Experiment-Driven Development of a GWAP for Marking Segments in Text

Chris Madge
University of Essex
Wivenhoe Park, Colchester
CO4 3SQ
cjmadg@essex.ac.uk

Jon Chamberlain
University of Essex
Wivenhoe Park, Colchester
CO4 3SQ
jchamb@essex.ac.uk

Udo Kruschwitz
University of Essex
Wivenhoe Park, Colchester
CO4 3SQ
udo@essex.ac.uk

Massimo Poesio
University of Essex
Wivenhoe Park, Colchester
CO4 3SQ
poesio@essex.ac.uk

CHI PLAY'17 Extended Abstracts, October 15–18, 2017, Amsterdam, Netherlands

ACM ISBN 978-1-4503-5111-9/17/10.
https://doi.org/10.1145/3130859.3131332

Abstract
This paper describes *TileAttack*, an innovative highly configurable game-with-a-purpose (GWAP) designed to gather annotations for text segmentation tasks whilst exploring the effects of different game mechanics on GWAP for NLP (Natural Language Processing) problems, with a view to improving both quality of player contributions and player uptake. In this work we present a pilot experiment that shows *TileAttack* labelling "mentions" and being used to test the effects of in game time constraints on accuracy and player engagement. We present the results of this experiment using a set of metrics derived from those used for evaluating Free-To-Play (F2P) games.

Author Keywords
Games With A Purpose; Game Design; Natural Language Processing; Graphical User Interfaces

ACM Classification Keywords
H.5.2 [User Interfaces]: Graphical user interfaces (GUI); Evaluation/methodology; K.8.0 [General]: Games; J.4 [Social and Behavioral Sciences]: Psychology

Introduction
Many Natural Language Processing (NLP) tasks require large amounts of annotated text to train statistical models. These are often hand-annotated contributions [7]. This pro-

cess can be time consuming, expensive and tedious. Consequently, this requirement for annotated data remains an obstacle to progression for NLP. One proven method of reducing the time to gather the annotations is crowdsourcing [12]. However, this doesn't scale very well. When attempting to build large corpora gamification can be cheaper [8].

Gamification and GWAPs offer entertainment in exchange for contribution rather than a financial incentive. GWAPs have been successful in many applications attracting large numbers of users to label datasets and solve real world problems [4]. Examples include image labelling with *The ESP Game* [15], or protein folding with *FoldIt* [11]. Gamification has worked well for text problems (e.g. *Phrase Detectives* [8]), but there are limited examples of GWAPs for NLP. Presenting such challenges as a GWAP rather than applying gamification is a greater challenge, as it requires mapping the problem completely into a game, rather than adding selected game mechanics. However, GWAPs have the potential for greater rewards over gamified applications in terms of higher player engagement.

Games such as *Puzzle Racer* have shown that it is possible to create an engaging GWAP that produces annotations of a high quality at a reduced cost [3]. However, they have yet to achieve a player uptake or number of judgements comparable to GWAPs in other domains. GWAPs for text problems often present additional unique challenges compared to those for image labelling and other similar tasks e.g. users can differentiate between images features easily, but not so easily with text features [18]. The linguistic complexity of some text annotation tasks may not be immediately obvious or difficult to map into a game domain. Additionally, it may be challenging to find a representation that appeals to the users both in terms of entertainment and understanding.

In this work we present the game *TileAttack*[1]. TileAttack is designed to gather "mentions", a crucial step of the coreference resolution pipeline which discovers potential referring expressions including noun-phrases and possessive pronouns [5]. The following example shows the nested mentions enclosed in braces (taken from the Phrase Detectives corpus) [2]:
{A Wolf} had been gorging on {an animal {he} had killed}

Aside from text segmentation, *TileAttack* is designed with consideration to providing a controlled and highly configurable platform that attempts to answer questions on how the design of such games can be improved. For example, *do players prefer playing this type of game under time constraints?* Although it has long been hypothesised that time constraints provide a compelling and fun mechanic for games [6], this may not necessarily be the case with GWAPs for NLP [8]. When annotating text it is necessary to consider a sentence or possibly even wider context. Users may not appreciate being judged on their ability to read and consider that context. Furthermore, this may be detrimental to annotation quality.

