Using Games to Create Language Resources: Successes and Limitations of the Approach

Chamberlain, J., Fort, K., Kruschwitz, U., Lafourcade, M. and Poesio, M.

Abstract One of the more novel approaches to collaboratively creating language resources in recent years is to use online games to collect and validate data. The most significant challenges collaborative systems face are how to train users with the necessary expertise and how to encourage participation on a scale required to produce high quality data comparable with data produced by "traditional" experts. In this chapter we provide a brief overview of collaborative creation and the different approaches that have been used to create language resources, before analysing games used for this purpose. We discuss some key issues in using a gaming approach, including task design, player motivation and data quality, and compare the costs of each approach in terms of development, distribution and ongoing administration. In conclusion, we summarise the benefits and limitations of using a gaming approach to resource creation and suggest key considerations for evaluating its utility in different research scenarios.

Jon Chamberlain
University of Essex, Wivenhoe Park, Colchester CO4 3SQ, England, e-mail: jchamb@essex.ac.uk

Karën Fort
INIST-CNRS / LIPN, 2, allée de Brabois, 54500 Vandoeuvre-lès-Nancy, France, e-mail: karen.fort@inist.fr

Udo Kruschwitz
University of Essex, Wivenhoe Park, Colchester CO4 3SQ, England, e-mail: udo@essex.ac.uk

Mathieu Lafourcade
LIRMM, UMR 5506 - CC 477, 161 rue Ada, 34392 Montpellier Cedex 5, France e-mail: mathieu.lafourcade@lirmm.fr

Massimo Poesio
University of Essex, Wivenhoe Park, Colchester CO4 3SQ, England, e-mail: poesio@essex.ac.uk

1 Introduction

Recent advances in human language technology have been made possible by groups of people collaborating over the Internet to create large-scale language resources. This approach is motivated by 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. This is demonstrated in simulations of collective behaviour in self-organising systems [34].

Web-based systems such as Wikipedia[1] and similar large initiatives have shown that a surprising number of individuals can be willing to participate in projects.

One of the more novel approaches to collaboratively creating language resources in recent years is to use online games to collect and validate data. The ESP Game[2], the first mass market online *game-with-a-purpose (GWAP)*, highlighted the potential for a game-based approach to resource creation (in this case image tagging). Since then, new games have been developed for different tasks including language resource creation, search verification and media tagging.

The most significant challenges collaborative systems face are how to train users with the necessary expertise and how to encourage participation on a scale required to produce large quantities of high quality data comparable with data produced by "traditional" experts.

In this chapter, we provide insights into GWAP for language resource creation, focusing on the successes and limitations of the approach by investigating both quantitative and qualitative results.

This study will use data from the Phrase Detectives game[3], developed by the University of Essex (England) to gather annotations on anaphoric co-reference, and the JeuxDeMots game[4], developed by Laboratoire d'Informatique, de Robotique et de Microelectronique de Montpellier (LIRMM, France) to create a lexico-semantic network.

We first provide a brief overview of collaborative creation and the different approaches that have been used to create language resources. We then provide details of Phrase Detectives and JeuxDeMots, followed by other notable efforts of GWAP for language resource creation. Next we discuss some key issues in using a gaming approach, focusing on task design, player motivation, and data quality. Finally we look at the costs of each approach in terms of development, distribution and ongoing administration. In conclusion, we summarise the benefits and limitations of the games-with-a-purpose approach.

[1] http://www.wikipedia.org

[2] http://www.gwap.com/gwap

[3] http://www.phrasedetectives.com

[4] http://www.jeuxdemots.org

2 Collaborative Creation and Collective Intelligence

Collaboration is a process where two or more people work together to achieve a shared goal. From the point of view of collaborative creation of language resources, the resources are the goal, and they are created or modified by at least two people, who work incrementally, in parallel or sequentially on the project.

In the latter case language resources are developed with people working on the same project but never exactly on the same part of it. Parallel work is necessary to evaluate the validity of the created resource. For example, inter-annotator agreement, using parallel annotations, was used in the Penn Treebank [48]. Incremental work involves adjudication, either by an expert, or by consensus.

Several attempts have been made recently to bring order to the rapidly developing field of collaborative creation on the Internet [62, 46, 80]. Wikipedia showed that allowing users free reign of encyclopaedic knowledge not only empowers mass participation but also that the resulting creation is of a very high quality. This can be seen as a good example of the broad term *collective intelligence* where groups of individuals do things collectively that seem intelligent [46].

Collective intelligence can be shown in many domains including Computer Science, Economics and Biology[5] but here we focus on coordinating collective action in computational systems that overcome the bottleneck in creating and maintaining language resources which would normally have to be done by paid administrators.

The utility of collective intelligence came to the fore when it was proposed to take a job traditionally performed by a designated employee or agent and outsource it to an undefined large group of Internet users through an open call. This approach, called *crowdsourcing* [31], revolutionised the way traditional tasks could be completed and made new tasks possible that were previously inconceivable due to cost or labour limitations.

One use for crowdsourcing can be as a way of getting large amounts of human work hours very cheaply as an alternative to producing a computerised solution that may be expensive or complex. However, it may also be seen as a way of utilising human processing power to solve problems that computers, as yet, cannot solve, termed *human computation* [72]. Human computation has particular appeal for *natural language processing (NLP)* because computer systems still need large resources for training algorithms that aim to understand the meaning of human language.

By combining collective intelligence, crowdsourcing and human computation it is possible to enable a large group of collaborators to work on linguistic tasks normally done by highly skilled (and highly paid) annotators and to aggregate their collective answers to produce a more complex dataset that not only is more robust than an individual answer but allows for linguistic ambiguity. Enabling groups of people to work on the same task over a period of time is likely to lead to a collectively intelligent decision [68].

---

[5] http://scripts.mit.edu/∼cci/HCI

Three variations of this type of collaboration over the Internet have been successful in recent years and are distinguished by the motivations of the participants.

The first variation is where the motivation for the users to participate already exists. This could be because the user is inherently interested in contributing, for example in the case of Wikipedia or GalaxyZoo[6], or intrinsically motivated because they need to accomplish a different task, for example the reCAPTCHA[7] authentication system.

Unfortunately, most linguistic tasks are neither interesting (for the majority of people) nor easy to integrate into another system. Therefore, a second variation of crowdsourcing called microworking was proposed, where participants are paid small amounts of money to perform tasks. Although the payments are small, the total cost for a language resource produced in this way will increase proportionately with its size. Therefore, it is being used more in NLP for the fast annotation of small to medium sized corpora and for some types of linguistic evaluation [9].

This approach demonstrates the difficulties in producing the size of resources needed for modern linguistic tools, so a third approach was proposed to make the motivation for the user be entertainment rather than money. The *games-with-a-purpose (GWAP)* approach showed enormous initial potential and has been used for a variety of data collection and annotation tasks where the task has been made fun. In this chapter we focus on games used to create language resources.

3 Approaches to Creating Language Resources

3.1 Traditional, Entirely Validated Annotation

In order to evaluate crowdsourcing approaches to language resource creation it is necessary to also consider more traditional approaches. When we talk about traditional annotation, we think of the methodology used, for example, to create the OntoNotes corpus[8], containing multilingual annotated news articles, dialogue transcriptions and weblogs, and the SALSA corpus[9] of syntactically annotated German newspaper articles.

In this approach, a formal coding scheme is developed, and often extensive agreement studies are carried out. Every document is annotated twice according to the coding scheme by two professional annotators under the supervision of an expert, typically a linguist, followed by merging and adjudication of the annotations. These projects also generally involve the development of suitable annotation tools or at least the adaptation of existing ones.

---

[6] http://www.galaxyzoo.org

[7] http://www.google.com/recaptcha

[8] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2011T03

[9] http://www.coli.uni-saarland.de/projects/salsa

3.2 Traditional, Partly Validated Annotation

This type of annotation also involves the development of a formal coding scheme and training of annotators but most items will be typically annotated only once, for example in the ARRAU [57] and GNOME [56] corpora for anaphoric co-reference.

Approximately 10% of items are double-annotated to identify misunderstandings and improve the annotation guide [8]. In many cases, the annotations will have to be corrected, possibly extensively. Annotation is typically carried out by trained annotators, generally students, under the supervision of an expert annotator.

3.3 Microwork Crowdsourcing

Amazon Mechanical Turk (AMT)[10] pioneered microwork crowdsourcing: using the Web as a way of reaching very large numbers of workers (sometimes referred to as turkers) who get paid to complete small items of work called *human intelligence tasks (HITs)*. This is typically very little – in the order of 0.01 to 0.20 US$ per HIT.

Some studies have shown that the quality of resources created this way are comparable to that of resources created in the traditional way, provided that multiple judgements are collected in sufficient number and that enough post-processing is done [67, 9]. Other studies have shown that the quality does not equal that provided by experts [6] and for some tasks does not even surpass that of automatic language technology [76]. It is beyond the scope of this chapter to go into great depth about the quality attainable from AMT, rather we simply compare reported results with that reported from other approaches.

A further reported advantage of AMT is that the work is completed very fast. It is not uncommon for a HIT to be completed in minutes, but this is usually for simple tasks. In the case of more complex tasks, or tasks where the worker needs to be more skilled, e.g., translating a sentence in an uncommon language, it can take much longer [55].

AMT is very competitive with traditional resource creation methods from a financial perspective. Whilst AMT remains a very popular microworking platform some serious issues regarding the rights of workers, minimum wage and representation have been raised [24]. Other microworking platforms, such as Samasource[11], guarantee workers a minimum payment level and basic rights.

Microwork crowdsourcing is becoming a standard way of creating small-scale language resources but even this approach can become prohibitively expensive to create resources of the size that are increasingly required in modern linguistics, i.e., in the order of 100 million annotated words.

---

[10] http://www.mturk.com

[11] http://samasource.org

3.4 Games With A Purpose (GWAP)

Generally speaking, a game-based crowdsourcing approach uses entertainment rather than financial payment to motivate participation. The approach is motivated by the observation that every year an estimated 9 billion person-hours are spent by people playing games on the Web [72]. If even a fraction of this effort could be redirected towards useful activity that has a purpose, as a side effect of having people play entertaining games, there would be an enormous human resource at our disposal.

GWAP come in many forms; they tend to be graphically rich, with simple interfaces, and give the player an experience of progression through the game by scoring points, being assigned levels and recognising their effort. Systems are required to control the behaviour of players: to encourage them to concentrate on the tasks and to discourage them from malicious behaviour. This is discussed in more detail later.

The GWAP approach showed enormous initial potential, with the first, and perhaps most successful, game called the ESP Game. In the game two randomly chosen players are shown the same image. Their goal is to guess how their partner will describe the image (hence the reference to extrasensory perception or ESP) and type that description under time constraints. If any of the strings typed by one player matches the strings typed by the other player, they both score points. The descriptions of the images provided by players are very useful to train content-based image retrieval tools.

The game was very popular, attracting over 200,000 players who produced over 50 million labels [72]. The quality of the labels has been shown to be as good as that produced through conventional image annotation methods. The game was so successful that a license to use it was bought by Google, who developed it into the Google Image Labeler which was online from 2006 to 2011.

