A new life for a dead parrot: Incentive structures in the Phrase Detectives game

Jon Chamberlain
University of Essex
School of Computer Science
and Electronic Engineering
jchamb@essex.ac.uk

Massimo Poesio
University of Essex
School of Computer Science
and Electronic Engineering
poesio@essex.ac.uk

Udo Kruschwitz
University of Essex
School of Computer Science
and Electronic Engineering
udo@essex.ac.uk

He's passed on! This parrot is no more! He has ceased to be! He's expired and gone to meet his maker! He's kicked the bucket, he's shuffled off his mortal coil, run down the curtain and joined the bleedin' choir invisibile! THIS IS AN EX-PARROT! [1]

ABSTRACT

In order for there to be significant improvements in certain areas of natural language processing (such as anaphora resolution) large linguistically annotated resources need to be created which can be used to train, for example, machine learning systems. Annotated corpora of the size needed for modern computational linguistics research cannot however be created by small groups of hand-annotators. Simple Web-based games have demonstrated how it might be possible to do this through Web collaboration. This paper reports on the ongoing work of *Phrase Detectives*, a game developed in the ANAWIKI project designed for collaborative linguistic annotation on the Web. In this paper we focus on how we recruit and motivate players, incentivise high quality annotations and assess the quality of the data.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human factors; Human information processing; H.2.7 [**Artificial Intelligence**]: Natural Language Processing

Keywords

Web-based games, incentive structures, user motivation, distributed knowledge acquisition, anaphoric annotation

1. INTRODUCTION

The statistical revolution in natural language processing (NLP) has resulted in the first NLP systems and components which are usable on a large scale, from part-of-speech (POS) taggers to parsers [7]. However it has also raised the problem of creating the large amounts of annotated linguistic data needed for training and evaluating such systems. Potential solutions to this problem include semi-automatic annotation and machine learning methods that make better use of the available data. Unsupervised or semi-supervised techniques hold great promise, but for the foreseeable future at least, the greatest performance improvements are still likely to come from increasing the amount of data to be used by supervised training methods. These crucially rely on hand-annotated data. Traditionally, this requires trained annotators, which is prohibitively expensive both financially and in terms of person-hours (given the number of trained annotators available) on the scale required.

Recently, however, Web collaboration has emerged as a viable alternative. Wikipedia and similar initiatives have shown that a surprising number of individuals are willing to help with resource creation and scientific experiments. The Open Mind Common Sense project [16] demonstrated that such individuals are also willing to participate in the creation of databases for Artificial Intelligence (AI), and von Ahn showed that simple Web games are an effective way of motivating participants to annotate data for machine learning purposes [23].

The goal of the ANAWIKI project[1] is to experiment with Web collaboration as a solution to the problem of creating large-scale linguistically annotated corpora, both by developing Web-based annotation tools through which members of the scientific community can participate in corpus creation and through the use of game-like interfaces. We will present ongoing work on *Phrase Detectives*[2], a game designed to collect judgments about anaphoric annotations. We will also report results which include a substantial corpus of annotations already collected.

2. RELATED WORK

Related work comes from a range of relatively distinct research communities including, among others, Computational Linguistics / NLP, the games community and researchers working in the areas of the Semantic Web and knowledge representation.

Large-scale annotation of low-level linguistic information (part-of-speech tags) began with the Brown Corpus, in which very low-tech and time consuming methods were used. For the creation of the British National Corpus (BNC), the first 100M-word linguistically annotated corpus, a faster methodology was developed consisting of preliminary annotation with automatic methods followed by partial hand-correction [1]. This was made possible by the availability of relatively high quality automatic part-of-speech taggers (CLAWS).

With the development of the first high-quality chunkers, this methodology became applicable to the case of syntactic annotation. It was used for the creation of the Penn

[1] `http://www.textfiles.com/media/petshop`

WWW2009, April 20-24, 2009, Madrid, Spain.
.

[1] `http://www.anawiki.org`

[2] `http://www.phrasedetectives.org`

Treebank [10] although more substantial hand-checking was required.

Medium and large-scale semantic annotation projects (for wordsense or coreference) are a recent innovation in Computational Linguistics. The semi-automatic annotation methodology cannot yet be used for this type of annotation, as the quality of, for instance, coreference resolvers is not yet high enough on general text. Nevertheless the semantic annotation methodology has made great progress with the development, on the one end, of effective quality control methods [4] and on the other, of sophisticated annotation tools such as Serengeti [20].

