# Testing game mechanics in games with a purpose for NLP applications

**Chris Madge, Udo Kruschwitz, Jon Chamberlain, Richard Bartle, Massimo Poesio**
School of Computer Science and Electronic Engineering
University Of Essex
{cjmadg,udo,jchamb,rabartle,poesio}@essex.ac.uk

## Abstract

This paper describes a highly configurable game-with-a-purpose (GWAP) designed to explore the effects of different game mechanics on GWAP for NLP problems with a view to improving both quality of annotation and player uptake. The details of the game are discussed along with some of the questions the game hopes to answer.

## Introduction

GWAPs have been successful in many applications attracting large numbers of users to label datasets and solve real world problems. Examples include games such as image labelling with *The ESP Game* (Von Ahn and Dabbish, 2004), or protein folding with *FoldIt* (Seth Cooper et al, 2010; Firas Khatib et al, 2011). Gamification has worked well in text (e.g. *Phrase Detectives* (Poesio et al., 2013; Chamberlain et al., 2008)), 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, moving away from gamification applications has potential for greater rewards 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 (Jurgens and Navigli, 2014). However, they have yet to achieve a player uptake or number of judgements comparable to GWAPs in other domains. GWAPs in text often present additional unique challenges compared to those around image labelling and other similar tasks e.g. users can differentiate between images easily, not so easily with text (Winter, Mason et al, 2010). 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.

The motivation for the development of this game is to provide a controlled and highly configurable platform that presents a valid GWAP to answer questions on how such games can be improved. For example, such questions may include:

- do players prefer playing this type of game under time constraints?
- do players prefer turn based play (like chess)?
- how are accuracy and play are affected by different reward policies?

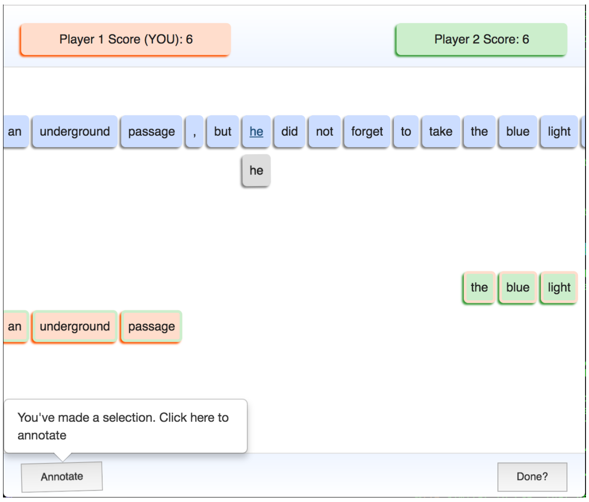
## Methodology

*Tile Attack* is a two player blind game in which players are awarded points based on player agreement on tokens that they identify as being noun-phrases. The design of the game is inspired by scrabble with a tile like visualisation shown in figure 1. The game includes a point system and leaderboard that is shown to the player between rounds.

As by means of a control, the game design starts from as close as possible to a working recipe yielding a game that is in many respects analogous to *The ESP Game*, but for text annotation, testing what lessons learned from games similar to *The ESP Game* still apply with text annotation games, and how, in the domain of text annotation, these lessons can be expanded upon.

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, if there is a timer, they are told how long they will have.

Similar to *The ESP Game*, that uses recorded player moves in the event that a suitable player is not available, this game also uses an AI (artificial intelligence) as a substitute. In its most basic form, this will make known moves from a gold standard and will slow its pace slightly for beginner players.



You've made a selection. Click here to annotate

Figure 1: Game Screenshot

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 simplistic un-themed template for the game provides a plain canvas for future experiments relating to individual user personalisation (where they may express themselves by applying their own designs) or theming the game in line with current trends (e.g. spaceships, zombies, football).

Designed to be mobile first and work at a variety of resolutions, the players interact with the interface by simply selecting the start and end token of the item they wish to mark using the blue selection tokens. That selection is then shown in grey immediately beneath. To make that an annotation, they may either click the grey selection that is shown, or click the “Annotate” button. The annotation is then shown in the player’s colour. When a match is made, the tiles are shown in the colour of the player that first annotated the tiles, with a border surround colour of the player that agreed. The players’ scores are shown at the top of the screen.

When the game is complete the player may click the “Done” button to finish the round.

The game is designed to have variable rules and objectives between experiments with the goal of discovering the effect they have. Players will be split into two groups, and A/B testing carried out to investigate different variables such as adding a time constraint, enforcing turn based play, and using different reward policies.

Throughout experiments, various metrics will be measured, including accuracy of the players annotations against the gold standard, the number of judgements they make, the speed of their annotations and the number of games they choose to play. Once they have finished playing, players will be asked to take a usability survey.

## Conclusion

This work has discussed a game prototype that hopes to serve as a suitable base for asking many questions about how using GWAPs for NLP can be improved.

Through an iterative development and experiment cycle, coupled with deployment in a large scale online setting the goal is to both gather NLP data and refine an improved formula for development of other games.

## Acknowledgements

This research was supported in part by the EPSRC CDT in Intelligent Games and Game Intelligence (IGGI), EP/L015846/1; in part by the DALI project, ERC Grant 695662.

## References

Jon Chamberlain, Massimo Poesio, and Udo Kruschwitz. 2008. Phrase detectives: A web-based collaborative annotation game. In *I-Semantics*.

Firas Khatib et al. 2011. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature structural & molecular biology*, 18(10):1175–1177.

David Jurgens and Roberto Navigli. 2014. It’s all fun and games until someone annotates: Video games with a purpose for linguistic annotation. *TACL*, 2:449–464.

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):3.

Seth Cooper et al. 2010. Predicting protein structures with a multiplayer online game. *Nature*, 466(7307):756–760.

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

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