Proceedings of the LREC 2020 Workshop "Citizen Linguistics in Language Resource Development", pages 26–34
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Speaking Outside the Box: Exploring the Benefits of Unconstrained Input in Crowdsourcing and Citizen Science Platforms

Jon Chamberlain[a], Udo Kruschwitz[b] & Massimo Poesio[c]
[a] University of Essex, Wivenhoe Park, Colchester, Essex UK. jchamb@essex.ac.uk
[b] Universität Regensburg, 93040 Regensburg, Germany. udo.kruschwitz@ur.de
[c] Queen Mary University of London, Mile End Rd, Bethnal Green, London UK. m.poesio@qmul.ac.uk

Abstract

Crowdsourcing approaches provide a difficult design challenge for developers. There is a trade-off between the efficiency of the task to be done and the reward given to the user for participating, whether it be altruism, social enhancement, entertainment or money. This paper explores how crowdsourcing and citizen science systems collect data and complete tasks, illustrated by a case study from the online language game-with-a-purpose Phrase Detectives. The game was originally developed to be a constrained interface to prevent player collusion, but subsequently benefited from posthoc analysis of over 76k unconstrained inputs from users. Understanding the interface design and task deconstruction are critical for enabling users to participate in such systems and the paper concludes with a discussion of the idea that social networks can be viewed as form of citizen science platform with both constrained and unconstrained inputs making for a highly complex dataset.

Keywords: crowdsourcing, citizen science, unconstrained, interface design, verbatim, input type, natural language interface

1. Introduction

The popularity of crowdsourcing approaches in recent years, encompassing everything from microworking to citizen science and all systems in between, has proved a difficult design challenge for system developers. Primarily such systems are designed to collect, label or in some way engage human participants in solving problems that cannot be done computationally (and to help train systems to perform tasks better). There is a trade-off between the efficiency of the task to be done and the reward given to the user for participating, whether it be altruism, social enhancement, entertainment or money. This trade-off is key to ensuring systems work for both the *requester* (the party that wants the task to be completed) and the *worker* (the party that does the task). From the point of view of the requester, the most efficient way to collect the data required is to constrain the worker to a pre-defined set of responses that can be easily processed, aggregated and analysed, with poor performing users identified against a gold standard and excluded from contributing. However, from the point of view of the worker, the pre-defined set of solution options may be ambiguous and they may not be able to fully express their intent and solution to the task.

In a toy example, consider a theatre booking website that requires a user to enter a date to book a ticket for a show. The *requester* (the theatre) requires a date (the task) to be entered into the system so it can be matched to a date in the database of remaining tickets for sale and automatically processed to issue the ticket. Hence, a set of predefined dropdown select boxes are offered to the user on the booking form (or an interactive calendar selection popup). The result is that the user can only enter a date that the system can recognise. However, the user may find that the constrained input does not allow them to query the system in a way they would find natural, for example, they may wish to use natural language to express their intent ('tomorrow', 'next Monday', or 'the first Saturday in June') or provide an ambiguous answer more aligned to their intention, e.g., 'next Saturday but if fully booked then the Saturday after'. In the trade-off between precise booking and user experience, the former approach is more commonly used than the latter, although the rise of chatbots for a more personalised booking experience may indicate the beginnings of a paradigm shift to a more human-centred interface (Elsholz et al., 2019).

This paper explores how crowdsourcing and citizen science systems collect data and complete tasks by characterising the type of task and style of interface used in popular systems (Section 2). Section 3 presents a case study of research from the online language game-with-a-purpose *Phrase Detectives*, originally developed to be a constrained interface to prevent player collusion, but subsequently benefited from posthoc analysis of unconstrained input from users. Section 4 generalises further how the interface design and task deconstruction are critical for enabling users to participate in such systems and explores the idea that social networks can be viewed as form of citizen science platform with both constrained and unconstrained inputs making for a highly complex dataset.

2. Related Work

Crowdsourcing (Howe, 2008) has become ubiquitous in systems where tasks need to be completed by human workers that are too difficult for computers to perform accurately. This section provides a brief overview of the most common types of crowdsourcing systems and characterises them by how the task is processed.

Peer production Peer production is a way of completing tasks that relies on self-organising communities of individuals in which effort is coordinated towards a shared outcome (Benkler and Nissenbaum, 2006). The willingness of Web users to collaborate in peer production can be seen in the creation of resources such as Wikipedia. English Wikipedia numbers (as of Feb 2020) over 6M articles, con-

tributed to by over 38M users.[1] The key aspects that make peer production so successful are the openness of the data resource being created and the transparency of the community that is creating it (Lakhani et al., 2007; Dabbish et al., 2014).

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 (Yang and Lai, 2010). This motivation is also behind the success of **citizen science** projects, such as the *Zooniverse* collection of projects[2], in which the scientific research is conducted mainly by amateur scientists and members of the public (Clery, 2011). The costs of ambitious data annotation tasks are also kept to a minimum, with expert annotators only required to validate a small portion of the data (which is also likely to be the data of most interest them).

