjoint design, the target of the annotation and the task. To our knowledge there is still no overlap between the sentence level annotation category and the disjoint design category.

Game	**Disjoint design**	**Task type**
Phrase Detectives (Poesio et al., 2013)	No	**Sentence level**
Zombilingo (Fort et al., 2014)	No	**Sentence level**
Sentimentator (Öhman and Kajava, 2018)	No	**Sentence level**
Argotario (Habernal et al., 2017)	No	**Sentence level**
Wordrobe (Venhuizen et al., 2013)	No	Word level
JeuxDeMots (Joubert et al., 2018)	No	Word level
OnToGalaxy (Krause et al., 2010)	**Yes**	Word level
Infection (Vannella et al., 2014)	**Yes**	Word level
Ka-Boom! (Jurgens and Navigli, 2014)	**Yes**	Word level
Puzzle Racer (Jurgens and Navigli, 2014)	**Yes**	Word level
The Knowledge Towers (Vannella et al., 2014)	**Yes**	Word level

Table 1: Feature summary.

3. Abusive Language Annotation

The goal of the game we implement is to collect data for hate speech detection (Fortuna and Nunes, 2018). Due to the increasing popularity of social media platforms such as Facebook, Twitter and Instagram, it has indeed become of crucial importance to automatically detect abusive messages online with the aim to suspend accounts, delete hate speech messages, etc. While existing hate speech detection systems have achieved good results on resource-rich languages using deep learning techniques (Basile et al., 2019), these data-intensive approaches require large amounts of high-quality annotated data for training, which are typically expensive and time-consuming to create. We therefore develop the first GWAP with the goal to annotate data to be used for hate speech detection.

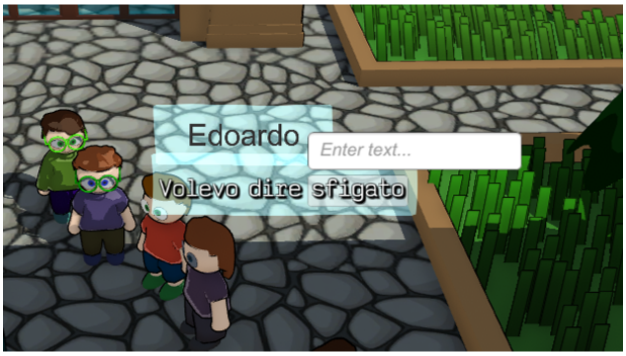
We distinguish between two different linguistic tasks: the goal of the first one is to collect a set of abusive and not abusive sentences. The goal of the second task is to identify, in an abusive sentence, which expressions or words are offensive, so to have a fine-grained annotation of the sentence, isolating only the offensive strings. For both cases, the game takes in input a corpus of sentences that may contain abusive language, with the goal to annotate them. For our first experimentation, we use the Italian WhatsApp corpus of cyberbullying interactions (Sprugnoli et al., 2018), containing 10 chats for a total of 14,600 tokens. The messages had been manually annotated as offensive or not, and the semantic type of the offense was also specified (e.g. body shame, sexism, blackmail, etc.). For our game, the existing annotation has not been taken into account, but it can be used to check whether the information on offensive messages collected through the game matches the gold standard. Since users are exposed to vulgar language in this game, a disclamer is put at the beginning where they are informed about the potential harm.

The input format for the game is rather straightforward: a standard .txt/.xml file containing a conversation (made up of *name* + space + *sentence* turns if it is a .txt file). The game engine takes this file in input, splits the turns, assigns random names to the speakers and represents the chat in the game as students talking to each other.

4. Tasks

4.1. Task 1: Sentence level annotation

The protagonist of our game, a high-school student, has been given a special device by a scientist and has been appointed the mission to lower the level of bullying in the school. This level is represented by a 'security meter' in the form of a classic health bar near the player's avatar in the heads-up display. The device makes it possible to tap into other people's minds to change what they are going to say. This mechanism in particular allows to annotate sentences and constitutes Task 1. In this task players have to change what a bully says, if it contains abusive language, in order to make the expression inoffensive. This is done by clicking on the tokens that represent what the bully is thinking. The purpose of the task is twofold. The main goal is to annotate the sentence as containing abusive language or not (if it does, it is fed to task 2). The secondary goal is to obtain pairs of abusive and non-abusive sentences. The dialogue phase unfolds as follows: when the player goes near a certain group of students, it is possible to overhear their conversation. Before every message, the player is able to read the speaker's mind: a cloud is shown where tokens are freely modifiable; when the change has been made, the bullies say what the player has told them to say, then they look puzzled and run away. The task implements disjoint design in the sense that what the players do is they *make sentences inoffensive* while the underlying task mechanics consist of *marking* sentences and *providing pairs of abusive and non-abusive sentences*. The task goal is *driven* by the surface goal. Both the modified sentence and the original sentence are kept in order to have positive and negative examples. The new sentence can be similar to the original one or rewritten from scratch, since the focus is on knowing if, not how, the sentences have been modified. The game leaves players rather free to change all the tokens they want. However, it is possible that users will only change the one or two tokens required to render the sentence less offensive. This would actually help us collect pairs of sentences where the difference is minimal, so that the classifier can learn from these examples to recognise offensive messages also when they are similar to not-offensive ones.