GWAP have been used for many different types of crowdsourced data collection [70] including:

- image annotation such as the ESP Game, Matchin, FlipIt, Phetch, Peekaboom, Squigl, Magic Bullet and Picture This;
- video annotation such as OntoTube, PopVideo, Yahoo's VideoTagGame and Waisda;
- audio annotation such as Herd It, Tag a Tune and WhaleFM;
- biomedical applications such as Foldit, Phylo and EteRNA;
- transcription such as Ancient Lives and Old Weather;
- improving search results such as Microsoft's Page Hunt;
- social bookmarking such as Collabio.

Links to the GWAP listed above can be found in Appendix A.

GWAP have a different goal to *serious games*, where the purpose is to educate or train the player in a specific area such as learning a new language or secondary school level topics [51]. Serious games can be highly engaging, often in a 3D world, and have a directed learning path for the user as all of the data is known to the system

beforehand. Therefore, the user can receive immediate feedback as to their level of performance and understanding at any point during the game.

GWAP aim to entertain players whilst they complete tasks that the system does not know, for the most part, the correct answer, and in many cases there may not even be a "correct" answer. Hence, providing feedback to users on their work presents a major challenge and understanding the motivation of players in this scenario is a key to the success of a GWAP.

4 Using Games to Create Language Resources

This section looks in detail at the design and reported results from two GWAP for NLP: Phrase Detectives and JeuxDeMots. For completeness, we mention other notable GWAP used for linguistic purposes and a summary, with links where available, is in Appendix B.

4.1 Phrase Detectives

Phrase Detectives (PD) is a single-player GWAP designed to collect data about English (and subsequently Italian) anaphoric co-reference [14, 59]. The game architecture is structured around a number of tasks that use 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.

The game uses two styles of text annotation for players to complete a linguistic task. Initially text is presented in Annotation Mode (called *Name the Culprit* in the game - see Figure 2). This is a straightforward annotation mode where the player makes an annotation decision about a highlighted markable (section of text). 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 - see Figure 3). The players in Validation Mode have to agree or disagree with the interpretation.

Players are trained with training texts created from a gold standard (a text that has been annotated by a linguistic annotation expert). 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 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,

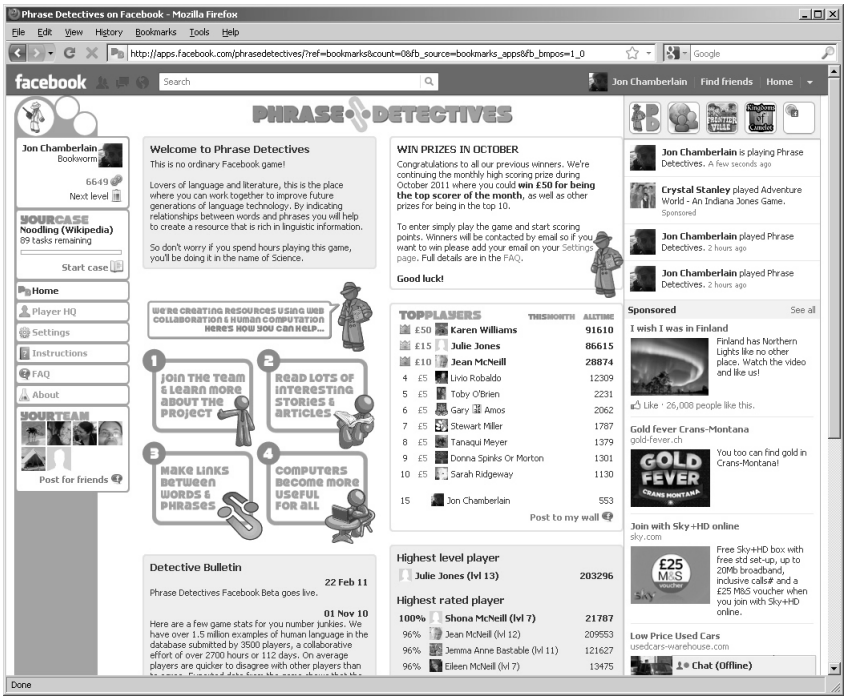

Fig. 1 Screenshot of the Phrase Detectives Facebook homepage.

including personal (achieving game goals and scoring points), social (competing with other players) and financial (small prizes), are employed.

Text used in PD comes from two main domains: Wikipedia articles selected from the 'Featured Articles' page[12] and the page of 'Unusual Articles'[13]; and narrative text from Project Gutenberg[14] including simple short stories (e.g., Aesop's Fables, Grimm's Fairy Tales, Beatrix Potter's tales) and more advanced narratives such as several Sherlock Holmes stories by A. Conan-Doyle, *Alice in Wonderland*, and several short stories by Charles Dickens.

The goal of the game was not just to annotate large amounts of text, but also to collect a large number of judgements 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 validation to tools for analysing the behaviour of players (see Figure 7).

[12] http://en.wikipedia.org/wiki/Wikipedia:Featured_articles

[13] http://en.wikipedia.org/wiki/Wikipedia:Unusual_articles

[14] http://www.gutenberg.org

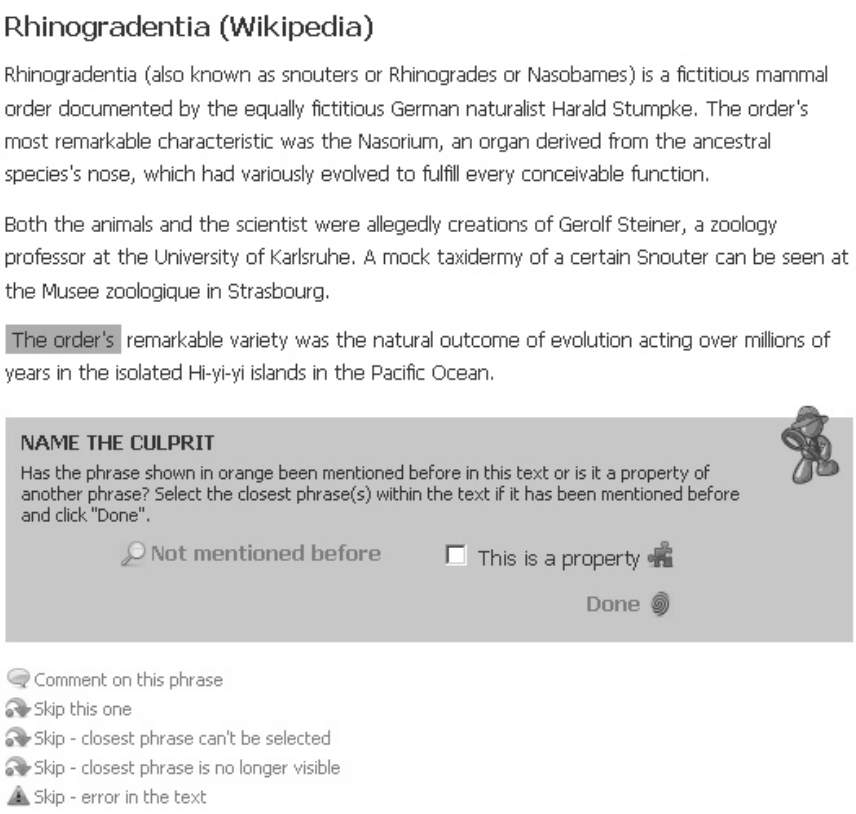

Fig. 2 Detail of a task presented in Annotation Mode in Phrase Detectives on Facebook.

A version of PD was developed for Facebook[15] that maintained the previous game architecture whilst incorporating a number of new features developed specifically for the social network platform (see Figure 1).

The game was developed with PHP SDK[16] (an API for accessing user data, friend lists, wall posting, etc) and integrates seamlessly within the Facebook site. Both implementations of the game run simultaneously on the same corpus of documents.

This version of the game makes full use of socially motivating factors inherent in the Facebook platform. Any of the player's friends from Facebook, who are also 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 both score additional points.

Player levels have criteria, including total points scored, player rating and total wall posts made from the game. The player must activate their new level once the

[15] http://www.facebook.com

[16] http://developers.facebook.com/docs/reference/php

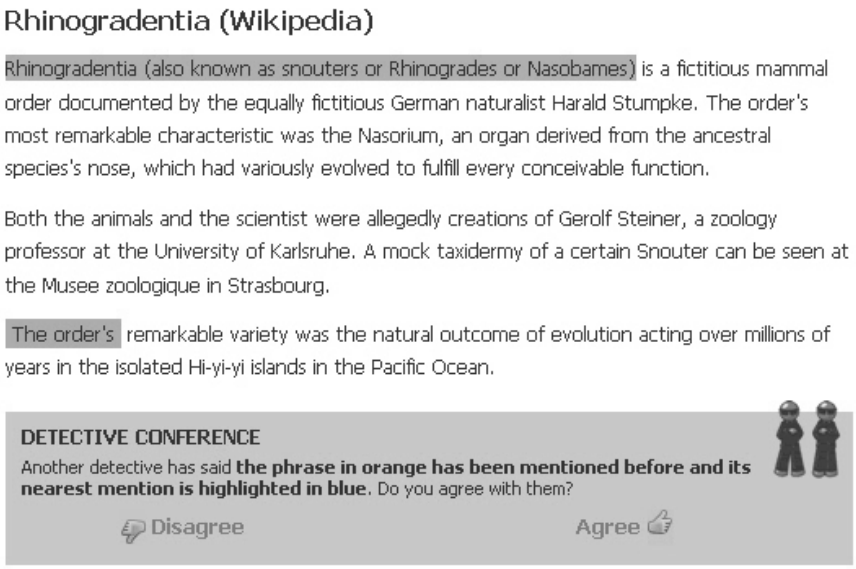

Fig. 3 Detail of a task presented in Validation Mode in Phrase Detectives on Facebook.

criteria are met. In addition to the monthly and all-time leaderboards, the Facebook version has leaderboards for the highest level players, highest rated players and the players with the biggest team.

The purpose of redeveloping the game for Facebook was to investigate the utility of social networking sites in achieving high visibility and to explore different ways players can collaborate.

The first game was released in December 2008, with the Facebook version released in February 2011. Both games continue to collect data but results reported here are from December 2008 to February 2012 or are from previously published papers [12, 13, 15, 59].

4.2 JeuxDeMots

JeuxDeMots (JDM) is a two player GWAP, launched in September 2007, that aims to build a large lexico-semantic network composed of terms (nodes) and typed relations (links between nodes) [42] – see Figure 4. It contains terms and possible refinements in the same spirit as WordNet [21], although it is organised as decision trees. There are more than 50 different relation types, the occurrences of which are weighted.