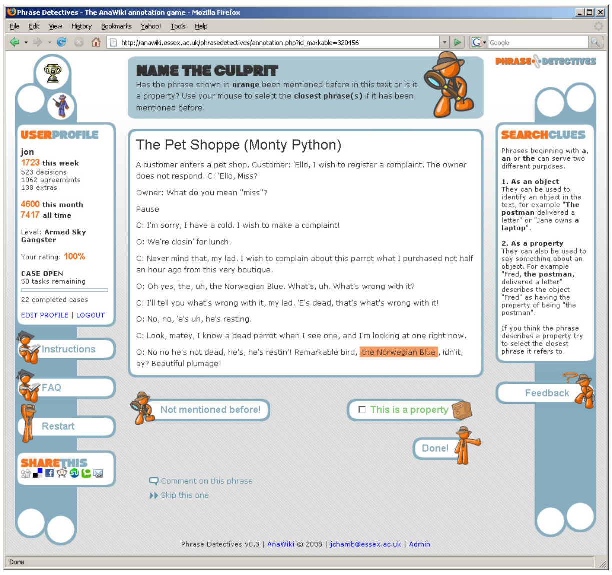
These developments have made it possible to move from the small-scale semantic annotation projects, the aim of which was to create resources of around 100K words in size [14], to the efforts made as part of US initiatives such as Automatic Context Extraction (ACE), Translingual Information Detection, Extraction and Summarization (TIDES), and GALE to create 1 million word corpora. Such techniques could not be expected to annotate data on the scale of the BNC.

Collaborative resource creation on the Web offers a different solution to this problem. The motivation for this is the observation that a group of individuals can contribute to a collective solution, which has a better performance and is more robust than an individual's solution as demonstrated in simulations of collective behaviours in self-organizing systems [6].

Wikipedia is perhaps the best example of collaborative resource creation, but it is not an isolated case. The gaming approach to data collection, termed *games with a purpose*, has received increased attention since the success of the ESP game [22]. Subsequent games have attempted to collect data for multimedia tagging (*OntoTube*[3], *Tag a Tune*[4]) and language tagging (*Verbosity*[5], *OntoGame*[6], *Categorilla*[7], *Free Association*[8]). As Wikipedia has demonstrated however, there is not necessarily the need to turn every data collection task into a game. Other current efforts in attempting to acquire large-scale world knowledge from Web users include Freebase[9] and True Knowledge[10].

The *games with a purpose* concept has now also been adopted by the Semantic Web community in an attempt to collect large-scale ontological knowledge because currently "the Semantic Web lacks sufficient user involvement almost everywhere" [17].

It is a huge challenge to recruit enough users to make data collection worthwhile and, as we will explore later, it is also important to attract the right kind of player. Previous games have attracted exceptional levels of participation such as the ESP game (13,500 players in 4 months) [22], Peekaboom (14,000 players in 1 month) [24] and OpenMind (15,000 users) [16] which encourages one to believe mass participation might be possible for similar projects.



http://anawiki.essex.ac.uk/phrasedetectives/annotation.php?id_markable=320456

Has the phrase shown in **orange** been mentioned before in this text or is it a property? Use your mouse to select the **closest phrase(s)** if it has been mentioned before.

jon
1723 **this week**
523 decisions
1062 agreements
138 extras

4600 **this month**
7417 **all time**

Level: **Armed Sky Gangster**

Your rating: 100%

CASE OPEN
50 tasks remaining

22 completed cases

EDIT PROFILE | LOGOUT

The Pet Shoppe (Monty Python)

A customer enters a pet shop. Customer: 'Ello, I wish to register a complaint. The owner does not respond. C: 'Ello, Miss?

Owner: What do you mean "miss"?

Pause

C: I'm sorry, I have a cold. I wish to make a complaint!

O: We're closin' for lunch.

C: Never mind that, my lad. I wish to complain about this parrot what I purchased not half an hour ago from this very boutique.

O: Oh yes, the, uh, the Norwegian Blue. What's, uh. What's wrong with it?

C: I'll tell you what's wrong with it, my lad. 'E's dead, that's what's wrong with it!

O: No, no, 'e's uh, he's resting.

C: Look, matey, I know a dead parrot when I see one, and I'm looking at one right now.

O: No no he's not dead, he's, he's restin'! Remarkable bird, the Norwegian Blue, idn'it, ay? Beautiful plumage!