Edoardo

Volevo dire sfigato

Figure 1: Game screenshot of task 1: Modifying offensive sentences.

4.2. Task 2: Word level annotation

This task consists in erasing offensive expressions off a blackboard or a wall. The snippets of texts that make up the graffiti are taken from sentences annotated in Task 1 as offensive, so this also serves as a validation phase. Players can erase tokens they think are offensive by using a sponge or a wiper. The erasing mechanics adds a layer to the interaction, since erasing by rubbing an object against a surface in correspondence of a token is different than simply clicking on a token. Again, the idea is to make the task less direct but more satisfactory. Words are considered erased when more than 2/3 of the word surface has been wiped. In order to prevent the player from erasing too many inoffensive words, we put a limit to the available game resources involved in these mechanics (such as soap) and reward low waste.

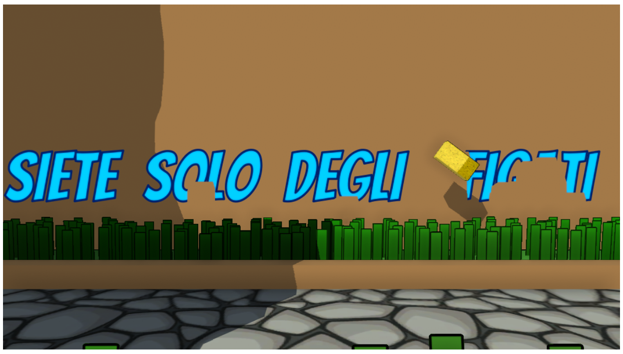


Figure 2: Game screenshot of task 2: Erasing graffiti with a sponge.

4.3. Score and quality control

To control the annotation quality, three methods are being implemented. The first one consists of randomly presenting players with gold standard annotated sentences. Players who show deviation from the gold standard are given hard feedback about their performance, with advice on how to improve it.

Another way of assessing whether players are good annotators, especially if no gold standard is available yet, is to check their response time with respect to the sentences presented. If players systematically skip sentences after a very short time, we can infer that their motivation or interest is low and rate their reliability accordingly. One way to cope with this is to either exclude the annotations or submit them to other players in the form of a specific validation task. Finally, agreement between players who annotate the same sentences will be used to add to their score. Regardless, a base score will always be given to players in both tasks, according to the amount of sentence skipping and time dedicated to the annotation. This score is partly represented in the security meter and partly used to calculate the experience points that allow the player to level up.

5. Game Design

5.1. Gameplay

The game world is intended to be, to an extent, free to roam, which means the player is allowed to explore freely, progressing with the story only when they feel ready to. During the exploration phase, it is possible to overhear conversations and intervene when hate speech is used, or erase abusive language off of walls and blackboards. These two instances of tasks reiterate themselves indefinitely, or until the player has reached a certain amount of discipline in the school that let them advance with the story.

A crucial issue is how to keep players engaged as progress is made through the story. A common datum is that games gradually increase the difficulty to keep the player challenged. This is modeled in Flow Theory applied to video games (Csikszentmihalyi, 1997; Cowley et al., 2008). However, many successful games (see *Minecraft*) do not implement difficulty as an upward curve. Rather, the player is motivated by the possibility to do more, to build more, to explore more. The difficulty changes according to the player's strategy and play style. In a game where the tasks consist of linguistic annotation we think that this is the best model. Rewards are primarily of power-ups, equip items, new mechanics and new areas to explore. However, as players advance, we plan to give them the possibility to annotate more ambiguous sentences, that is, sentences that have received mixed interpretations and are thus more difficult to classify.