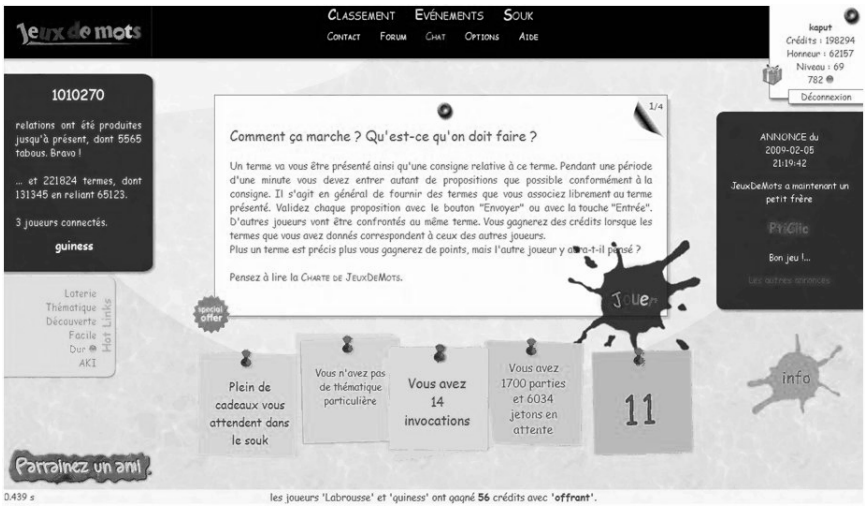

Fig. 4 Screenshot of the JeuxDeMots homepage. From here the player has status information and can launch a game by clicking on the *jouer* (play) button.

When a player begins a game, instructions concerning the type of lexical relation (synonyms, antonym, domain, etc.) are displayed, as well as a term *T*, chosen from the database or offered by other players. The player has limited time to enter terms which, to their mind, correspond to term *T* and the lexical relation. The maximum number of terms a player can enter is limited, thus encouraging the player to think carefully about their choices. A screenshot of the game is shown in Figure 5.

The same term *T*, along with the same instructions, are later given to another player for whom the process is identical. To make the game more fun, the two players score points for words they both choose. Score calculation was designed to increase both precision and recall in the construction of the database [35]. In the context of the lexical network, precision is related to the set of the most immediate and activated relations of a given term that are uttered by native speakers. Recall is related to the set of the numerous but relevant low activation relations (also known as the long tail) [43].The more original a proposition given by both players, the more it is rewarded. Answers given by both players are displayed, those common to both players are highlighted, as are their scores (see Figure 6).

For a target term *T*, common answers from both players are inserted into the database. Answers given by only one of the two players are not, thus reducing noise. The semantic network is therefore constructed by connecting terms by typed and weighted relations, validated by pairs of players. These relations are labelled according to the instructions given to the players and weighted according to the number of pairs of players who choose them.

Initially, prior to putting the game online, the database was populated with 140,000 terms (nodes) from French dictionaries, however if a pair of players suggest a non-existing term, a new node is added to the database. Since populating the

Fig. 5 Screenshot of an ongoing game in JDM with the target word *laver* (to wash). Several propositions have been given by the player and are listed on the right hand side.

database the players have added 110,000 new terms however these include spelling mistakes, plurals, feminine forms, numbers, dates and foreign words.

In the interest of quality and consistency, it was decided that the validation process would involve anonymous players playing together. A relation is considered valid only if it is given by at least one pair of players. This validation process is similar to the process for indexing images [73] and, more recently, to collect common sense knowledge [45] and for knowledge extraction [65].

The activity of the players in JDM constructs a lexical network which contains over 50 types of ontological relations such as generic relations (hypernyms), specific relations (hyponyms), part and whole relations, matter and substance, domain, synonyms and antonyms (the latter also being strongly lexical). The ongoing process of the network construction leads to the identification of word usages for disambiguating terms of the constructed ontology.

4.3 Other GWAP for Language Resources

4.3.1 Knowledge Acquisition

1001 Paraphrases [17], one of the first GWAP whose aim was to collect corpora, was developed to collect training data for a machine translation system that needs to recognise paraphrase variants. In the game, players have to produce paraphrases of an expression shown at the top of the screen, such as “this can help you”. If they guess one of the paraphrases already produced by another player, they get the

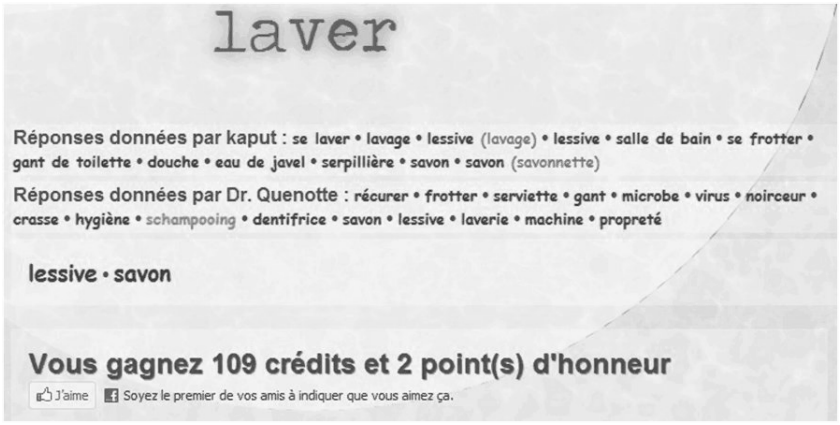

Fig. 6 Screenshot of the result of a game in JDM. Two words *lessive* and *savon* were given by both players for the term *laver* and hence scores them both points.

number of points indicated in the window; otherwise the guess they produced is added to those already collected by the system, the number of points they can win is decreased, and they can try again. Many of the ideas developed in 1001 Paraphrases, and the earlier LEARNER system, are extremely useful, in particular the idea of validation.

Other games for collecting common sense knowledge include FACTory, Verbosity, Categorilla and Free Association.

4.3.2 Text Annotation

The game most directly comparable with PD is PlayCoref, developed at Charles University in Prague [29]. PlayCoref is a two-player game in which players mark coreferential pairs between words in a text (no phrases are allowed). They mark the coreferential pairs as undirected links. During the session, the number of words the opponent has linked into the coreferential pairs is displayed to the player. The number of sentences with at least one coreferential pair marked by the opponent is displayed to the player as well. A number of empirical evaluations have been carried out showing that players find the game very attractive but the game has not yet been put online to collect data on a large scale.

PhraTris is a GWAP for syntactic annotation developed by Giuseppe Attardi's lab at the University of Pisa using a general-purpose GWAP development platform called GALOAP. PhraTris, based on the traditional game Tetris, has players arrange sentences in a logical way, instead of arranging falling bricks, and won the Insemtives Game Challenge 2010. The game is not online but can be downloaded and installed locally.

PackPlay [28] was another attempt to build semantically-rich annotated corpora. The two game variants *Entity Discovery* and *Name That Entity* use slightly different

approaches in multi-player games to elicit annotations from players. Results from a small group of players showed high precision and recall when compared to expert systems in the area of named entity recognition, although this is an area where automated systems also perform well.

4.3.3 Sentiment Analysis

Human language technology games integrated into social networking sites such as Sentiment Quiz [63] on Facebook show that social interaction within a game environment does motivate players to participate. The Sentiment Quiz asks players to select a level of sentiment (on a 5 point scale) associated with a word taken from a corpus of documents regarding the 2008 US Presidential election. The answer is compared to another player and points awarded for agreement.

4.3.4 Generation

A family of GWAP which have been used to collect data used in computational linguistics are the GIVE games developed in support of the GIVE-2 challenge for generating instructions in virtual environments, initiated in the Natural Language Generation community [40]. GIVE-2 is a treasure-hunt game in a 3D world. When starting the game, the player sees a 3D game window, which displays instructions and allows the players to move around and manipulate objects. In the first room players learn how to interact with the system; then they enter into an evaluation virtual world where they perform the treasure hunt, following instructions generated by one of the systems participating in the challenge. The players can succeed, lose, or cancel the game and this outcome is used to compute the task success metric, one of the metrics used to evaluate the systems participating in the challenge.

GIVE-2 was extremely successful as a way to collect data, collecting over 1825 game sessions in three months, which played a key role in determining the results of the challenge. This is due, in part, to the fact that it is an extremely attractive game to play.

4.3.5 Ontology Building

The OntoGame, based around the ESP Game data collection model, aims to build ontological knowledge by asking players questions about sections of text, for example whether it refers to a class of object or an instance of an object. Other Web-based systems include Open Mind Word Expert [52], which aims to create large sense-tagged corpora, and SemKey [47] which makes use of WordNet and Wikipedia to disambiguate lexical forms referring to concepts.

5 Defining Collaborative Approaches

There have been several recent attempts to define and classify collaborative approaches in collective intelligence and distributed human computation [62, 46]. We focus on 3 dimensions proposed for crowdsourcing projects [77] that are essential considerations when designing GWAP for NLP:

- Task Character
- Player Motivation
- Annotation Quality

5.1 Task Character

5.1.1 Game Interface

Most GWAP tend to have fairly simple interfaces making it easy for first time users to start playing, with a short timespan (i.e., arcade style) and online delivery. This constrains the game to small tasks in a programmatically simple framework which is suitable for the goal of collecting data. A game deployed on the Web should observe all the normal guidelines regarding browser compatibility, download times, consistency of performance, spatial distance between click points, etc.[17]

Game interfaces should be graphically rich, although not at the expense of usability, and aimed at engaging the target audience (i.e., a game aimed at children may include more cartoon or stylised imagery in brighter colours than a game aimed at adults). The game should also provide a consistent metaphor within the gaming environment. For this PD used a detective metaphor, with buttons stylised with a cartoon detective character and site text written as if the player was a detective solving cases. The game task should be integrated in such a way that task completion, scoring and storyline form a seamless experience.

Three styles of game scenario have been proposed [74]:

1. Output-agreement, where the players must guess the same output from one input;
2. Inversion-problem, where one player describes the input to a second player who must guess what it is;
3. Input-agreement, where two players must guess whether they have the same input as each other based on limited communication.

The Output-agreement game scenario is the most straight forward to implement and collect data from, however, other scenarios can make the game more interesting for the players and increase their enjoyment.

[17] http://www.usability.gov/guidelines

5.1.2 Task Design

Whilst the design of the game interface is important, it is the design of the task that determines how successfully the player can contribute data. In PD the player is constrained to a set of predefined options to make annotations, with freetext comments allowed (although this is not the usual mode of play in the game). The preprocessing of text allows the gameplay in PD to be constrained in this way but is subject to errors in processing that also need to be fixed.

JDM requires players to type text into a freetext box which allows for the collection of novel inputs but will also collect more noise from players through spelling mistakes and similar inputs. These can be filtered out using post-processing and validation, however it makes the collection of novel and ambiguous data more difficult.

The task design has an impact on the speed at which players can complete tasks, with clicking being faster than typing. A design decision to use radio buttons or freetext boxes can have a significant impact on performance [1].

The interface of AMT is predefined and presents limitations that constitute an important issue for some tasks, for example to annotating noun compound relations using a large taxonomy [71]. In a word sense disambiguation task considerable redesigns were required to get satisfactory results [30]. These examples show how difficult it is to design NLP tasks for crowdsourcing within a predefined system.

5.2 Player Motivation

The motivation of the players is an important issue both in terms of data analysis and of return on investment (and therefore cost).

Incentives that motivate players to participate can be categorised into 3 groups: personal; social; and financial [15]. These directly relate to other classifications of motivations in previous research: Love; Glory; and Money [46].

Given that GWAP attempts to avoid direct financial incentives (as found in microwork crowdsourcing) the game must motivate the player with entertainment and enjoyment.

There may also be other motivational considerations, such as the desire to contribute to a scientific project or for self enrichment and learning.