Phrases beginning with **a**, **an** or **the** can serve two different purposes.

1. As an object
They can be used to identify an object in the text, for example "**The postman** delivered a letter" or "Jane owns **a laptop**".

2. As a property
They can also be used to say something about an object. For example "Fred, **the postman**, delivered a letter" describes the object "Fred" as having the property of being "the postman".

If you think the phrase describes a property try to select the closest phrase it refers to.

Phrase Detectives v0.3 | AnaWiki © 2008 | jchamb@essex.ac.uk | Admin

Figure 1: A screenshot of the Annotation Mode.

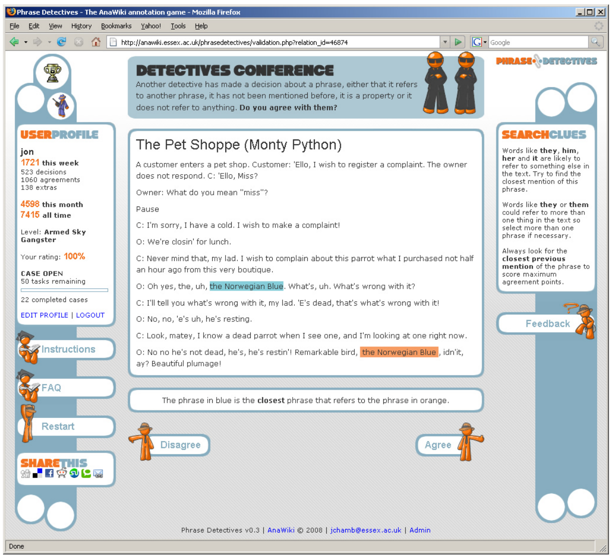
3. THE PHRASE DETECTIVES GAME

Phrase Detectives is a game offering a simple interface for non-expert users to learn how to annotate text and to make annotation decisions [2]. The goal of the game is to identify relationships between words and phrases in a short text. An example of a task would be to highlight an anaphor-antecedent relation between the markables (sections of text) *'This parrot'* and *'He'* in *'This parrot is no more! He has ceased to be!'* Markables are identified in the text by automatic pre-processing. There are two ways to annotate within the game: by selecting a markable that corefers to another one (Annotation Mode, called *Name the Culprit* in the game); or by validating a decision previously submitted by another player (Validation Mode, called *Detectives Conference* in the game).

Annotation Mode (see Figure 1) is the simplest way of collecting judgments. The player has to locate the closest antecedent markable of an anaphor markable, i.e. an earlier mention of the object. By moving the cursor over the text, markables are revealed in a bordered box. To select it the player clicks on the bordered box and the markable becomes highlighted. They can repeat this process if there is more than one antecedent markable (e.g. for plural anaphors such as *'they'*). They submit the annotation by clicking the *Done!* button. The player can also indicate that the highlighted markable has not been mentioned before (i.e. it is not anaphoric), that it is non-referring (for example, *'it'* in *'Yeah, well it's not easy to pad these Python files out to 150 lines, you know.'*) or that it is the property of another markable (for example, *'a lumberjack'* being a property of *'I'* in *'I wanted to be a lumberjack!'*). Players can also make a comment about the markable (for example, if there is an error in the automatic text processing) or skip the markable and move on to the next one.

In Validation Mode (see Figure 2) the player is presented with an annotation from a previous player. The anaphor markable is shown with the antecedent markable(s) that the previous player chose. The player has to decide if he agrees with this annotation. If not he is shown the Annotation

[3] http://www.ontogame.org/ontotube
[4] http://www.gwap.com/gwap/gamesPreview/tagatune
[5] http://www.gwap.com/gwap/gamesPreview/verbosity
[6] http://www.ontogame.org
[7] http://wordgame.stanford.edu/categorilla.html
[8] http://wordgame.stanford.edu/freeAssociation.html
[9] http://www.freebase.com
[10] http://www.trueknowledge.com



Phrase Detectives - The AnaWiki annotation game - Mozilla Firefox
File Edit View History Bookmarks Yahoo! Tools Help
http://anawiki.essex.ac.uk/phrasedetectives/validation.php?relation_id=46874
Google
PHRASE DETECTIVES
USERPROFILE
SEARCHCLUES
EDIT PROFILE | LOGOUT
Instructions
FAQ
Restart
SHARETHIS
Feedback
Disagree
Agree
Done

DETECTIVES CONFERENCE

Another detective has made a decision about a phrase, either that it refers to another phrase, it has not been mentioned before, it is a property or it does not refer to anything. **Do you agree with them?**

jon
1721 **this week**
523 decisions
1060 agreements
138 extras

4598 **this month**
7415 **all time**

Level: **Armed Sky Gangster**

Your rating: 100%

CASE OPEN
50 tasks remaining

22 completed cases

The Pet Shoppe (Monty Python)

A customer enters a pet shop. Customer: 'Ello, I wish to register a complaint. The owner does not respond. C: 'Ello, Miss?