5.2. Genre and setting

Choosing the right genre is important since it has an impact on how text is presented during game play. Role-playing games (RPGs) are a viable option when it comes to moderately high amounts of text since they naturally present players with lots of messages from non-playing characters. Since the tasks that have been implemented are based on hate speech and the corpus was created by young students, we decided to set the game in a school. The architecture and aesthetics were inspired by Mt Tacoma High School in Washington, USA. The model of the school is under construction but it is intended to be fully explorable when it is finished, allowing a certain amount of free roam.

5.3. Graphics

The game environment is a 3D world rendered in a cartoon style (called cel shading or toon shading), which is quite common in commercial video games. Thanks to the versatility of Unity and Blender and their widespread documentation, it is relatively easy to create 3D environments, as long as the models are kept simple. To match the basic style of the 3D models, we implemented a cel shader with black outlines. The final result was achieved by modifing an existing shader available for free on the Unity Asset Store. This choice was also influenced by the fact that some of the most successful commercial games of the last decade, and 3D games by Nintendo in general, use colorful graphics: *Fortnite* by Epic Games, *The Legend of Zelda: Breath of The Wild*, the *Super Mario* franchise and more independent experimental games like *Untitled Goose Game*, to name a few.

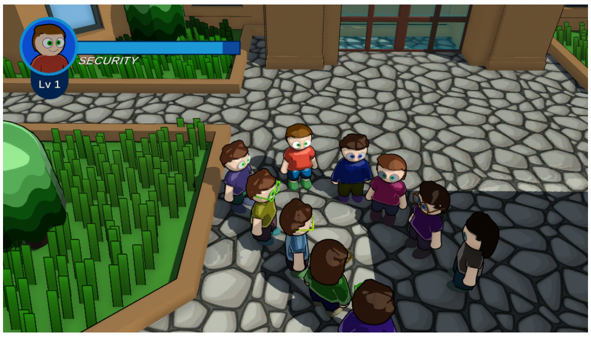


Figure 3: A view of the school yard.

5.4. Player representation

A core feature of many RPGs is the avatar customizability (especially in massive mutiplayer online RPGs, but also in traditional RPGs to a lesser extent). In our game the player is representend as a customizable 3D character. At the beginning of the game, players have the opportunity to create a character with the appearance they prefer. The game lets the players customize their avatar without asking for their gender: it is sufficient to choose the preferred hair style and clothes.

It is worth noticing that this feature is not limited to RPGs and recently there have been attempts to bring character customization even to genres where the player appearance is of minimum importance in terms of gameplay experience, like driving games (see *Forza Horizon 4* or even *Farming Simulator 2019*). This feature in particular seems to drive user motivation remarkably. It is not infrequent to see users online reporting having spent hours just in the character creation interface screen. Customization improves our sense of control over the game outcomes and makes it more likely that we continue playing (Turkay and Adinolf, 2015). Overall, the freedom to modify one's own avatar contributes to the sense of agency and autonomy, which is one of the three psychological needs theorized in self-determination theory: autonomy, competence and relatedness (Ryan et al., 2006).

5.5. Development tools

The game is currently being developed in Unity[2], in C#, relying on Blender for the 3D modeling. Both programs boast huge online documentation and Unity has many build options, including mainstream gaming consoles and WebGL, allowing easy multi-platform releases. Most importantly they are free to use, at least within a certain amount of profit in the case of Unity, and Blender is open source.

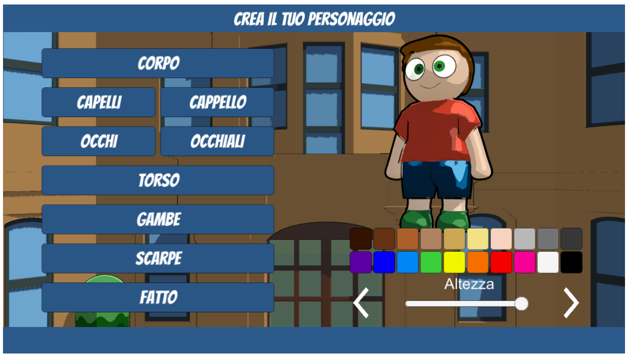


Figure 4: Character customization interface.

6. Conclusion and Future Work

In this paper we have presented a work-in-progress 3D role-playing video game for abusive language annotation that uses disjoint design as its core design strategy. This feature allows the designer to hide a task making the whole experience more gameful. This project aims at being a first step towards the use of disjoint design in a gamified application for sentence-level linguistic annotation. While we did not devise this game with a particular educational purpose in mind, it is certainly a welcome byproduct to be able to raise awareness about the topic of abusive language and cyberbullying.