All incentives should be applied with caution as rewards have been known to decrease annotation quality [53].

It is important to distinguish between *motivation to participate* (why people start doing something) and *motivation to contribute or volition* (why they continue doing something) [23]. Once both conditions are satisfied we can assume that a player will continue playing until other factors such as fatigue or distraction break the game cycle. This has been called *volunteer attrition*, where a player's contribution diminishes over time [45].

Although incentives can be categorised, in reality they form a complex psychology in participants that is best discussed by focusing on a particular game consideration:

- The concept of enjoyment as a motivator;
- How timing tasks affects player motivation;
- Altruism and citizen science;
- Indirect financial incentives in games;
- Publicity, virality and social networks.

5.2.1 Enjoyment as an Incentive

GWAP focuses on one main type of incentive: enjoyment. There is substantial literature on what makes games fun [41] and models of enjoyment in games (called *the game flow*) identify 8 criteria for evaluating enjoyment [69] (the model being based on a more generic theory [19]):

1. Concentration - Games should require concentration and the player should be able to concentrate on the game;
2. Challenge - Games should be sufficiently challenging and match the player's skill level;
3. Player skills - Games must support player skill development and mastery;
4. Control - Players should feel a sense of control over their actions in the game;
5. Clear goals - Games should provide the player with clear goals at appropriate times;
6. Feedback - Players must receive appropriate feedback at appropriate times;
7. Immersion - Players should experience deep but effortless involvement in the game;
8. Social interaction - Games should support and create opportunities for social interaction.

The main method used by GWAP to make players enjoy the task is by providing them with a challenge. This is achieved through mechanisms such as requiring a timed response, keeping scores that ensure competition with other players, and having players of roughly similar skill levels play against each other. In JDM, the challenge is both the combination of timed response and word-relation pairs of various difficulties.

For the players of PD, they can choose to read texts that they find interesting and have some control over the game experience. Whilst some texts are straightforward, others can provide a serious challenge of reading comprehension and completion of linguistic tasks. Players can also comment on the gaming conditions (perhaps to identify an error in the game, to skip a task or to generate a new set of tasks) and contact the game administrators with questions.

One of the simplest mechanisms of feedback is scoring. By getting a score the player gains a sense of achievement and some indication as to how well they are doing in the game.

GWAP tend to be short, arcade style games so immersion is achieved by progression through the game: by learning new types of tasks; becoming more proficient at current tasks; and by assigning the player a named level, starting from novice and going up to expert.

Social incentives are also provided by the scoring mechanism. Public leaderboards reward players by 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 [74]. An interesting phenomenon has 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) [75], however analysis of data from PD on Facebook did not support this [13].

5.2.2 Time-based Challenges in Language Tasks

The timing of tasks is usually required in the game format, either as motivational feature or as a method of quality control checking (or both). von Ahn and his colleagues view timing constraints as a key aspect of what makes games exciting [74], and built them into all their games. This is also the case for many other GWAP including JDM.

In PD, however, there are no timing constraints, although the time taken to perform a task is used to assess the quality of annotations. As the task in PD is text based (rather than image based in the ESP Game), it was considered important to give players time to read documents at a relatively normal speed whilst completing tasks.

This was supported by the results of the first usability study of PD. In the game prototype used in that study, players could see how long it had taken to do an annotation. On the contrary to suggestions that timing provides an incentive, the subjects complained that they felt under pressure and that they did not have enough time to check their answers, even though the time had no influence on the scoring. As a result, in all following versions of PD the time it takes players to perform a task is recorded but not shown.

Several players found the timing element of JDM stressful and in one case a player gave up the game for this reason. Most players in this game consider a timed task as normal and exciting and can buy extra time when needed (a game feature).

The time limitation tends to elicit spontaneous answers in a way that is not possible without a time limit where the players can give a more considered response. The design of the task must balance the increase in excitement a timed element can offer with the need to allow players time to give good quality answers.

Related to this are the concepts of “throughput” and “wait time”, discussed in more detail later, that are used to assess the efficiency of an interface. By increasing the speed at which the players are working, by using a timed element, you also increase the speed at which you can collect data.

5.2.3 Altruism and Participation in a Scientific Community

People who contribute information to Wikipedia are motivated by personal reasons such as the desire to make a particular page accurate, or the pride in one's knowledge in a certain subject matter [79]. This motivation is also behind the success of *citizen science* projects, such as the Zooniverse collection of projects[18], where the scientific research is conducted mainly by amateur scientists and members of the public.

GWAP may initially attract collaborators (e.g., other computational linguists) by giving them the sense that they are contributing to a resource from which a whole discipline may benefit and these are usually the people that will be informed first about the research. However, in the long term, most of the players of GWAP will never directly benefit from the resources being created. It is therefore essential to provide some more generic way of expressing the benefit to the player.

For example, this was done in PD with a BBC radio interview by giving examples of NLP techniques used for Web searching. Although this is not a direct result of the language resources being created by this particular GWAP, it is the case for efforts of the community as a whole, and this is what the general public can understand and be motivated by.

The purpose of data collection in GWAP has an advantage over microworking in AMT, where the workers are not connected to the requester, in that there is a sense of ownership, participation in science, and generally doing something useful. When players become more interested in the purpose of the GWAP than the game itself it becomes more like a citizen science approach where players are willing to work on harder tasks, provide higher quality data and contribute more.

In JDM, the collected data is freely available and not restricted in use (under the Creative Commons licence). The players do not have to know they are part of a research project although it is written in the rules of the game. Players reported that they were more interested in playing the JDM game than knowing what the data was used for. However, for some players (around 20) the purpose of the GWAP approach became more important than the game. These players played more on incomplete and difficult term-relation couples. The fact that the data constructed is freely available does matter for these types of players.

5.2.4 Indirect Financial Incentives

Indirect financial incentives in GWAP are distributed as prizes which are not directly related to the amount of work being done by the player, unlike microworking where a small sum of money is paid for the completion of a particular task.

In PD, financial incentives were offered in the form of a daily or weekly lottery, where each piece of work stood an equal chance of winning, or for high scoring players. These were distributed as Amazon vouchers emailed to the winning player.

[18] https://www.zooniverse.org

The ESP Game occasionally offers financial incentives in a similar way. JDM and most GWAP do not offer financial incentives.

Whilst financial incentives seem to go against the fundamental idea behind GWAP (i.e., that enjoyment is the motivation), it actually makes the enjoyment of potentially winning a prize part of the motivation. 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. By using a lottery style financial prize the hard working players are more likely to win, but the players who only do a little work are still motivated.

Indirect financial incentives can be a cost-effective way to increase participation in GWAP, i.e., the increase of work completed per prize fund is comparable to the cost of other approaches.

5.2.5 Attracting Players

In order to attract the number of participants required to make a success of the GWAP approach, it is not enough to develop attractive games; it is also necessary to develop effective forms of advertising. The number of online games competing for attention is huge and without some effort to raise a game's profile, it will never catch the attention of enough players. The importance of this strategy was demonstrated by von Ahn's lab. The ESP Game was constantly advertised in the press and also on TV. Other methods to reach players included blogs and being discussed on gaming forums. Part of the success of PD was down to the advertising of the game on blogs, language lists, conferences, tutorials and workshops as well as traditional media (via press releases). JDM on the other hand relied exclusively on word of mouth.

Not all advertising methods are equally successful and it is important to evaluate which works best for the game task, language or country.

Indirect financial incentives have been shown to be a good way to attract new players, however it is other motivational elements that keep players contributing to a game [66].

5.2.6 Virality and Social Networks

Social incentives can be made more effective when the game is embedded within a social networking platform such as 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 each other.

The PD game on Facebook allowed players to make posts to their wall (or news feed). Posting is considered a very important factor in recruiting more players as

surveys have shown that the majority of social game players start to play because of a friend recommendation.[19] [20]

Posts were automatically generated in PD and could be created by a player by clicking a link in the game. They could either be *social* in nature, where the content describes what the player is doing or has done, or *competitive*, where the content shows achievements of the player. Results showed that players preferred to make social posts, i.e., about the document they were working on or had just completed (52%). This compares to competitive posts when they went up a level (13%), when their rating was updated (10%) or to post about their position in the leaderboard (12%). The remaining 13% of posts were players making a direct request for their friends to join the game. This indicates that social motivations are more important than competitive motivations, at least on this platform.

In JDM, some social network features exist as achievements (scoring, winning some words, etc) displayed on Facebook however the real impact of such features is uncertain.

5.3 Annotation Quality

Whereas the designers of standard online games only need to motivate players to participate, the designers of GWAP also need to motivate the players to contribute good quality work. Obtaining reliable results from non-experts is also a challenge for other crowdsourcing approaches, and in this context strategies for dealing with the issue have been discussed extensively [39, 2, 3, 22].

In the case of microworking, the main strategy for achieving good quality labelling is to aggregate results from many users to approximate a single expert's judgements [67].

However, for the task of word-sense disambiguation, a small number of well-trained annotators produces much better results than a larger group of AMT workers [6] which illustrates that higher quality cannot always be achieved by simply adding more workers.

GWAP for linguistic annotation is not motivated solely by the desire to label large amounts of data. Web collaboration could also be used to gather data about the interpretation of natural language expressions, which all too often is taken to be completely determined by context, often without much evidence [60]. From this perspective it is important to attempt to avoid poor quality individual judgements.

The strategies for quality control in GWAP address four main issues:

- Training and Evaluating Players
- Attention Slips
- Multiple Judgements and Genuine Ambiguity
- Malicious Behaviour

[19] http://www.infosolutionsgroup.com/2010_PopCap_Social_Gaming_Research_Results.pdf

[20] http://www.lightspeedresearch.com/press-releases/it's-game-on-for-facebook-users

5.3.1 Training and Evaluating Players

GWAP usually begin with a training stage for players to practice the task and also to show that they have sufficiently understood the instructions to do a real task. However, the game design must translate the language task into a game task well enough for it still to be enjoyable, challenging and achievable. GWAP need to correlate good performance in the game with producing good quality data, but this is not an easy thing to do.

The level of task difficulty will drive the amount of training that a player will need. Simple tasks like image tagging need very little instruction other than the rules of the game, whereas more complex judgements such as those required by PD may require the players to be either more experienced or to undergo more training. The training phase has been shown to be an important factor in determining quality and improvement in manual annotation [20].

Most GWAP, at least initially, will have a core of collaborators to test and perform tasks and these are most likely to be friends or colleagues of the task designers. It can therefore be assumed that this base of people will have prior knowledge of the task background, or at least easy access to this information. These pre-trained collaborators are not the "crowd" that crowdsourcing needs if it is to operate on a large scale nor are they the "crowd" in the wisdom of the crowd.

Training should assume a layman's knowledge of the task and should engage the participant to increase their knowledge to become a pseudo-expert. The more they participate, the more expert they become. This graduated training is difficult to achieve and makes a rating system (where the user is regularly judged against a gold standard) essential to give appropriately challenging tasks.

As previously discussed, players can be motivated by a myriad of complex reasons. The desire to progress in the game may become more important to the player than to provide good quality work and this may lead to the desire to cheat the system.