Owner: What do you mean "miss"?

Pause

C: I'm sorry, I have a cold. I wish to make a complaint!

O: We're closin' for lunch.

C: Never mind that, my lad. I wish to complain about this parrot what I purchased not half an hour ago from this very boutique.

O: Oh yes, the, uh, the Norwegian Blue. What's, uh. What's wrong with it?

C: I'll tell you what's wrong with it, my lad. 'E's dead, that's what's wrong with it!

O: No, no, 'e's uh, he's resting.

C: Look, matey, I know a dead parrot when I see one, and I'm looking at one right now.

O: No no he's not dead, he's, he's restin'! Remarkable bird, the Norwegian Blue, idn'it, ay? Beautiful plumage!

The phrase in blue is the **closest** phrase that refers to the phrase in orange.

Words like **they**, **him**, **her** and **it** are likely to refer to something else in the text. Try to find the closest mention of this phrase.

Words like **they** or **them** could refer to more than one thing in the text so select more than one phrase if necessary.

Always look for the **closest previous mention** of the phrase to score maximum agreement points.

Phrase Detectives v0.3 | AnaWiki © 2008 | jchamb@essex.ac.uk | Admin

Figure 2: A screenshot of the Validation Mode.

Mode to enter a new annotation. The Validation Mode not only sorts ambiguous, incorrect and/or malicious decisions but also provides a social training mechanism [9].

When the users register they begin with the training phase of the game. Their answers are compared with Gold Standard texts to give them feedback on their decisions and to get a user rating, which is used to determine whether they need more training. Contextual instructions are also available during the game.

The corpus used in the game is created from short texts including: Wikipedia articles selected from the 'Featured Articles' and the page of 'Unusual Articles'; stories from Project Gutenberg including Aesop's Fables, Sherlock Holmes and Grimm's Fairy Tales; and dialogue texts from Textfile.com including Monty Python's Dead Parrot sketch. Selections from the GNOME and ARRAU corpora are also included to analyse the quality of the annotations.

4. THE SCORING SYSTEM

One of the most significant problems when designing a game that collects data is how to reward a player's decision when the correct answer is not known (and in some cases there may not be just one correct answer). Our solution is to motivate players using comparative scoring (awarding points for agreeing with the Gold Standard) and collaborative scoring (increasing the reward the more the players agree with each other).

In the game groups of players work on the same task over a period of time as this is likely to lead to a collectively intelligent decision [21]. An initial group of players are asked to annotate a markable. For each decision the player receives a 'decision' point. If all the players agree with each other then they are all awarded an additonal 'agreement' point and the markable is considered complete.

However it is likely that the first group of players will not agree with each other (62% of markables are given more than one relationship). In this case each unique relationship for the markable is validated by another group of players. The validating players receive an 'agreement' point for every player from the first group they agree with (either by agreeing or disagreeing). The players they agree with also receive an 'agreement' point.

This scoring system motivates the initial annotating group of players to choose the best relationship for the markable because it will lead to more points being added to their score later. The validating players are motivated to agree with these relationships as they will score more agreement points.

Contrary to expectations [3] it took players almost twice as long to validate a relationship than to annotate a markable (14 seconds compared to 8 seconds).

5. INCENTIVE STRUCTURES

The game is designed to use 3 types of incentive structure: personal, social and financial. All incentives were applied with caution as rewards have been known to decrease annotation quality [12]. The primary goal is to motivate the players to provide high quality answers, rather than large quantities of answers.