Historically, the success of GWAPs has largely been measured based on their accuracy, with few figures published with regards to player engagement. However, just as important in GWAPs, is their ability to attract players and retain them. As a consequence, at the moment, it can be challenging to determine the true success of a game in this respect, or compare a GWAP with its counterparts. In this work, metrics from Free-To-Play (F2P) games are used with the goal of painting a clear and easily comparable picture of how the game performed in engaging players.

[1] http://tileattack.com

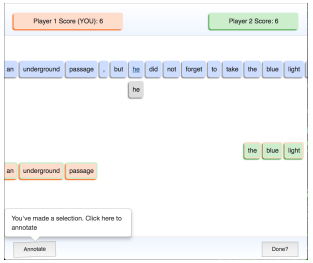

Figure 1: Game screenshot

Related Work

One very successful example of gamification of text annotation for an NLP task is Phrase Detectives [8] in which players annotate and validate anaphora. Aside from its game like detective theme, the game borrows design concepts from games such as points, leaderboards and levels.

Puzzle Racer is a recent example of a GWAP for an NLP annotation task. *Puzzle Racer* is an image-sense annotation game in which players tie images to word senses by racing through a series of gates, attempting to pass through gates that match a certain word sense [3]. Whilst a great example of a GWAP for NLP annotation, the task itself has an image component that leaves the task not too far from being image labelling rather than a typical NLP annotation task and didn't achieve a number of judgements that would be feasible for large scale annotation.

A Freemium or Free to Play game takes payment from the user during the game, typically in small "micro-transactions" in exchange for further features or functionality [1]. They are similar to GWAPs in that both have the common goal of requiring the user to continue to play to gain a return on their original investment, making F2P metrics suitable for evaluating the performance of GWAPs. To tailor the monetization strategy in a F2P game, the game is evaluated using a set of Key Performance Indicators (KPI) [1] that look not only at cost, but player engagement. Examples of player engagement metrics include the *Monthly Active Users*, and *Retention Rate*, that measures how many players continued to play over a given time period.

Despite their apparent relevance, such metrics are rarely reported in GWAPs. Von Ahn et al discuss the benefits of evaluation of GWAPs, and propose three measures. Throughput (annotated items per hour), Average Lifetime Play (average time a player spends playing) and Expected Contribution (throughput multiplied by ALP) [16]. Additionally, a detailed cost analysis is performed by Vanella et al of their GWAPs, *The Knowledge Towers* and *Infection* [14]. These metrics are similar to the aforementioned F2P metrics, particularly Throughput, but there is no universal standard for GWAPs.

TileAttack

TileAttack is a web-based two player blind game in which players are awarded points based on player agreement of the tokens they mark. The visual design of the game is inspired by *Scrabble*, with a tile like visualisation (shown in Figure 1).

In the game, players perform a text segmentation task which involves marking spans of tokens represented by tiles.

The game is designed to have variable rules and objectives between experiments with the goal of discovering the effect they have. Aside from being able to share or like the game on Facebook, there is further integration that allows players to log in to the game via Facebook.

Before being taken to the game, players are shown a short introduction that includes an explanation of the items they will be marking, the interface, the controls and properties of the game unique to the specific experiment taking place. For example, when there is a timer, they are told how long they will have.