5.3.2 Attention Slips

Players may occasionally make a mistake and press the wrong button. Attention slips need to be identified and corrected by validation, where players can examine other players' work and evaluate it. Through validation, poor quality interpretations should be voted down and high quality interpretations should be supported (in the cases of genuine ambiguity there may be more than one). Validation thus plays a key role as a strategy for quality control.

Unlike collaboration in Wikipedia, it is not advisable to allow players of GWAP to go back and correct their mistakes, otherwise a player could try all possible variations of an answer and then select the one offering the highest score. In this sense the way players work together is more "collective" than "collaborative".

5.3.3 Multiple Judgements and Genuine Ambiguity

Ambiguity is an inherent problem in all areas of NLP [36]. Here, we are not interested in solving this issue, but in using collaborative approaches to capture ambiguity where it is appropriate. Therefore, language resources should not only aim to select the best, or most common, annotation but also to preserve all inherent ambiguity, leaving it to subsequent processes to determine which interpretations are to be considered spurious and which instead reflect genuine ambiguity. This is a key difference between GWAP for NLP and other crowdsourcing work.

Collecting multiple judgements about every linguistic expression is a key aspect of PD. In the present version of PD eight players are asked to express their judgements on a markable. If they do not agree on a single interpretation, four more players are then asked to validate each interpretation.[21]

Validation has proven very effective at identifying poor quality interpretations. The value obtained by combining the player annotations with the validations for each interpretation tends to be zero or negative for all spurious interpretations. This formula can also be used to calculate the best interpretation of each expression, which we will refer to in what follows as the *game interpretation*.

Anaphoric judgements can be difficult, and humans will not always agree with each other. For example, it is not always clear from a text whether a markable is referential or not; and in case it is clearly referential, it is not always clear whether it refers to a new discourse entity or an old one, and which one. In PD we are interested in identifying such problematic cases: if a markable is ambiguous, the annotated corpus should capture this information.

5.3.4 Malicious Behaviour

Controlling cheating may be one of the most important factors in GWAP design. If a player is motivated to progress in a game, e.g., by scoring points and attaining levels, they may also become motivated to cheat the system and earn those rewards without completing the tasks as intended.

All crowdsourcing systems attract spammers, which can be a very serious issue [22, 50, 38]. However, in a game context we can expect spamming to be much less of an issue because the work is not conducted on a pay-per-annotation basis.

Nevertheless, several methods are used in PD to identify players who are cheating or who are providing poor annotations. These include checking the player's IP address (to make sure that one player is not using multiple accounts), checking annotations against known answers (the player rating system), preventing players from resubmitting their decisions [16] and keeping a blacklist of players to discard all their data [72].

[21] It is possible for an interpretation to have more annotations and validations than required if a player enters an existing interpretation after disagreeing or if several players are working on the same markables simultaneously.

	System	Good player	Bad player
ANNOTATIONS			
Total Annotations:	1423078	4587	11018
Average Annotation Time:	00:00:07	00:00:07	00:00:04
Total (Ratio) DN:	955520 (0.67)	1495 (0.33)	10935 (0.99)
Total (Ratio) DO:	378256 (0.27)	2696 (0.59)	58 (0.01)
Total (Ratio) PR:	79172 (0.06)	334 (0.07)	24 (0)
Total (Ratio) NR:	13395 (0.01)	64 (0.01)	2 (0)
VALIDATIONS			
Total Validations:	608982	3848	5256
Total (Ratio) Agree:	200174 (0.33)	1186 (0.31)	8 (0)
Ave Agree Time:	00:00:09	00:00:08	00:00:18
Total (Ratio) Disagree:	408808 (0.67)	2662 (0.69)	5248 (1)
Ave Disgree Time:	00:00:08	00:00:07	00:00:02
OTHER			
Total Skips:	51616	142	26
Skip per annotation:	0.04	0.03	0
Total Comments:	26593	229	0
Comment per annotation:	0.02	0.05	0

Fig. 7 Player profiling in Phrase Detectives, showing the game totals and averages (left), a good player profile (centre) and a bad player profile (right) taken from real game profiles. The bad player in this case was identified by the speed of annotations and that the only responses were DN in Annotation Mode and Disagree in Validation Mode. The player later confessed to using automated form completion software.

A method of profiling players was also developed for PD to detect unusual behaviour. The profiling compares a player's decisions, validations, skips, comments and response times against the average for the entire game - see Figure 7. It is very simple to detect players who should be considered outliers using this method (this may also be due to poor task comprehension as well as malicious input) and their data can be ignored to improve the overall quality.

6 Evaluating the Gaming Approach to Creating Language Resources

Evaluating a gaming approach to collaborative resource creation needs to be done in conjunction with other approaches. In order to make things comparable all costs are converted to US$, the lowest level of linguistic labelling is called an *annotation* and an action that the player is asked to perform (that may result in several annotations at once) is called a *task*. To this end, we compare three main areas:

- Participants - How are participants motivated? How much do participants contribute? Do certain participants contribute more?
- Task - How fast is the data being produced? What is the quality of the contributions when aggregated? What is the upper limit of quality that can be expected?
- Implementation - How much does the data collection cost? Which approach represents the best value for money?

The first two areas of comparison correspond to the elements of collective intelligence [46]: the first covering the "who" and "why"; and the latter covering the "how" and "what". The third area of comparison is a more pragmatic view on the approaches, where the choice of approach may be made based on how much budget there is for the data collection, what level of quality is needed for the data to be of use or how much data is required.

6.1 Participants

As previously discussed, participant motivation in collaborative approaches is a very complex issue that has implications on data quality. We consider the case of GWAP without financial incentives, with indirect financial incentives and reported results for other approaches.

6.1.1 Motivating Participation

We measure the success of advertising and the motivation to join the game by how many players have registered over the period of time the game was online. The first version of PD recruited 2000 players in 32 months (62 players/month) and PD on Facebook recruited 612 players in 13 months (47 players/month). JDM recruited 2,700 players in 56 months (48 players/month).

This level of recruitment, whilst not in the same league as the ESP Game which enjoyed massive recruitment in its first few months online, could be seen as what you could expect if some effort was made to advertise a GWAP and motivate people to play it.

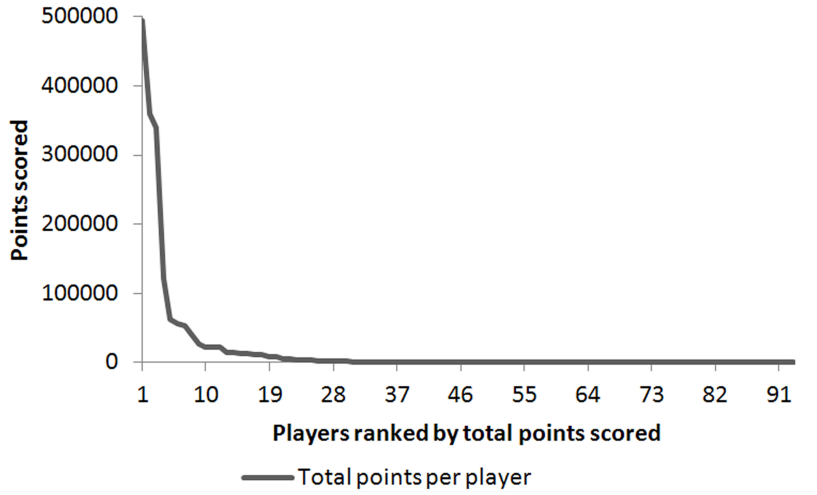

Fig. 8 Chart showing the scores of players (approximately equivalent to workload) in the Phrase Detectives game on Facebook.

There are 5,500 registered reviewers of Wikipedia articles[22], which is the equivalent to a player in a GWAP, however there is an unknown (presumably very large) number of unregistered contributors.

The number of active AMT workers has been estimated as between 15,059 and 42,912 [24]. This explains the difficulty in finding workers with specific skills, such as native speakers for some languages [55], or who can perform large tasks [33].

The total number of participants is useful for evaluating the relative success of recruitment efforts. However, it is not a good predictor of how much work will be done, how fast it will be completed or of what quality it will be. Therefore further analysis of the players themselves is required.

6.1.2 Motivating Contributions

Participation (or volition) of collaborators to contribute is another way to assess whether the incentives of an approach are effective. We measure player motivation to contribute by the average lifetime play. In the case of PD it was 35 minutes (the average sum of all tasks) and in the case of JDM it was 25 minutes (the average length of a session for approximatively 20 games).

The average weekly contribution for Wikipedia is just over 8 hours [54] however this is for contributing users of Wikipedia, not for casual browsers of the website. This indicates that when a user starts contributing to Wikipedia they are highly motivated to contribute. In AMT the contribution rate is a little lower, between 4-6

[22] http://en.wikipedia.org/wiki/Wikipedia:Wikipedians

hours [33], and it can also be expected that the user, once registered, will be highly motivated to contribute.

Obviously, there is a huge complexity and spread of user types within the AMT user base, however it is interesting to note that for 20% of the workers, AMT represents their primary source of income (and for 50%, their secondary source of income), and they are responsible for completing more that one third of all the HITs [32]. Participating for leisure is important for only 30% of workers. So the motivations for participating to AMT are very different from that of Wikipedia.

An observation in most crowdsourcing systems is the uneven distribution of contribution per person, often following a Zipfian power law curve. In PD, it was reported that the ten highest scoring players (representing 1.3% of total players) had 60% of the total points on the system and had made 73% of the annotations [15].

In the Facebook version of PD, the ten highest scoring players (representing 1.6% of total players) had 89% of the total points and had made 89% of the annotations - see Figure 8 [13].

Similarly in JDM the top 10% of the player represents 90% of the activity and studies of AMT also find that only 20% of the users are doing 80% of the work.[23]

These results show that the majority of the workload is being done by a minority of players. However, the influence of players who only contribute a little should not be undervalued as in some systems it can be as high as 30% of the workload [37] and this is what makes the collective decision making robust.

6.1.3 The Effect of Incentives on Participation and Contribution

Further to the figures for motivation and participation, Figure 9 shows the growth of PD on Facebook. Months where there was active promotion of the site (February, July and December 2011) show increases in new players, as one would expect.

Based on the assumption that the first promotion month, when the site went live, was an exception as players of the previous game joined the new version, there is an indication that financial incentives increase recruitment to the game, if sufficiently advertised.

It is noticeable that the number of active players (a player who made more than one annotation in a particular month) stayed consistent and does not seem to increase with recruitment or financial incentives. Whilst it could be expected that the number of active players steadily increases over time as more players are recruited, the results show that most players will play the game for a short period of time and only a small number continue to play every month.

Indirect financial incentives do appear to be a strong motivating factor when considering how much work the active players do. Months with prizes have considerably more new annotations than those without, but with a similar number of active players.