- Document topic
- Task speed
- User contributed documents
- Leaderboards
- Collaborative scoring
- Weekly and monthly prizes

5.1 Personal incentives

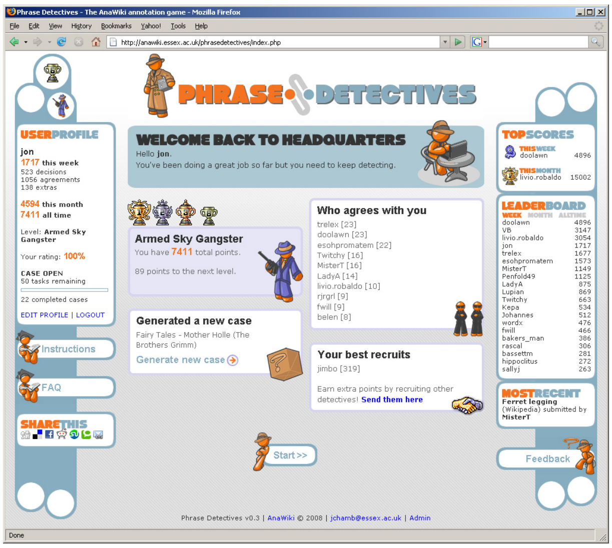
Personal incentives are evident when simply participating is enough of a reward for the user. For example, a Web user submitting information to Wikipedia does not usually receive any reward for what they have done but are content to be involved in the project. Similarly the progress of a player through a computer game will usually only be of interest to themselves, with the reward being the enjoyment of the game.

Generally, the most important personal incentive is that the user feels they are contributing to a worthwhile project. News and links to the research were posted on the homepage to reinforce the credibility of the project.

Also important for the players of Phrase Detectives is that they read texts that they find interesting. The choice of documents is important in getting users to participate in the game, to understand the tasks and to keep playing. Players can specify a preference for particular topics, however only 4% do so. This could be an indication that the corpus as a whole was interesting but it is more likely that they simply didn't change their default options [11].

It is also important for the players to read the documents at a relatively normal speed whilst still being able to complete the tasks. By default the tasks are generated randomly (although displayed in order) and limited (50 markable tasks selected from each document) which allows a normal reading flow. Players are given bonus points if they change their profile settings to select every markable in each document (which makes reading slower). Only 5% of players chose to sacrifice readability for the extra points.

In early versions of the game the player could see how long they had taken to do an annotation. Although this had no



jon
1717 this week
523 decisions
1056 agreements
138 extras

4594 this month
7411 all time

Level: Armed Sky Gangster

Your rating: 100%

CASE OPEN
50 tasks remaining

22 completed cases

WELCOME BACK TO HEADQUARTERS

Hello jon.
You've been doing a great job so far but you need to keep detecting.

Armed Sky Gangster

You have 7411 total points.

89 points to the next level.

Generated a new case

Fairy Tales - Mother Holle (The Brothers Grimm)

Who agrees with you

trelex [23]
doolawn [23]
esohpromatem [22]
Twitchy [16]
MisterT [16]
LadyA [14]
livio.robaldo [10]
rjrgrl [9]
fwill [9]
belen [8]

Your best recruits

jimbo [319]

Earn extra points by recruiting other detectives! Send them here

THISWEEK doolawn 4896

THISMONTH livio.robaldo 15002

doolawn	4896
VB	3147
livio.robaldo	3054
jon	1717
trelex	1677
esohpromatem	1573
MisterT	1149
Penfold49	1125
LadyA	875
Lupian	869
Twitchy	663
Kepa	534
Johannes	512
wordx	476
fwill	466
bakers_man	386
rascal	306
bassettm	281
hippoclitus	272
sallyj	263

Ferret legging (Wikipedia) submitted by MisterT

Phrase Detectives v0.3 | AnaWiki © 2008 | jchamb@essex.ac.uk | Admin

Figure 3: A screenshot of the player's homepage.

influence on the scoring, players complained that they felt under pressure and that they didn't have enough time to check their answers. This is in contrast to previous suggestions that timed tasks motivate players [23]. The timing of the annotations is now hidden from the players but still recorded with annotations. The relationship between the time of the annotation, the user rating and the agreement will be crucial in understanding how a timed element in a reading game influences the data that is collected.

The throughput of Phrase Detectives is 450 annotations per human hour (compared to the ESP game at 233 labels per human hour [23]). There is, however, a difference in data input between the 2 games, the former only requiring clicks on pre-selected phrases and the latter requiring the user to type in a phrase. The design of a game task must consider the speed at which the player can process the input source (e.g. text, images) and deliver their response (e.g. a click, typing) in order to maximise throughput and hence the amount of data that is collected.

