

Novel Incentives for Phrase Detectives

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz and Chris Madge

University of Essex, Language and Computation Group

Abstract

The *Phrase Detectives* Game-With-A-Purpose for anaphoric annotation is a moderately successful example of use of novel incentives to create resources for computational linguistics. In this paper we summarize the *Phrase Detectives* experience in terms of incentives and discuss our future plans to improve such incentives.

1. Introduction

Phrase Detectives (Chamberlain et al., 2008; Poesio et al., 2013) an interactive online **game with a purpose** (von Ahn, 2006) for creating anaphorically annotated corpora through web collaboration, is a moderately successful example of use of novel incentives to create resources for computational linguistics. *Phrase Detectives* has been live since December 2008, collecting almost 3 million judgments on the anaphoric expressions in texts in two languages (English and Italian) from over 40,000 players, resulting in a corpus of over 500 documents and over 300,000 tokens. In this paper we briefly discuss the incentives provided by *Phrase Detectives*, assess their contribution, and discuss future work to address some of the current shortcomings. For further discussion of the incentive structure in *Phrase Detectives* and a more detailed evaluation, see (Chamberlain et al., 2009; Chamberlain et al., 2012; Chamberlain, 2016)

2. A Brief Description of the Game

Phrase Detectives is a single-player GWAP developed to collect data about English (and subsequently Italian) anaphoric reference (Poesio et al., 2013) The game architecture is articulated around a number of tasks and uses scoring, progression and a variety of other mechanisms to make the activity enjoyable. The game design is based on a detective theme, relating to the how the player must search through the text for a suitable annotation (Chamberlain et al., 2008).

The players have to carry out two different tasks. Initially text is presented in Annotation Mode (called *Name the Culprit* in the game - see Figure 1). This is a straightforward annotation mode where the player makes an annotation decision about a highlighted **markable** (section of text). (The annotation scheme used in *Phrase Detectives* is a simplified version of the anaphoric annotation scheme used in the ARRAU corpus (Poesio and Artstein, 2008).)

If different players enter different interpretations for a markable then each interpretation is presented to more players in Validation Mode (called *Detectives Conference* in the game). The players in Validation Mode have to agree or disagree with the interpretation.

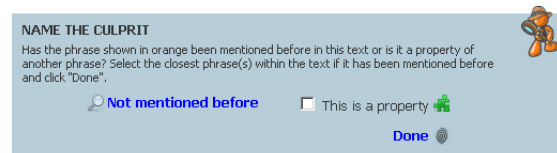
Players are trained with texts from a gold standard. Players always receive a training text when they first start the game. Once the player has completed all of the training tasks they are given a rating (the percentage of correct decisions out of the total number of training tasks). If the rating is above a certain threshold (currently 50%) the player progresses on

Rhinogradentia (Wikipedia)

Rhinogradentia (also known as snouters or Rhinogrades or Nasobarnes) is a fictitious mammal order documented by the equally fictitious German naturalist Harald Stumpke. The order's most remarkable characteristic was the Nasorium, an organ derived from the ancestral species's nose, which had variously evolved to fulfill every conceivable function.

Both the animals and the scientist were allegedly creations of Gerolf Steiner, a zoology professor at the University of Karlsruhe. A mock taxidermy of a certain Snouter can be seen at the Musee zoologique in Strasbourg.

The order's remarkable variety was the natural outcome of evolution acting over millions of years in the isolated Hi-yi-yi islands in the Pacific Ocean.



- Comment on this phrase
- Skip this one
- Skip - closest phrase can't be selected
- Skip - closest phrase is no longer visible
- Skip - error in the text

Figure 1: Detail of a task presented in Annotation Mode.

to annotating real documents, otherwise they are asked to do a training document again. The rating is recorded with every future annotation that the player makes as the rating is likely to change over time. The scoring system is designed to reward effort and motivate high quality decisions by awarding points for retrospective collaboration. A mixture of incentives, from the personal (scoring, levels) to the social (competing with other players) to the financial (small prizes) are employed.