Our approach was to start with a game design that begins from as close as possible to an existing working recipe. We chose a design that is in many respects analogous to *The ESP Game*, but for text annotation. This provides the opportunity to test what lessons learned from games similar to *The ESP Game* still apply with text annotation games, and

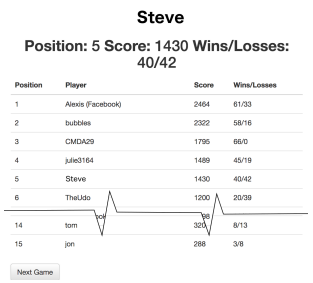

Figure 2: Leaderboard (midsection cut for brevity)

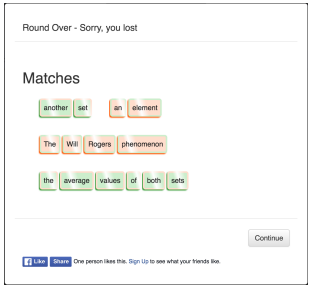

Figure 3: End of round summary

how, in the domain of text annotation, these lessons can be expanded upon. Like *The ESP Game* we use the "output-agreement" format for the game, in which two players or agents are paired, and must produce the same output, for a given input [16].

Interface

The game deliberately omits any specific design themes that may appeal to a subset of the players in order to focus only on the game mechanics being tested. This clean Scrabble inspired template for the game provides a canvas for future experiments relating to individual user personalisation or theming the game in line with current trends (e.g. spaceships, zombies, football).

A mobile-first responsive interface has been used with quick methods of interaction. Selections can be made with minimal taps over large tiles to make it easy to tap the tiles on a mobile device. The sentence can be scrolled on the phone by swiping to the left or right. When displayed on a small portrait screen the scores resize and are stacked vertically.

Great care has been taken in the selection and application of visual game design concepts to effectively communicate operation of the game through the interface using multiple channels including colour, object movement and text. For example, items that are in an interactive state display a subtle animated wobbling effect. This can be seen when the player makes a selection in the preview selection bar and buttons, when appropriate. Consequences of positive actions are shown using a horizontally moving glinting effect (Figure 3). This can be seen when the players match moves. A simple colour scheme provides context to the user as to which aspects of the game relate to them, and which relate to their opponent.

Gameplay

In each round, the player is shown a single sentence to annotate. The players can chose to select a span from the sentence by simply selecting the start and end token of the item they wish to mark using the blue selection tokens. A preview of their selection is then shown immediately below. To confirm this annotation, they may either click the preview selection or click the *Annotate* button. The annotation is then shown in the player's colour. When the two players match on a selection, the tiles for the selection in agreement are shown with a glinting effect, in the colour of the player that first annotated the tiles and a border colour of the player that agreed. The players' scores are shown at the top of the screen.

Players receive a single point for marking any item. If a marked item is agreed between the two players, the second player to have marked the item receives the number of points that there are tokens in the selection, and the first player receives double that amount. The player with the greatest number of points at the end of the round wins.

When a player has finished, they click the *Done* button, upon which they will not be able to make any more moves, but will see their opponents moves. Their opponent is also notified they have finished and invited to click *Done* once they have finished. Once both players have clicked *Done*, the round is finished and both players are shown a round summary screen (Figure 3). This screen shows the moves that both players agreed on, and whether they won or lost the round.

Clicking *Continue* then takes the player to a leaderboard (Figure 2), where they are shown their current position, score, wins, losses and the current top fifteen players. From this page they may click the *Next Game* button, to start another round.

Player Recruitment Sources

- *Reddit /r/LanguageTechnology* - a board known as a "subreddit" of a publicly viewable bulletin board style system dedicated to the topic of Language Technology. At the time of posting, the board had 6,633 subscribers [9];
- *Reddit /r/gamification* - another "subreddit", but on the topic of Gamification. 934 subscribers at the time of posting [10];
- *Facebook* - A closed Facebook group for PhD students research games with approximately 60 subscribers;
- *Corpora Mailing List* - A mailing list managed by the University of Bergen. No subscriber count is published [13];
- *YouTube* - A video was posted to YouTube demonstrating the interface. The video had 47 views at the time of writing [19].

Pilot Experiment

Task

This experiment will test if players prefer playing this type of game under time constraints. In this game, players mark "mentions". These entities would normally be collected by a mention detection system and are typically used as part of larger NLP pipelines such as relation extraction systems or co-reference resolution systems [5]. To determine how successfully players are annotating the corpus, they are given sentences from the gold standard Phrase Detectives corpus [2] to annotate.