We allowed users to submit their own text to the corpus. This would be processed and entered into the game. We anticipated that, much like Wikipedia, this would motivate users to generate content and become much more involved in the game. Unfortunately this was not the case, with only one user submitting text. We have now stopped advertising this incentive however the concept may still hold promise for games where the user-submitted content is more naturally created (e.g. collaborative story writing).

5.2 Social incentives

Social incentives reward users by improving their standing amongst their peers (in this case their fellow players).

Phrase Detectives features the usual incentives of a computer game, including weekly, monthly and all-time leaderboards, cups for monthly top scores and named levels for reaching a certain amount of points (see Figure 3). Interesting phenomenon have been reported with these reward mechanisms, namely that players gravitate towards the cut-off points (i.e. they keep playing to reach a level or high score before stopping) [24]. The collaborative agreement scoring in Phrase Detectives prevents us from effectively analysing this (as players continue to score even when they have stopped playing) however our high-scoring players can be regularly seen outscoring each other on the leaderboards.

In addition to the leaderboards that are visible to all players, each player can also see a leaderboard of other players who agreed with them. Although there is no direct incentive (as you cannot influence your own agreement leaderboard) it reinforces the social aspect of how the scoring system works. The success of games integrated into social networking sites like Sentiment Quiz[11] on Facebook indicates that visible social interaction within a game environment motivates the players to contribute more.

5.3 Financial incentives

Financial incentives reward effort with money. We introduced a weekly prize where a player is chosen by randomly selecting an annotation made during that week. This prize motivates low-scoring players because any annotation made during the week has a chance of winning (much like a lottery) and the more annotations you make, the higher your chance of winning.

We also introduced monthly prizes for the 3 highest scorers of the month. The monthly prize motivates the high-scoring players to compete with each other by doing more work, but also motivates some of the low-scoring players in the early parts of the month when the high score is low.

The weekly prize was £15 and the monthly prizes were £75, £50 and £25 for first, second and third places. The prizes were sent as Amazon vouchers by email.

6. QUALITY OF DATA

The psychological impact of incentive structures, especially financial ones, can create a conflict of motivation in players (i.e. how much time they should spend on their decisions). They may decide to focus on ways to maximise rewards rather than provide high quality answers. The game's scoring system and incentive structures are designed to reduce this to a minimum. We have identified four aspects that need to be addressed to control annotation quality: ensuring users understand the task; attention slips; malicious behaviour; and genuine ambiguity of data [9].

Further analysis will reveal if changing the number of players in the annotating and validating groups will effect the quality of the annotations. The game currently uses 8 players in the annotating group and 4 in the validating group with an average of 18 players looking at each markable. Some types of task can achieve high quality annotations with as few as 4 annotators [18] but other types of tasks (e.g anaphor resolution) may require more [15].

7. ATTRACTING & MOTIVATING USERS

The target audience for the game are English-speakers who spend significant amounts of time online, either playing computer games or casually browsing the Internet.

In order to attract the number of participants required to make a success of this methodology it is not enough to develop attractive games, but also successful advertising. Phrase Detectives was written about in local and national press, on science websites, blogs, bookmarking websites and

[11] `http://www.modul.ac.at/nmt/sentiment-quiz`

gaming forums. The developer of the game was also interviewed by the BBC. At the same time a pay-per-click advertising campaign was started on the social networking website Facebook, as well as a group connected to the project.

We investigated the sources of traffic since live release using Google Analytics. Incoming site traffic didn't show anything unusual: direct (46%); from a website link (29%); from the Facebook advert (13%); from a search (12%). However the bounce rate (the percentage of single-page visits, where the user leaves on the page they entered on) revealed how useful the traffic was. This showed a relatively consistent figure for direct (33%), link (29%) and search (44%) traffic. However for the Facebook advert it was significantly higher (90%), meaning that 9 out of 10 users that came from this source did not play the game. This casts doubt over the usefulness of pay-per-click advertising as a way of attracting participants to a game.

The players of Phrase Detectives were encouraged to recruit more players by giving them extra points every time they referred a player and whenever that player gained a level. The staggered reward for referring new players was to discourage players from creating new accounts themselves in order to get the reward. The scores of the referred players are displayed to the referring player on the recruits leaderboard. 4% of players have been referred by other players.