The goal of the game was not just to annotate large amounts of text, but also to collect a large number of judgments about each linguistic expression. This led to the deployment of a variety of mechanisms for quality control which try to reduce the amount of unusable data beyond those created by malicious users, from the level mechanism itself to validation to a number of tools for analysing the behavior of players.

A Facebook version of *Phrase Detectives*,¹ launched in February 2011, makes full use of socially motivating factors inherent in the Facebook platform (Chamberlain et al., 2012). For instance, any of the player's friends who are playing the game form the player's team, which is visible in the left hand menu. Whenever a player's decision agrees with a team member they score additional points. The most

¹<http://apps.facebook.com/phrasedetatives>

interesting finding from this work is that although fewer players play it, the quality and quantity of their work is significantly superior to that of the players of the original game; more in general, knowing the identity of the player leads to much better quality (Chamberlain, 2016).

Phrase Detectives is one of the most successful GWAPs for computational linguistics. Started in December 2008, it is still being played. As of April 2016, over 40,000 players have registered; of these, 4,000 passed the training phase—around 1,000 of which on *Facebook Phrase Detectives*. Over 2.3 million annotation judgments have been collected and 466,000 validations. 549 documents have been completely annotated for a total of around 330,000 words (the complete corpus will be of 1.2 million words). These annotations are being turned into a publically available corpus (Chamberlain et al., 2016).

3. Incentives in Phrase Detectives

The primary incentives in a GWAP for collective resource creation are enjoyment and scientific interest, but we experimented with a number of other types incentives as well. We discuss each in turn.

3.1. Enjoyment

The primary motivation for someone to use *Phrase Detectives* is supposed to be enjoyment: having fun while playing the game. The game was thus designed to incorporate several mechanisms that are meant to make a game fun (Koster, 2005). One of the simplest such mechanisms is **scoring**: by getting a score the player gains a sense of achievement. A second common method to entertain players is to have them experience a **progression through the game**, whether by learning new types of tasks, becoming more proficient at current tasks, or gaining recognition for their effort (see below). A common form of progression is by assigning the player a named **level**, starting from novice and going up to expert (Koster, 2005; von Ahn et al., 2006). (Although we will not discuss quality control here, the level mechanism also provides one form of quality control.) Last but not least, great care was taken in **choosing texts** to annotate that players would find interesting, helped in this by the decision to concentrate on text genres that are under-used in computational linguistics, in particular fiction. We also included a number of documents from Wikipedia, but all chosen for their quirkiness.

3.2. Design

When designing any interface it is essential to know your target audience. Individual, social and socio-technical factors will all determine how successful the interface is at engaging users and what type of data will be contributed. We believe that a key part of the success of *Phrase Detectives* is due to the attractive design of its interface. Game interfaces should be graphically rich, although not at the expense of usability, and aimed at engaging a specific audience (i.e., a game aimed at children may include more cartoon or stylised imagery in brighter colours than a game aimed at adults). Interfaces should also provide a consistent metaphor and work flow. *Phrase Detectives* used a detective metaphor, with buttons stylised with a cartoon detective

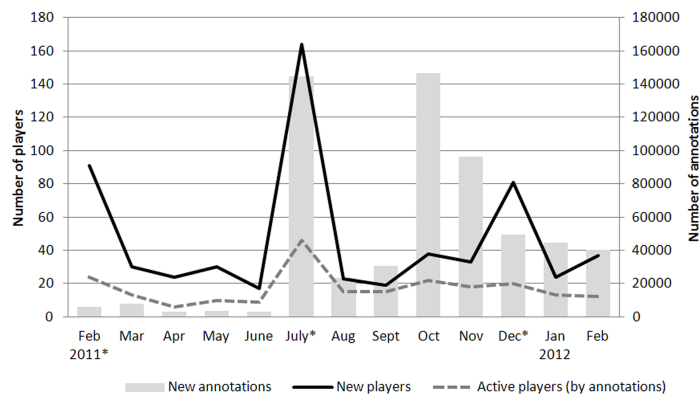


Figure 2: Chart showing the effect of prizes on the workload of *Phrase Detectives* players.

character and site text written as if the player was a detective solving cases. The tasks should be integrated in such a way that task completion, user evaluation and work flow form a seamless experience.