[23] http://groups.csail.mit.edu/uid/deneme/?p=502

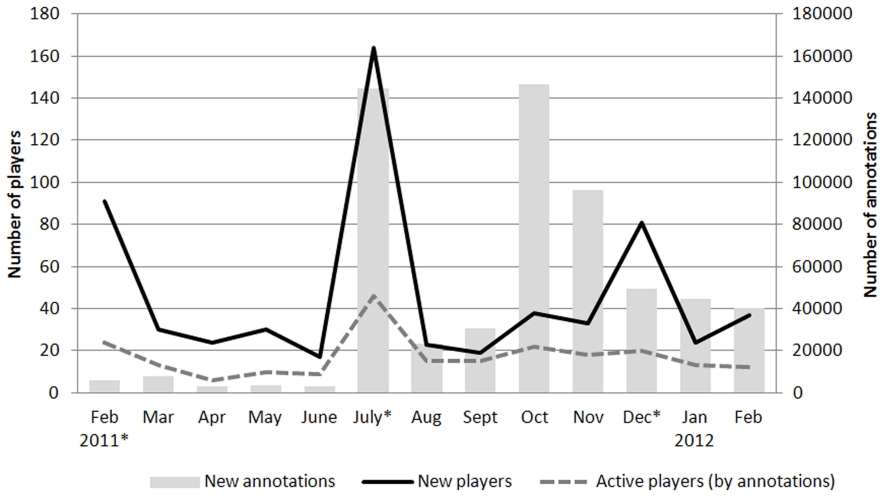

Fig. 9 Chart showing new annotations plotted with new players and active players in Phrase Detectives on Facebook. Prizes were available in the game from July 2011 to February 2012. * indicates a month with active promotion for the game.

This suggests that active players are motivated to contribute more by financial incentives, however the large amount of game play in October and November 2011 indicates that other motivating factors, such as personal and social incentives are, to some extent, also successful. Whilst financial incentives are important to recruit new players, a combination of all three types of incentives is essential for the long term success of a game.

6.1.4 Gender of participants

Crowdsourcing approaches with AMT and games tend to be dominated by female participants. In JDM 60% of players were female. In PD on Facebook female players represented 65% of the total and the top female players contributed significantly more work than the top male players. This suggests that not only are female players more likely to participate, they are also more likely to actively contribute than male players of GWAP.

A survey of AMT workers initially showed a similar gender divide in participants when the system was mainly populated by US workers [33] (due, in part, to payment only being possible to a US bank account). More recent surveys show that the changing demographics of the workers, driven by allowing payment to Indian workers in rupees, now have more male workers from India who use microworking as a primary source of income [64] and the gender split is almost even [33].

The changing demographics of crowdsourcing participants will have an impact on the types of incentives and types of tasks offered. For example a further study

of AMT users performing two tasks showed female dominance, but with preference for word puzzle tasks (74% female) over image sorting tasks (58.8% female) [50].

Conversely, it has been reported that only 12% of contributors to Wikipedia are female [27]. This prompted significant research into the gender bias in the authorship of the site [44].

It has been shown that diverse groups are better at solving tasks and have higher collective intelligence (termed *c*) than more homogeneous groups. A balanced gender divide within a group also produces a higher *c* as females demonstrate higher social sensitivity towards group diversity and divergent discussion [78]. However, this may not have such an impact where the collaboration is indirect.

6.2 Task

6.2.1 Throughput

A measure of efficiency of the interface and task design is how fast tasks are being completed or annotations being generated. This measure is called *throughput*, the number of labels (or annotations) per hour [74].

The throughput of PD is 450 annotations per human hour, which is almost twice as fast as the throughput of 233 labels per human hour reported for the ESP Game.

There is a crucial difference between the two games: PD only requires clicks on pre-selected markables, whereas in the ESP Game the user is required to type in the labels. However, the throughput for JDM is calculated to be 648, where the players also had to type labels, so throughput may also be an indication of task difficulty and cognitive load on the player.

Designers of GWAP who are considering making their task timed should therefore carefully 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 maximize throughput and hence the amount of data that is collected without making the game unplayable.

The throughput of AMT has been reported to be close to real time (within 500ms of a HIT being posted) however this is usually for very simple tasks [7]. More complex tasks can take up to a minute to complete giving a throughput range from 1 to 7,200 labels per hour, while some may never be completed. Whilst these figures are not especially helpful, it highlights the potential speed of this approach if the task can be presented in an efficient way.

Related to throughput is the *wait time* for tasks to be done. Most crowdsourcing systems allow data collection in parallel (i.e., many participants can work at once on the same tasks), although validation requires users to work in series (i.e., where one user works on the output of another user). So whilst the throughput may give us a maximum speed from the system, it is worth bearing in mind that the additional time spent waiting for a user to be available to work on the task may slow the system considerably.

This is where the AMT approach, with a large worker pool, has an advantage and some task requesters even pay workers a retainer to be on demand [5]. With GWAP it is possible to prioritise tasks to maximise completion of corpora, but for open collaboration like Wikipedia it is much more difficult to direct users to areas that need contribution. This can be seen by comparing popular pages that have considerable work, such as for the film Iron Man[24] with 8,000 words, with less popular pages, such as Welsh poetry[25] with only 300 words.

6.2.2 Annotation Quality

Annotation quality is usually assessed by comparing the work to a gold standard or to an expert's opinion. However it is worth noting that there is an upper boundary of quality with these resources as gold standards may occasionally contain errors and experts do not always agree.

In PD agreement between experts is very high although not complete: 94%, for a chance-adjusted κ value [18, 11], of $\kappa = .87$ which can be considered good for coreference tasks [4, 61]. This value can be seen as an upper boundary on what we might get out of the game.

Agreement between experts and the PD game interpretation is also good. We found 84.5% percentage agreement between Expert 1 and the game ($\kappa = .71$) and 83.9% agreement between Expert 2 and the game ($\kappa = .70$). In other words, in about 84% of all annotations the interpretation specified by the majority vote of non-experts was identical to the one assigned by an expert.

These values are comparable to those obtained when comparing an expert with the trained annotators (usually students) that are typically used to create *Traditional, Partly Validated Annotation* resources.

For JDM, there is no similar resource that could be used as gold standard and it is difficult to assign an expert role for common sense knowledge acquisition.

AKI[26], a guessing game, was designed as an indirect evaluation procedure. The goal of the game is to make the system (AKI) guess what the player has in mind from given clues, with the system making a proposal after each clue. The game goes on until the system finds the proper answer or fails to do so. In this task, the AKI system finds the right answer in 75% of the cases. For the same task, humans get the right answer in 48% of cases.

The data used as a knowledge base is strictly the lexical network constructed with JDM, without any modification or preprocessing.

[24] http://en.wikipedia.org/wiki/Iron_Man

[25] http://en.wikipedia.org/wiki/Welsh_poetry

[26] http://www.jeuxdemots.org/AKI.php

Table 1 Agreement on annotations in Phrase Detectives, broken down by annotation type.

	Expert 1 vs. Expert 2	**Expert 1 vs. Game**	**Expert 2 vs. Game**
Overall agreement	94.1%	84.5%	83.9%
Discourse-new (DN) agreement	93.9%	96.0%	93.1%
Discourse-old (DO) agreement	93.3%	72.7%	70.0%
Non-referring (NR) agreement	100.0%	100.0%	100.0%
Property (PR) agreement	100.0%	0.0%	0.0%

6.2.3 Task Difficulty

There is a clear difference in quality when we look at the difficulty of the task in GWAP [12]. Looking separately at the agreement on each type of markable annotation in PD (see Table 1), we see that the figures for a discourse-new (DN) annotation are very close for all three comparisons, and well over 90%. Discourse-old (DO) interpretations are more difficult, with only 71.3% average agreement.

Of the other two types, the 0% agreement between experts and the game on property (PR) interpretations suggests that they are very hard to identify, or possibly the training for that type is not effective. Non-referring (NR) markables on the other hand, although rare, are correctly identified in every single case with 100% precision.

This demonstrates the issue that quality is not only affected by player motivation and interface design but also by the inherent difficulty of the task. As we have seen, users need to be motivated to rise to the challenge of difficult tasks and this is where financial incentives may prove to be too expensive on a large scale.

The quality of the work produced by AMT, with appropriate post-processing, seems sufficient to train and evaluate statistical translation or transcription systems [10, 49]. However, it varies from one task to another according to the parameters of the task. Unsurprisingly, workers seem to have difficulties performing complex tasks, such as the evaluation of summarisation systems [26].

6.3 Implementation

6.3.1 Cost

When evaluating the costs of the different approaches to collaboratively creating language resources, it is important to also consider other constraints, namely the speed at which data needs to be produced, the size of the corpus required, and the quality of the final resource. In order to compare the cost effectiveness we make some generalisations, convert all costs to US$ and calculate an approximate figure for the number of annotations per US$. Where we have factored in wages for software development and maintenance we have used the approximate figure of US$

54,000 per annum for a UK-based post doc research assistant.[27] Additional costs that may be incurred include maintenance of hardware, software hosting, and institutional administrative costs but as these are both difficult to quantify and apply to all approaches they will not be included in the estimates below.

Traditional, Entirely Validated Annotation requires in the order of US$ 1 million per 1 million tokens.[28] On average English texts contain around 1 markable every 3 tokens, so we get a cost of 0.33 markables/US$.

Traditional, Partly Validated Annotation, from estimates of projects by the authors in the UK and Italy, are in the order of US$ 400,000 per 1 million tokens, including the cost of expert annotators. This equates to 0.83 markables/US$.

Both of the above figures are generalisations that include the costs for administering the data collection and developing tools for the task if required. The timescale of data collection is usually several years.

Costs with AMT depend on the amount paid per HIT, which is determined by the task difficulty, the availability of workers with sufficient skills to do the task, and on the extent of redundancy. The literature suggests that US$ 0.01 per HIT is the minimum required for non-trivial tasks, and for a more complex linguistic task like anaphoric co-reference or uncommon language translation, the cost is upwards from US$ 0.1 per HIT. Redundancy for simple tasks is usually around 5 repetitions per task, although in practice we find that 10 repetitions is more likely to be required to attain reasonable quality and filter out poor quality decisions. AMT allows requesters of HITs to set a performance threshold for participants based on whether previous work has been acceptable to other requesters. By using this method there would be less need for redundancy, however the cost of the HIT may need to increase to attract the better workers.

In the case of simple tasks where quality is not a priority the cost would be in the region of 20 markables/US$. In the case of more complicated tasks it would be more like a cost of 1 markable/US$. This is a more conservative estimate than what has previously been cited for early studies with AMT at 84 markables/US$ [67] however, we feel this is more realistic given a more developed microwork platform and workforce.

AMT has the advantage that it is fast to develop, deploy and collect data on a small scale. Typically it may take 1 month for a researcher to create an interface for AMT (US$ 4,500), and perhaps 2 months to collect the data (US$ 9,000) for a small resource.

The advantage of GWAP over other approaches is that once the system is set up, annotations do not cost anything to collect data.

PD took approximately 6 months to develop (US$ 27,000) and a further 3 months to develop the Facebook interface (US$ 13,500). Approximately US$ 9,000 was spent over 38 months in prizes and advertising for the game (approximately US$ 235 per month) with a researcher maintaining the system part-time (at 20%, equivalent to US$ 900 per month, totalling US$ 34,200). 2.5 million annotations have been

[27] http://www.payscale.com/research/UK/Job=Research_Scientist/Salary

[28] This figure was obtained by informally asking several experienced researchers involved in funding applications for annotation projects.

collected by PD, which is gives a cost of 30 annotations/US$. 84,000 markables were completely annotated (although in reality many more were partially annotated) giving a conservative estimate of 1 markables/US$.