Attracting large numbers of players to a game is only part of the problem. It is also necessary to attract players who will make significant contributions. Since its release the game has attracted 750 players but we found that the top 10 players (5% of total) had 60% of the total points on the system and had made 73% of the annotations. This indicates that only a handful of users are doing the majority of the work, which is consistent with previous findings [18], however the contribution of one-time users should not be ignored [8]. Most of the players who have made significant contributions have a language-based background.

Players are invited to report on their experiences either through the feedback page or by commenting on a markable. Both methods send a message to the administrators who can address the issues raised and reply to the player if required. General feedback included suggestions for improvements to the interface and clarification of instructions and scoring. Frequent comments included reporting markables with errors from the pre-processing and discussing ambiguous or difficult markable relations.

It was intended to be a simple system of communication from player to administrator that avoids players colluding to gain points. However it is apparent that a more sophisticated community message system would enhance the player experience and encourage the development of a community.

8. IMPLEMENTATION

Phrase Detectives is running on a dedicated Linux server. The pre-processed data is stored in an MySQL database and most of the scripting is done via PHP.

The Gold Standard is created in Serengeti (a Web-based annotation tool developed at the University of Bielefeld [20]) by computational linguists. This tool runs on the same server and accesses the same database.

The database stores the textual data in Sekimo Generic Format (SGF) [19], a multi-layer representation of the original documents that can easily be transformed into other common formats such as MAS-XML and PAULA. We apply a pipeline of scripts to get from raw text to SGF format. For English texts this pipeline consists of these main steps:

- A pre-processing step normalises the input, applies a sentence splitter and runs a tokenizer over each sentence. We use the *openNLP*[12] toolkit to perform this process.
- Each sentence is analysed by the *Berkeley Parser*[13].
- The parser output is interpreted to identify markables in the sentence. As a result we create an XML representation which preserves the syntactic structure of the markables (including nested markables, e.g. noun phrases within a larger noun phrase).
- A heuristic processor identifies a number of additional features associated with markables such as person, case, number etc. The output format is MAS-XML.

The last two steps are based on previous work within the research group at Essex University [15]. Finally, MAS-XML is converted into SGF. Both MAS-XML and SGF are also the formats used to export the annotated data.

9. RESULTS

Before going live we evaluated a prototype of the game interface informally using a group of randomly selected volunteers from the University of Essex [2]. The beta version of *Phrase Detectives* went on-line in May 2008, with the first live release in December 2008. Over 1 million words of text have been added to the live game.

In the first 3 months of live release the game collected over 200,000 annotations and validations of anaphoric relations. To put this in perspective, the GNOME corpus, produced by traditional methods, included around 3,000 annotations of anaphoric relations [13] whereas OntoNotes[14] 3.0, with 1 million words, contains around 140,000 annotations.

The analysis of the results is an ongoing issue. However, by manually analyzing 10 random documents we could not find a single case in which a misconceived annotation was validated by other players. This confirms the assumptions we made about quality control. It will need to be further investigated by more thorough analysis methods which will be part of the future work.

10. CONCLUSIONS

The incentives structures used in Phrase Detectives were successful in motivating the users to provide high quality data. In particular the collaborative and social elements (agreement scoring and leaderboards) seem to offer the most promise if they can be linked with existing social networks.

The methodology behind collaborative game playing has become increasingly more widespread. Whilst the good-will of Web volunteers exists at the moment, there may be a point of saturation, where it becomes significantly more difficult to attract users and more novel incentive structures will need to be developed.

[12] `http://opennlp.sourceforge.net`

[13] `http://nlp.cs.berkeley.edu`

[14] `http://www.ldc.upenn.edu`

11. FUTURE WORK

We are progressively converting text for use in the game with the aim of having 100 million words. So far, mainly narrative texts from Project Gutenberg and encyclopedic texts from Wikipedia have been converted. We also plan to include further data from travel guides, news articles, and the American National Corpus [5].

It has become evident that working with a corpus of that size will require additional types of users. New tasks need to be developed, some as game tasks and others as admin player tasks that allow the management of players and documents to be handled by the users themselves. Motivating admin players will require very different incentive structures than have been used so far in the game.

The data collected by the game will be made available to the community through the Anaphoric Bank[15].

Ultimately, the usefulness of the annotated data will need to be shown by, for example, successfully training anaphora resolution algorithms that perform better than existing systems.

Acknowledgments

ANAWIKI is funded by a grant from the Engineering and Physical Sciences Research Council (EPSRC), grant number EP/F00575X/1. Thanks to Daniela Goecke, Nils Diewald, Maik Stührenberg and Daniel Jettka (University of Bielefeld), Mark Schellhase (University of Essex) and all the players who have contributed to the project, in particular *livio.robaldo*, *trelex*, *VB*, *TLS* and *Lupian*.