3.3. Contributing to Science

An important incentive for players of GWAPs is the opportunity to participate in a project producing something of relevance to a (scientific) community. This type of incentive did play a role in attracting players to *Phrase Detectives* and retaining them: many of the players of the game are computational linguists who heard about the game through presentations and lectures, or thanks to the mention of *Phrase Detectives* in computational linguistics blogs with a substantial following such as those by Mark Liberman² or Bob Carpenter.

3.4. Prizes

Offering substantial direct payment to the players would defeat the purpose of using GWAPs to reduce the cost of generating resources. But a very low-cost reward structure can be built into online games through the mechanism of **prizes**. In *Phrase Detectives* a variety of prizes in the form of Amazon vouchers for a maximum value of £50 have often been offered. Prizes for high scoring players will motivate hard working or high quality players but the prize soon becomes unattainable for the majority of other players. We also offered therefore lottery style financial prizes, whose winner is randomly selected. In this way the hardest-working players are more likely to win, but the players who only do a little work are still motivated. These prizes have proven extremely effective. Figure 2 shows the effect of prizes on *Facebook Phrase Detectives*. Months where there was active promotion of the site via prizes (February, July and December 2011) show substantial increases in new players, annotations, and active players.

3.5. Social Incentives

A different sort of social incentive is provided by the scoring mechanism. **Public leaderboards** reward players by

²<http://languagelog.ldc.upenn.edu/n11/?p=2050>

improving their standing amongst their peers (in this case their fellow players). Using leaderboards and assigning levels for points has been proven to be an effective motivator, with players often using these as targets (von Ahn and Dabish, 2008). An interesting phenomenon has been reported with these reward mechanisms, namely that players gravitate towards the cutoff points (i.e. they keep playing to reach a level or high score before stopping) (von Ahn et al., 2006).

Both types of social incentives can be made even more effective when the game is **embedded in a social networking platform** like Facebook. In such a setting, the players motivated by the desire to contribute to a communal effort may share their efforts with their friends, whereas those motivated by a competitive spirit can compete against them. This was one of the motivations behind the Facebook version of *Phrase Detectives*.

4. Beyond Phrase Detectives: the DALI Project

The incentives to annotation provided by *Phrase Detectives* could already be defined as having been reasonably successful. The game has motivated a reasonable number of players to annotate a corpus of respectable size. And the corpus already has a significant advantage in comparison with other existing corpora in terms of judgments per markable, with over 20 judgments per markable on average. This said, the ambitions motivating the development of a GWAP are much higher both in terms of number of players (some of von Ahn's games attracted over 100,000 players) and in terms of corpus size (our ambition is to fully annotate over 100 million words). In the soon-to-start DALI project, a collaboration between the University of Essex and LDC funded by ERC, we intend to improve the current incentive structure in a number of ways.

4.1. Making the game more enjoyable

Although many current players enjoy the game, most of those tend to do so because they are interested in the linguistics of anaphora or find the texts quirky, rather than because they find the game enjoyable. Our first objective in DALI will be to develop a new game, or games, which are genuinely enjoyable. Among the ideas we intend to pursue is incorporating in our games a stronger sense of progression, by providing intrinsic rewards to players that achieve a higher status such as the ability to choose more interesting icons for higher status players. We will also develop more attractive ways for players to express their judgments (e.g., clicking on icons associated with discourse entities). We also intend to make smartphones the main platform through which to play the games. While the main motivation for this move is increasing their accessibility, we expect it to make them more enjoyable as well.