Experimental Design

Players are split into two groups (A and B) evenly upon registration, alternating, in the order they arrive. The Group A players have a 3 minute time limit on their round length, after which the round will finish automatically. This limit has been chosen so as not to drastically impact the available time they have to complete the game, but study the effects of the presence of the timer. Whilst playing, the remaining time is displayed at the top of the screen. Group B has no timer and may play each round for as long as they please. All players play against an AI opponent.

Links to the game were posted to a combination of public and private places. Taking a sum of the sources for which the number of subscribers is published, the approximate combined audience reach is 7,675. No financial incentives or paid marketing were used to attract players.

Results

The results discussed here take a snapshot of the period from the 1st March 2017, to the 31st March 2017. Excluded are, incomplete games, games played by the game's creator, and results from users that did not complete any games. This is a total of 654 games and 46 users.

	Both Groups	Group A	Group B
Lifetime Judgements (LTJ)	14.22 sd(20.9)	12.5 sd(21.85)	15.79 sd(19.85)
Average Judgements per Player (AJpP)	8.84 sd(10.49)	8.09 sd(10.49)	9.48 sd(10.44)
Average Lifetime Play (mm:ss)	08:42	07:36	09:38
Monthly Active Users (MAU)	46	22	24
Retention (per day) (%)	8 (17.39%)	4 (18.18%)	4 (16.67%)
Throughput (per hour)	0.90	0.38	0.53

Table 1: F2P Metrics

Player Engagement

Table 1 shows the evaluation of the game using the adapted F2P metrics. The timing of the games does not have a statistically significant result (LTJ, $p = 0.6$, unpaired t-test; AJpP, $p = 0.57$, unpaired t-test). Players did not regularly reach the time limit. The overall average time spent on a game round was 42 seconds (rounded to the nearest second), and only 13 game rounds lasted longer than 2 minutes. On average, players spent 8:42 (mins:secs) playing per session. In comparison, the well developed *Verbosity* game achieved an average session length of 23.48 minutes [17]. Eight players in total returned after 24 hours to play again.

Annotation Quality

The player's annotations are compared with that from the expert annotated Phrase Detectives corpus [2]. This corpus provides expert annotated data as corrections to an

Adapted F2P Metrics
Lifetime Judgements (LTJ) is the average number of sentences annotated per player over their lifetime of play. *Average Judgements per Player (AJpP)* is the average number of sentences marked per player, per gaming session. *Average Lifetime Play* is the average session length in time. *Monthly Active Users (MAU)* is the number of users in a month, the active part refers specifically to those that finished a game. *Retention and churn* is the players that were kept and lost respectively, over some time period (in this work we use 24 hours). *Throughput* is the annotations received over some selected time period (in this work we use an hour).

	Both Groups	Group A (timed)	Group B
Precision	0.566	0.569	0.564
Recall	0.594	0.602	0.587
F-Measure	0.553	0.557	0.551
Games	363	143	220
Players	42	21	21

Table 2: Annotation quality against expert annotation. Average (mean) precision and recall over each game

automated pipeline. The game does not attempt to apply the corrections from the corpus. This analysis of annotation quality uses a subset of the sentences that were expert approved without requiring corrections. There was a total of 363 (of the 654 games) of these games played on these sentences specifically, by 42 of players (21 in each group).

Whilst Table 2 does not show high annotation quality, it would appear that players understood the task and that the game is effective in identifying markable items. There are multiple opportunities to filter out some of the data to raise precision and recall further. One example, is that in 9 games, players click the *Done* button without making a single annotation on sentences that have many moves available. With further player guidance, filtering of problem players, application of aggregation methods and validation, the system may produce useful annotation data in future.

Future Work

TileAttack will soon be distributed to a wider audience via marketing (e.g. online advertising). The greater sample of players will be used, to begin with, to test the other questions. For example, whether players prefer turn based play, and the effects of different reward policies. Future work will use the F2P metrics that analyse that cost in relation to return. These will include, *Cost Per Action/Acquistion/Conversion*, that may be used to measure the cost of acquiring a player or having them complete a unit of work. This will allow for direct comparison with the costs involved in other methods of soliciting annotation, such as crowdsourcing.