JDM took approximately 4 months to develop (US$ 18,000) and was maintained for 54 months by a researcher part-time (at 10%, equivalent to US$ 450 per month, totalling US$ 24,300). JDM did not spend any money on prizes or promotion. During this time 1.3 million individual relations were collected (but not validated in this game) giving an estimate of 53.5 unvalidated markables/US$.

From these estimates it is clear that creating language resources using traditional methods is expensive, prohibitively so beyond 1M words, however the quality is high. This approach is best suited for corpora where the quality of the data is paramount.

AMT for simple tasks is quick to set up and collect data and very cheap, however, more complex tasks are more expensive. The quality of such resources needs more investigation and the approach becomes prohibitively expensive when scaling beyond 10M words. Microworking approaches are therefore most suited for small to medium scale resources, or prototyping interfaces, where noisy data can be filtered.

The GWAP approach is expensive compared to AMT to set up, but the data collection is cheap. In a long term project it is conceivable to collect a 10M+ word corpus, with the main problem being the length of time it would take to collect the data. Over a long period of time the data collection would not only need continuous effort for player recruitment, but also the project requirements may change, requiring further development of the platform. With this in mind, this approach is most suited to a long term, persistent data collection effort that aims to collect very large amounts of data.

6.3.2 Reducing costs

One of the simplest ways of reducing costs is to reduce the amount of data needed and to increase the efficiency of the human computation. Pre-annotation of the data and bootstrapping can reduce the task load, increase the annotation speed and quality [25] and allow participants to work on more interesting tasks that are ambiguous or difficult. Bootstrapping has the downside of influencing the quality of usable output data and errors that exist in the input data multiply when used in crowdsourcing.

This was seen in the PD game, where occasional errors in the pre-processing of a document led to some markables having an incorrect character span. The game allowed players to flag markables with errors for correction by administrators (and to skip the markable if appropriate) however this created a bottleneck in itself. Currently there is no mechanism for players to correct the markables as this would have a profound impact on the annotations that have been collected. JDM did not have these problems as there was no preprocessing in the game.

As can be seen from the cost breakdown of PD, more savings can be made by reusing an existing GWAP platform; the development of the Facebook interface cost half that of the original game.

The advantage of GWAP over microworking is that personal and social incentives can be used, as well as financial, to minimise the cost and maximise the persistence of the system. The use of prizes can motivate players to contribute more whilst still offering value for money as part of a controlled budget.

However, we should be aware that the race towards reducing costs might have a worrying side-effect as short term AMT costs could become the standard. Funding agencies will expect low costs in future proposals and it will become hard to justify funding to produce language resources with more traditional, or even GWAP-based methodologies.

6.3.3 Data size and availability

In JDM, more than 1,340,000 relations between terms have been collected, the sum of the weights being over 150,000,000. More than 150,000 terms have at least one outgoing relation, and more than 120,000 have at least one incoming relation. The current JDM database is available from the game website.

In PD, 407 documents were fully annotated, for a total completed corpus of over 162,000 words, 13% of the total size of the collection currently uploaded for annotation in the game (1.2M words). The first release of the PD Corpus 0.1 [58] is about the size of the ACE 2.0 corpus[29] of anaphoric information, the standard for evaluation of anaphora resolution systems until 2007/08 and still widely used.

The size of the completed corpus does not properly reflect, however, the amount of data collected, as the case allocation strategy adopted in the game privileges variety over completion rate. As a result, almost all the 800 documents in the corpus have already been partially annotated. This is reflected, first of all, in the fact that 84,280 of the 392,120 markables in the active documents (21%) have already been annotated. This is already almost twice the total number of markables in the entire OntoNotes 3.0 corpus[30], which contains 1 million tokens, but only 45,000 markables.

The number of partial annotations is even greater. PD players produced over 2.5 million anaphoric judgements between annotations and validations; this is far more than the number of judgements expressed to create any existing anaphorically annotated corpus. To put this in perspective, the GNOME corpus, of around 40K words, and regularly used to study anaphora until 2007/08, contained around 3,000 annotations of anaphoric relations [56] whereas OntoNotes 3.0 only contains around 140,000 annotations.

Most of the reported resources created on AMT are small to medium size ones [24, 32]. Another issue raised by AMT is the legal status of intellectual property rights of the language resources created on it. Some US universities have insisted on institutional review board approval for AMT experiments.[31]

---

[29] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2003T11

[30] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2009T24

[31] From personal communication with K. Cohen.

7 Conclusions

In this chapter we have considered the successes and the limitations of the GWAP approach to collaboratively creating language resources, compared to traditional annotation methods and more recent approaches such as microwork crowdsourcing and Wikipedia-style open collaboration.

7.1 Game Interface and Task Design

The game interface should be attractive enough to encourage players to start playing and easy enough to use so they keep playing. Before building a GWAP it is essential to have an understanding of game concepts, such as game flow and creating entertaining game scenarios.

The design of the task itself will be determined in part by the complexity of the data being collected. By identifying the difficult or ambiguous tasks, the pre- and post-processing can be improved and the human input can be maximised to produce the highest quality resource possible given the inherent difficulty of the task. Participants may need to be motivated to rise to the challenge of difficult tasks and this is where financial incentives may prove to be too expensive on a large scale.

The task design should be streamlined for efficient collection of data as this is one of the simplest ways of reducing costs: by reducing the amount of data needed. The throughput (annotations per hour) of a GWAP is a good measure of how efficient it is at collecting data, however, it is worth bearing in mind that the additional time spent waiting for a user to be available to work on the task may slow the system.

7.2 Participants and Motivation

Generally speaking, GWAP will use entertainment as the motivating factor rather than direct financial incentives (as found in microwork crowdsourcing). There may also be other motivational considerations, such as the desire to contribute to a scientific project or for self enrichment and learning.

Most the players of GWAP will not benefit directly from the data being collected, however the player connection to the project and sense of contribution to science are strong motivating factors with the citizen science approach, where players are willing to work on harder tasks, provide higher quality data and contribute more.

Controlling cheating may be one of the most important factors in crowdsourcing design and is especially problematic for microworking.

An advantage of GWAP over microworking is that personal and social incentives can be used, as well as financial, to minimise the cost and maximise the persistence of the system. Indirect financial incentives can be a cost-effective way to increase participation in a game.

It is common for the majority of the workload to be done by a minority of players. Motivating the right kind of players is a complex issue, central to the design of the game interface and the task, and is as important as attracting large numbers of players because, although collective intelligence needs a crowd, that crowd also needs to do some work.

The more a player participates in a GWAP, the more expert they become. A system needs to correlate good performance at the task with good quality data and a ratings system (where the user is regularly judged against a gold standard) is essential to give appropriately challenging tasks.

Crowdsourcing approaches with microworking and games tend to be dominated by female participants, although this is not the case for Wikipedia. If crowdsourcing approaches ever hope to produce high quality data, the gender bias needs to be considered as it has been shown that diverse groups are better at solving tasks and have higher collective intelligence than more homogeneous groups.

7.3 Annotation Quality and Quantity

The issue of annotation quality is an area of continuous research. However, results with Phrase Detectives and JeuxDeMots are very promising. The ultimate goal is to show that language resources created using games and other crowdsourcing methods potentially offer higher quality and are more useful by allowing for linguistic ambiguity. By quantifying the complexity of the linguistic tasks, human participants can be challenged to solve computationally difficult problems that would be most useful to machine learning algorithms.

Creating language resources using traditional methods is expensive, prohibitively so beyond 1M words, however the quality is high. Whilst the initial costs of developing a GWAP are high, the game can persistently collect data, making it most suitable for long term, large scale projects. The speed and cost of a microworking approach make it most suitable for collecting small to medium scale resources or prototyping software for larger scale collection, however, some issues of requester responsibility and intellectual property rights remain unresolved.

Approaches that require financial motivation for the participants cannot scale to the size of resources that are now increasingly more essential for progress with human language technology. Only through the contribution of willing participants can very large language resources be created, and only GWAP or Wikipedia-style approach facilitate this type of collaboration.

Acknowledgements We would like to thank Jean Heutte (CREF-CNRS) for his help with the concepts of game flow and for the comments of the reviewers of this chapter. The contribution of Karën Fort to this work was realized as part of the Quæro Programme[32], funded by OSEO, French State agency for innovation. The original Phrase Detectives game was funded as part of the EPSRC AnaWiki project, EP/F00575X/1.

[32] http://quaero.org/

References

1. A. Aker, M. El-haj, D. Albakour, and U. Kruschwitz. Assessing crowdsourcing quality through objective tasks. In *Proceedings of LREC'12*, 2012.
2. O. Alonso and S. Mizzaro. Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. In *Proceedings of SIGIR '09: Workshop on The Future of IR Evaluation*, 2009.
3. O. Alonso, D. E. Rose, and B. Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15, 2008.
4. R. Artstein and M. Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.
5. M. S. Bernstein, D. R. Karger, R. C. Miller, and J. Brandt. Analytic methods for optimizing realtime crowdsourcing. *CoRR*, 2012.
6. V. Bhardwaj, R. Passonneau, A. Salleb-Aouissi, and N. Ide. Anveshan: A tool for analysis of multiple annotators' labeling behavior. In *Proceedings of the 4th linguistic annotation workshop (LAW IV)*, 2010.
7. J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White, and T. Yeh. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on user interface software and technology*, UIST '10, 2010.
8. H. Bonneau-Maynard, S. Rosset, C. Ayache, A. Kuhn, and D. Mostefa. Semantic annotation of the French media dialog corpus. In *Proceedings of InterSpeech*, 2005.
9. C. Callison-Burch. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 2009.
10. C. Callison-Burch and M. Dredze. Creating speech and language data with Amazon's Mechanical Turk. In *CSLDAMT '10: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.
11. J. Carletta. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22:249–254, 1996.
12. J. Chamberlain, U. Kruschwitz, and M. Poesio. Constructing an anaphorically annotated corpus with non-experts: Assessing the quality of collaborative annotations. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, 2009.
13. J. Chamberlain, U. Kruschwitz, and M. Poesio. Motivations for participation in socially networked collective intelligence systems. In *Proceedings of CI2012*, 2012.
14. 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)*, 2008.
15. J. Chamberlain, M. Poesio, and U. Kruschwitz. 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)*, 2009.
16. T. Chklovski and Y. Gil. Improving the design of intelligent acquisition interfaces for collecting world knowledge from web contributors. In *Proceedings of K-CAP '05*, 2005.
17. T. Chklovski. Collecting paraphrase corpora from volunteer contributors. In *Proceedings of K-CAP '05*, 2005.
18. J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.
19. M. Csikszentmihalyi. *Flow : The Psychology of Optimal Experience*. Harper and Row, 1990.
20. S. Dandapat, P. Biswas, M. Choudhury, and K. Bali. Complex linguistic annotation - No easy way out! A case from Bangla and Hindi POS labeling tasks. In *Proceedings of the 3rd ACL Linguistic Annotation Workshop*, 2009.
21. C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