4.2. Increased interaction with the computational linguistic community

As mentioned above, a great deal of the success of *Phrase Detectives*, particularly in the beginning, was due to the contribution of the computational linguistics community,

both in popularizing the game through blogs and in actually playing it. We intend to extend the collaboration with the community in collaboration with LDC, both by embedding the game in their future portal for community-created games, and by relying on their expertise in releasing annotated resources.

4.3. Educational Incentives

It can be argued that the most attractive aspect of the current version of *Phrase Detectives* is what it teaches its players about anaphora and its intricacies. This suggests that the game could find a use in teaching language. We intend to test this hypothesis in collaboration with the International Academy at the University of Essex, whose objective is to remedy any language skills shortcomings of future University of Essex students. To this purpose, they offer a variety of language courses that students can take prior to their starting their studies. These courses use a variety of computer-based practice exercises, including games. We recently piloted using *Phrase Detectives* as one of these practice games. We intend to continue and intensify this collaboration.

5. Conclusions

Games with a purpose can serve as a useful alternative for corpus annotation—in fact, as the only viable option when the aim is to create truly large-scale resources (Poesio et al., In press). But in order to realize this potential, sufficient players have to be enrolled through attractive incentives. The first years of the *Phrase Detectives* experience have taught us a lot about what works and what doesn't; we hope to take advantage of these lessons to develop new games that allow us to achieve our objective of creating truly large-scale annotated corpora for computational linguistics.

6. Acknowledgements

The initial funding for *Phrase Detectives* (2007/09) came from UK EPSRC project AnaWiki, EP/F00575X/1. Subsequent funding came from an EPSRC PhD studentship for Jon Chamberlain as well as from the EU project SENSEI. Funding for Chris Madge comes from the IGGI Doctoral Training Centre, funded by EPSRC. DALI will be funded by the European Research Council, ERC.

7. References

- J. Chamberlain, M. Poesio, and U. Kruschwitz. 2008. Phrase Detectives: A Web-based Collaborative Annotation Game. In *Proceedings of the International Conference on Semantic Systems (I-Semantics'08)*, Graz.
- J. Chamberlain, M. Poesio, and U. Kruschwitz. 2009. A new life for a dead parrot: Incentive structures in the Phrase Detectives game. In *Proceedings of the WWW 2009 Workshop on Web Incentives (WEBCENTIVES'09)*, Madrid.
- J. Chamberlain, U. Kruschwitz, and M. Poesio. 2012. Motivations for participation in socially networked collective intelligence systems. In *Proceedings of CI2012*.
- J. Chamberlain, M. Poesio, and U. Kruschwitz. 2016. Phrase detectives corpus 1.0: Crowdsourced anaphoric coreference. In *Proc. of LREC*, Portoroz, Slovenia.
- J. Chamberlain. 2016. *Harnessing Collective Intelligence on Social Networks*. Ph.D. thesis, University of Essex, School of Computer Science and Electronic Engineering.
- R. Koster. 2005. *A Theory of Fun for Game Design*. Paraglyph.
- M. Poesio and R. Artstein. 2008. Anaphoric annotation in the arrau corpus. In *Proceedings of the sixth International Conference on Language Resources and Evaluation*, Marrakesh, May.
- M. Poesio, J. Chamberlain, U. Kruschwitz, L. Robaldo, and L. Ducceschi. 2013. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Intelligent Interactive Systems*, 3(1).
- M. Poesio, J. Chamberlain, and U. Kruschwitz. In press. Crowdsourcing. In N. Ide and J. Pustejovsky, editors, *The Handbook of Annotation*. Springer.
- L. von Ahn and L. Dabbish. 2008. Designing games with a purpose. *Communications of the Association for Computing Machinery (ACM)*, 51(8):58–67.
- L. von Ahn, R. Liu, and M. Blum. 2006. Peekaboom: a game for locating objects in images. In *Proceedings of conference on Human Factors in computing systems*, pages 55–64.
- L. von Ahn. 2006. Games with a purpose. *Computer*, 39(6):92–94.