In response to the lower than expected precision and recall figures, two new game mechanics will be added. To dissuade players clicking randomly and raise precision, moves that are known to be invalid will result in the user receiving a negative score. To encourage users to not just select the obvious items and increase recall, selections that are commonly made will be highlighted as unavailable in the game.

Conclusion

This paper presented a highly configurable web based and mobile friendly game created for the purpose of gathering annotations of mentions that feature in co-reference chains and testing different game mechanics. More specifically, this experiment utilised the game to test the effects of applying time constraints. This work also applied a new approach to evaluation of NLP GWAP player engagement by applying F2P metrics.

This prototype achieves an average session length of 8:42 (mins:secs) on the very first experiment without players receiving any financial incentive on a challenging text annotation task, showing promise for future experiments on the same platform.

Whilst the experiment failed to produce significant result, it did demonstrate the games ability to to function in an experimental setting. We hope the planned future game mechanics played by a large sample of players may well more significant results.

REFERENCES

1. David Xicota. 2014. Free to play and its Key Performance Indicators. (2014). http://www.gamasutra.com/blogs/DavidXicota/20140527/218550/Free_to_play_and_its_Key_Performance_Indicators.php
2. Jon Chamberlain et al. 2016. Phrase Detectives Corpus 1.0 Crowdsourced Anaphoric Coreference.. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (23-28). European Language Resources Association (ELRA), Paris, France.
3. David Jurgens and Roberto Navigli. 2014. It's all fun and games until someone annotates: Video games with a purpose for linguistic annotation. *TACL* 2 (2014), 449–464.
4. Mathieu Lafourcade, Alain Joubert, and Nathalie Le Brun. 2015. *Games with a Purpose (GWAPS)*. John Wiley & Sons.
5. Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*. Association for Computational Linguistics, 28–34.
6. Thomas W Malone. 1982. Heuristics for designing enjoyable user interfaces: Lessons from computer games. In *Proceedings of the 1982 conference on Human factors in computing systems*. ACM, 63–68.
7. Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics* 31, 1 (2005), 71–106.
8. Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi. 2013. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM TiiS* 3, 1 (2013), 3.
9. Reddit. 2017a. TileAttack - Building an NLP model with a game : LanguageTechnology. (2017). https://www.reddit.com/r/LanguageTechnology/comments/5wvfm9/tileattack_building_an_nlp_model_with_a_game/
10. Reddit. 2017b. TileAttack - gamification for helping build natural language processing models : gamification. (2017). https://www.reddit.com/r/gamification/comments/5wvcqh/tileattack_gamification_for_helping_build_natural/
11. Seth Cooper et al. 2010. Predicting protein structures with a multiplayer online game. *Nature* 466, 7307 (2010), 756–760.
12. 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 the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 254–263.
13. University of Bergen. 2017. Corpora. (2017). http://mailman.uib.no/listinfo/corpora
14. Daniele Vannella, David Jurgens, Daniele Scarfini, Domenico Toscani, and Roberto Navigli. 2014. Validating and Extending Semantic Knowledge Bases using Video Games with a Purpose.. In *ACL (1)*. 1294–1304.

15. Luis Von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *SIGCHI*. ACM, 319–326.

16. Luis Von Ahn and Laura Dabbish. 2008. Designing games with a purpose. *Commun. ACM* 51, 8 (2008), 58–67.

17. Luis Von Ahn, Mihir Kedia, and Manuel Blum. 2006. Verbosity: a game for collecting common-sense facts. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 75–78.

18. Winter, Mason et al. 2010. Financial incentives and the performance of crowds. *ACM SigKDD Explorations Newsletter* 11, 2 (2010), 100–108.

19. YouTube. 2017. TileAttack - YouTube. (2017). `https://www.youtube.com/watch?v=fcmrsPkiMvA&feature=youtu.be`