22. D. Feng, S. Besana, and R. Zajac. Acquiring high quality non-expert knowledge from on-demand workforce. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, 2009.
23. F. Fenouillet, J. Kaplan, and N. Yennek. Serious games et motivation. In *4eme Conference francophone sur les Environnements Informatiques pour l'Apprentissage Humain (EIAH'09), vol. Actes de lAtelier "Jeux Serieux: conception et usages"*, 2009.
24. K. Fort, G. Adda, and K. B. Cohen. Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics (editorial)*, 37:413–420, 2011.
25. K. Fort and B. Sagot. Influence of pre-annotation on POS-tagged corpus development. In *Proceedings of the 4th ACL Linguistic Annotation Workshop (LAW)*, 2010.
26. D. Gillick and Y. Liu. Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.
27. R. Glott, P. Schmidt, and R. Ghosh. Wikipedia survey – Overview of results. *UNU-MERIT*, pages 1–11, 2010.
28. N. Green, P. Breimyer, V. Kumar, and N. F. Samatova. Packplay: Mining semantic data in collaborative games. In *Proceedings of the 4th Linguistic Annotation Workshop*, 2010.
29. B. Hladká, J. Mírovskỳ, and P. Schlesinger. Play the language: Play coreference. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, 2009.
30. J. Hong and C. F. Baker. How good is the crowd at "real" WSD? In *Proceedings of the 5th Linguistic Annotation Workshop*, 2011.
31. J. Howe. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. Crown Publishing Group, 2008.
32. P. Ipeirotis. Analyzing the Amazon Mechanical Turk marketplace. CeDER Working Papers, http://hdl.handle.net/2451/29801, September 2010.
33. P. Ipeirotis. Demographics of Mechanical Turk. CeDER Working Papers, http://hdl.handle.net/2451/29585, March 2010.
34. 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 6th International Conference on Artificial Life*, 1998.
35. A. Joubert and M. Lafourcade. Jeuxdemots : Un prototype ludique pour l'émergence de relations entre termes. In *Proceedings of JADT'2008, Ecole normale supérieure Lettres et sciences humaines*, 2008.
36. D. Jurafsky and J. H. Martin. *Speech and Language Processing- 2nd edition*. Prentice-Hall, 2008.
37. B. Kanefsky, N. Barlow, and V. Gulick. Can distributed volunteers accomplish massive data analysis tasks? *Lunar and Planetary Science*, XXXII, 2001.
38. G. Kazai. In search of quality in crowdsourcing for search engine evaluation. In *Proceedings of the 33rd European Conference on Information Retrieval (ECIR'11)*, 2011.
39. G. Kazai, N. Milic-Frayling, and J. Costello. Towards methods for the collective gathering and quality control of relevance assessments. In *Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval*, 2009.
40. A. Koller, K. Striegnitz, A. Gargett, D. Byron, J. Cassell, R. Dale, J. Moore, and J.Oberlander. Report on the 2nd NLG challenge on generating instructions in virtual environments (GIVE-2). In *Proceedings of the 6th INLG*, 2010.
41. R. Koster. *A Theory of Fun for Game Design*. Paraglyph, 2005.
42. M. Lafourcade. Making people play for lexical acquisition. In *Proceedings SNLP 2007, 7th Symposium on Natural Language Processing*, 2007.
43. M. Lafourcade and A. Joubert. A new dynamic approach for lexical networks evaluation. In *Proceedings of LREC'12: 8th International Conference on Language Resources and Evaluation*, 2012.
44. D. Laniado, C. Castillo, A. Kaltenbrunner, and M. Fuster-Morell. Emotions and dialogue in a peer-production community: The case of Wikipedia. In *Proceedings of the 8th International Symposium on Wikis and Open Collaboration (WikiSym'12)*, 2012.

45. H. Lieberman, S. D. A., and A. Teeters. Common consensus: A web-based game for collecting commonsense goals. In *Proceedings of IUI*, 2007.
46. T. Malone, R. Laubacher, and C. Dellarocas. Harnessing crowds: Mapping the genome of collective intelligence. Research Paper No. 4732-09, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, USA, February 2009.
47. A. Marchetti, M. Tesconi, F. Ronzano, M. Rosella, and S. Minutol. SemKey: A semantic collaborative tagging system. In *Proceedings of WWW 2007 Workshop on Tagging and Metadata for Social Information Organization*, 2007.
48. M. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus of English : The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.
49. M. Marge, S. Banerjee, and A. I. Rudnicky. Using the Amazon Mechanical Turk for transcription of spoken language. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010.
50. W. Mason and D. J. Watts. Financial incentives and the "performance of crowds". In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, 2009.
51. D. R. Michael and S. L. Chen. *Serious Games: Games That Educate, Train, and Inform*. Muska & Lipman/Premier-Trade, 2005.
52. R. Mihalcea and T. Chklovski. Open Mind Word Expert: Creating large annotated data collections with web users help. In *Proceedings of the EACL 2003 Workshop on Linguistically Annotated Corpora (LINC 2003)*, 2003.
53. 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*, 2008.
54. O. Nov. What motivates Wikipedians? *Communications of the ACM*, 50(11):60–64, 2007.
55. S. Novotney and C. Callison-Burch. Cheap, fast and good enough: Automatic speech recognition with non-expert transcription. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010.
56. M. Poesio. Discourse annotation and semantic annotation in the GNOME corpus. In *Proceedings of the ACL Workshop on Discourse Annotation*, 2004.
57. M. Poesio and R. Artstein. Anaphoric annotation in the ARRAU corpus. In *LREC'08*, 2008.
58. M. Poesio, J. Chamberlain, U. Kruschwitz, L. Robaldo, and L. Ducceschi. The Phrase Detective multilingual corpus, release 0.1. In *Proceedings of LREC'12 Workshop on Collaborative Resource Development and Delivery*, 2012.
59. M. Poesio, J. Chamberlain, U. Kruschwitz, L. Robaldo, and L. Ducceschi. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Interactive Intelligent Systems*, 2012 forthcoming.
60. M. Poesio, P. Sturt, R. Arstein, and R. Filik. Underspecification and anaphora: Theoretical issues and preliminary evidence. *Discourse Processes*, 42(2):157–175, 2006.
61. M. Poesio and R. Vieira. A corpus-based investigation of definite description use. *Computational Linguist*, 24(2), 1998.
62. A. Quinn and B. Bederson. Human computation: A survey and taxonomy of a growing field. *CHI*, 2011.
63. W. Rafelsberger and A. Scharl. Games with a purpose for social networking platforms. In *Proceedings of the 20th ACM conference on Hypertext and Hypermedia*, 2009.
64. J. Ross, L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson. Who are the crowdworkers?: Shifting demographics in Mechanical Turk. In *Proceedings of CHI EA '10*, 2010.
65. K. Siorpaes and M. Hepp. Games with a purpose for the semantic web. *IEEE Intelligent Systems*, 23(3):50–60, 2008.
66. F. Smadja. Mixing financial, social and fun incentives for social voting. *World Wide Web Internet And Web Information Systems*, 2009.
67. R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2008.
68. J. Surowiecki. *The Wisdom of Crowds*. Anchor, 2005.

69. P. Sweetser and P. Wyeth. Gameflow: A model for evaluating player enjoyment in games. *Computer Entertainment*, 3, July 2005.
70. S. Thaler, K. Siorpaes, E. Simperl, and C. Hofer. A survey on games for knowledge acquisition. Technical Report STI TR 2011-05-01, Semantic Technology Institute, 2011.
71. S. Tratz and E. Hovy. A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010.
72. L. von Ahn. Games with a purpose. *Computer*, 39(6):92–94, 2006.
73. L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2004.
74. L. von Ahn and L. Dabbish. Designing games with a purpose. *Communications of the ACM*, 51(8):58–67, 2008.
75. L. von Ahn, R. Liu, and M. Blum. Peekaboom: a game for locating objects in images. In *Proceedings of CHI '06*, 2006.
76. P. Wais, S. Lingamneni, D. Cook, J. Fennell, B. Goldenberg, D. Lubarov, D. Marin, and H. Simons. Towards building a high-quality workforce with Mechanical Turk. In *Proceedings of Computational Social Science and the Wisdom of Crowds (NIPS)*, 2010.
77. A. Wang, C. D. V. Hoang, and M. Y. Kan. Perspectives on crowdsourcing annotations for natural language processing. *Language Resources And Evaluation*, pages 1–19, July 2010.
78. A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, and T. W. Malone. Evidence for a collective intelligence factor in the performance of human groups. *Science*, 330:686–688, 2010.
79. H. Yang and C. Lai. Motivations of Wikipedia content contributors. *Computers in Human Behavior*, 26, 2010.
80. M. Yuen, L. Chen, and I. King. A survey of human computation systems playing / having fun. *Information Sciences*, 2009.

Appendix A

Categories of GWAP with links where available.

Image annotation	
ESP Game	http://www.gwap.com/gwap/gamesPreview/espgame
Matchin	http://www.gwap.com/gwap/gamesPreview/matchin
FlipIt	http://www.gwap.com/gwap/gamesPreview/flipit
Phetch	http://www.peekaboom.org/phetch
Peekaboom	http://www.peekaboom.org
Squigl	http://www.gwap.com/gwap/gamesPreview/squigl
Magic Bullet	http://homepages.cs.ncl.ac.uk/jeff.yan/mb.htm
Picture This	http://picturethis.club.live.com
Video annotation	
OntoTube	http://ontogame.sti2.at/games
PopVideo	http://www.gwap.com/gwap/gamesPreview/popvideo
Yahoo's VideoTagGame	http://sandbox.yahoo.com/VideoTagGame
Waisda	http://www.waisda.nl
Audio annotation	
Herd It	http://apps.facebook.com/herd-it
Tag a Tune	http://www.gwap.com/gwap/gamesPreview/tagatune
WhaleFM	http://whale.fm
Biomedical	
Foldit	http://fold.it/portal
Phylo	http://phylo.cs.mcgill.ca
EteRNA	http://eterna.cmu.edu
Transcription	
Ancient Lives	http://ancientlives.org
Old Weather	http://www.oldweather.org
Search results	
Page Hunt	http://pagehunt.msrlivelabs.com/PlayPageHunt.aspx
Social bookmarking	
Collabio	http://research.microsoft.com/en-us/um/redmond/groups/cue/collabio

Appendix B

Categories of GWAP used for NLP with links where available.

Knowledge acquisition	
1001 Paraphrases	
LEARNER	
FACTory	http://game.cyc.com
Verbosity	http://www.gwap.com/gwap/gamesPreview/verbosity
Categorilla	http://www.doloreslabs.com/stanfordwordgame/categorilla.html
Free Association	http://www.doloreslabs.com/stanfordwordgame/freeAssociation.html
Text annotation	
Phrase Detectives	http://www.phrasedetectives.com
Phrase Detectives on Facebook	http://apps.facebook.com/phrasedetectives
PlayCoref	
PhraTris	http://galoap.codeplex.com
PackPlay	
Sentiment analysis	
Sentiment Quiz	http://apps.facebook.com/sentiment-quiz
Generation	
GIVE games	http://www.give-challenge.org
Ontology building	
JeuxDeMots	http://www.jeuxdemots.org
AKI	http://www.jeuxdemots.org/AKI.php
OntoGame	http://ontogame.sti2.at/games