12. REFERENCES

[1] L. Burnard. The British National Corpus Reference guide. Technical report, Oxford University Computing Services, Oxford, 2000.

[2] J. Chamberlain, M. Poesio, and U. Kruschwitz. Phrase Detectives: A Web-based Collaborative Annotation Game. In *Proceedings of the International Conference on Semantic Systems (I-Semantics'08)*, Graz, 2008.

[3] T. Chklovski and Y. Gil. Improving the design of intelligent acquisition interfaces for collecting world knowledge from web contributors. In *K-CAP '05: Proceedings of the 3rd international conference on Knowledge capture*, pages 35–42, New York, NY, USA, 2005. ACM.

[4] E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. OntoNotes: The 90% Solution. In *Proceedings of HLT-NAACL06*, 2006.

[5] N. Ide and C. Macleod. The American National Corpus: A Standardized Resource of American English. In *Proceedings of Corpus Linguistics*, Lancaster, 2001.

[6] N. L. Johnson, S. Rasmussen, C. Joslyn, L. Rocha, S. Smith, and M. Kantor. Symbiotic Intelligence: Self-Organizing Knowledge on Distributed Networks Driven by Human Interaction. In *Proceedings of the Sixth International Conference on Artificial Life*. MIT Press, 1998.

[7] D. Jurafsky and J. H. Martin. *Speech and Language Processing- 2^{nd} edition.* Prentice-Hall, 2008.

[8] B. Kanefsky, N. Barlow, and V. Gulick. Can distributed volunteers accomplish massive data analysis tasks? *Lunar and Planetary Science*, XXXII, 2001.

[9] U. Kruschwitz, J. Chamberlain, and M. Poesio. (Linguistic) Science Through Web Collaboration in the ANAWIKI Project. In *Proceedings of WebSci'09*, Athens, 2009.

[10] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini. Building a large annotated corpus of english: the Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[11] K. Markey. Twenty-five years of end-user searching, Part 1: Research findings. *Journal of the American Society for Information Science and Technology (JASIST)*, 58(8):1071–1081, June 2007.

[12] J. Mrozinski, E. Whittaker, and S. Furui. Collecting a why-question corpus for development and evaluation of an automatic QA-system. In *Proceedings of ACL-08: HLT*, pages 443–451, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[13] M. Poesio. Discourse annotation and semantic annotation in the gnome corpus. In *Proceedings of the ACL Workshop on Discourse Annotation*, 2004.

[14] M. Poesio. The MATE/GNOME scheme for anaphoric annotation, revisited. In *Proceedings of SIGDIAL*, 2004.

[15] M. Poesio and R. Artstein. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation*, pages 76–83, 2005.

[16] P. Singh. The public acquisition of commonsense knowledge. In *Proceedings of the AAAI Spring Symposium on Acquiring (and Using) Linguistic (and World) Knowledge for Information Access*, Palo Alto, CA, 2002.

[17] K. Siorpaes and M. Hepp. Games with a purpose for the semantic web. *IEEE Intelligent Systems*, 23(3):50–60, 2008.

[18] R. Snow, O'Connor, D. Jurafsky, and A. Ng. Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. *Proceedings of EMNLP-08*, Jan 2008.

[19] M. Stührenberg and D. Goecke. SGF - An integrated model for multiple annotations and its application in a linguistic domain. In *Proceedings of Balisage: The Markup Conference*, Montreal, 2008.

[20] M. Stührenberg, D. Goecke, N. Diewald, A. Mehler, and I. Cramer. Web-based annotation of anaphoric relations and lexical chains. In *Proceedings of the ACL Linguistic Annotation Workshop*, pages 140–147, 2007.

[21] J. Surowiecki. *The Wisdom of Crowds.* Anchor, 2005.

[22] L. von Ahn. Games with a purpose. *Computer*, 39(6):92–94, 2006.

[23] L. von Ahn and L. Dabbish. Designing games with a purpose. *Communications of the ACM*, 51(8):58–67, 2008.

[24] L. von Ahn, R. Liu, and M. Blum. Peekaboom: a game for locating objects in images. In *Proceedings of CHI '06*, pages 55–64, 2006.

[15] http://www.anaphoricbank.org