One of our next steps will be to study a method to let players add their own content to be annotated later by other players. The exact way this will be made possible has not been defined yet. Some commercial games have already tried to gather text input by the players. An example from commercial games is *Kind Words*[3], where people are free to exchange supportive messages with each other, a mechanism that presents an obvious occasion for collecting corpora.

We are also planning a pilot study to evaluate the overall playability of the game and the task intrusiveness. A questionnaire to probe intrinsic motivation is being redacted, based on self-determination theory, to assess this aspect. An evaluation in terms of quality and cost of the annotations will also be made comparing our approach with the quality, time and cost of human annotation.

7. References

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F. M. R., Rosso, P., and Sanguinetti, M. (2019).

[2] Unity Technologies, https://unity.com/.

[3] Popcannibal, https://www.popcannibal.com/wp/

Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019*, pages 54–63.

Cowley, B., Charles, D., Black, M., and Hickey, R. (2008). Toward an understanding of flow in video games. *Computers in Entertainment*, 6(2):1, July.

Csikszentmihalyi, M. (1997). *Finding flow: The Psychology of Engagement with Everyday Life.* Finding flow: The psychology of engagement with everyday life. Basic Books, New York, NY, US.

Fort, K., Guillaume, B., and Chastant, H. (2014). Creating Zombilingo, a Game with a Purpose for Dependency Syntax Annotation. In *Proceedings of the First International Workshop on Gamification for Information Retrieval - GamifIR '14*, pages 2–6, Amsterdam, The Netherlands. ACM Press.

Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4):85:1–85:30, July.

Habernal, I., Hannemann, R., Pollak, C., Klamm, C., Pauli, P., and Gurevych, I. (2017). Argotario: Computational Argumentation Meets Serious Games. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 7–12, Copenhagen, Denmark. Association for Computational Linguistics.

Jamieson, P., Hall, J., and Grace, L. (2012). Research Directions for Pushing Harnessing Human Computation to Mainstream Video Games. In *Meaningful Play 2012*, East Lansing, MI.

Joubert, A., Lafourcade, M., and Brun, N. L. (2018). The JeuxDeMots Project is 10 Years Old: What We have Learned. In *Proceedings of the 2018 LREC Workshop "Games and Gamification for Natural Language Processing (Games4NLP)"*, pages 22–26, Miyazaki, Japan.

Jurgens, D. and Navigli, R. (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*, 2:449–464, December.

Krause, M., Takhtamysheva, A., Wittstock, M., and Malaka, R. (2010). Frontiers of a paradigm: exploring human computation with digital games. In *Proceedings of the ACM SIGKDD Workshop on Human Computation - HCOMP '10*, pages 22–25, Washington DC. ACM Press.

Öhman, E. and Kajava, K. (2018). Sentimentator: Gamifying Fine-grained Sentiment Annotation. In *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference (DHN 2019)*, volume 2084, pages 98–110, Helsinki, Finland, February.

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., and Ducceschi, L. (2013). Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Interactive Intelligent Systems*, 3(1):1–44, April.

Robertson, M. (2010). Can't Play, Won't Play. https://kotaku.com/cant-play-wont-play-5686393.

Ryan, R. M., Rigby, C. S., and Przybylski, A. (2006). The Motivational Pull of Video Games: A Self-Determination Theory Approach. *Motivation and Emotion*, 30(4):344–360, December.

Seaborn, K. and Fels, D. I. (2015). Gamification in theory and action: A survey. *International Journal of Human-Computer Studies*, 74:14–31, February.

Sprugnoli, R., Menini, S., Tonelli, S., Oncini, F., and Piras, E. (2018). Creating a WhatsApp Dataset to Study Pre-teen Cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59, Brussels, Belgium. Association for Computational Linguistics.

Tuite, K. (2014). GWAPs: Games with a Problem. In *Proceedings of the 9th International Conference on the Foundations of Digital Games*.

Turkay, S. and Adinolf, S. (2015). The effects of customization on motivation in an extended study with a massively multiplayer online roleplaying game. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 9(3).

Vannella, D., Jurgens, D., Scarfini, D., Toscani, D., and Navigli, R. (2014). Validating and Extending Semantic Knowledge Bases using Video Games with a Purpose. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1294–1304, Baltimore, Maryland. Association for Computational Linguistics.

Venhuizen, N. J., Evang, K., Basile, V., and Bos, J. (2013). Gamification for word sense labeling. In *Proceedings of the International Conference on Computational Semantics (IWCS)*, pages 397–403.