Question answering systems attempt to learn how to answer a question automatically from a human, either from structured data or from processing natural language of existing conversations and dialogue. Here we are more interested in **Community Question Answering (cQA)**, in which the crowd is the system that attempts to answer the question through natural language. Examples of cQA are sites such as StackOveflow[3] and Yahoo Answers.[4] Detailed schemas (Bunt et al., 2012) and rich feature sets (Agichtein et al., 2008) have been used to describe cQA dialogue and progress has been made to analyse this source of data automatically (Su et al., 2007).

Microworking Amazon Mechanical Turk[5] pioneered microwork crowdsourcing by using the Web as a way of reaching large numbers of workers (often 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. A reported advantage of microworking 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 in which the worker needs to be more skilled, e.g. translating a sentence in an uncommon language, it can take much longer (Novotney and Callison-Burch, 2010). Microwork crowdsourcing is becoming a standard way of creating small-scale resources, but is prohibitively expensive to create large-scale resources.

Gaming and games-with-a-purpose 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 people spend billions of hours playing games on the Web (von Ahn, 2006). A game-with-a-purpose (GWAP) can 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.

Social computing and social networks Social computing has been described as 'applications and services that facilitate collective action and social interaction online with rich exchange of multimedia information and evolution of aggregate knowledge' (Parameswaran and Whinston, 2007). It encompasses technologies that enable communities to gather online such as blogs, forums and social networks, although the purpose is largely not to solve problems directly. The open dialogue and self-organising structure of social networks[6] allow many types of human interaction, but here we are most interested in the idea of community problem solving, in which one user creates a task and the community solves it for them. As social networks mature the software is utilised in different ways, with decentralised and unevenly-distributed organisation of content, similar to how Wikipedia users create pages of dictionary content. Increasingly, social networks are being used to organise data, to pose problems, and to connect people who may have solutions that can be contributed in a simple and socially-convenient fashion. Facebook has been used as a way of connecting professional scientists and amateur enthusiasts with considerable success (Sidlauskas et al., 2011; Gonella et al., 2015). However, there are drawbacks with this method of knowledge sharing and problem solving: data may be lost to people interested in them in the future and they are often not accessible in a simple way, for example, with a search engine.

2.1. Features of crowdsourcing tasks

Crowdsourcing approaches can be distinguished by features related to the task. To clarify why these features apply to a particular approach an exemplar system is chosen for the approach that is perhaps the most prevalent or successful: Manual annotation is considered the benchmark where the task is completed by an expert; *GalaxyZoo* represents citizen science (although a detailed typology for citizen science projects also exists (Wiggins and Crowston, 2011)); *StackOverflow* represents Community Question Answering (cQA); *Wikipedia*'s main website is an example of a wiki-type approach; for microworking, *Amazon Mechanical Turk* is used; for GWAPs, the *ESP game* is used; and finally for social networks, *Facebook* itself is considered (rather than a system implemented on the platform).

The type of task that is presented covers the dimension of *how* the problem gets solved (Malone et al., 2009). One of the important features for distinguishing individual projects (rather than the approach) is to look at **task difficulty**, either as a function of the task (*routine, complex* or *creative* (Schenk and Guittard, 2011)) or as a function of worker *cognitive load* (Quinn and Bederson, 2011). Also useful for distinguishing between projects is the **centrality** of the

[1] `http://meta.wikimedia.org/wiki/List_of_Wikipedias`, accessed 18/2/2020.

[2] `https://www.zooniverse.org`

[3] `http://stackoverflow.com`

[4] `https://uk.answers.yahoo.com`

[5] `http://www.mturk.com`

[6] For the context of this paper we define a social network as the platform for communication, rather than a system deployed on the platform or the social network structure itself.

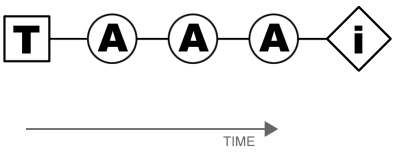

Figure 1: A task *T* can be completed in series in which each annotation *A* is dependent on the one before and leads to one interpretation *i* (Wikipedia, cQA and social networks).

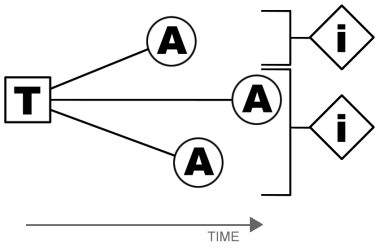

Figure 2: *T* can also be completed in parallel in which annotations can be entered simultaneously leading to multiple interpretations that require post-processing for a final output (microworking, GWAPs and manual annotation).

crowdsourcing in the system, i.e. is the crowdsourcing *core* to the system, such as creating content in Wikipedia, or is it *peripheral* such as rating articles (Organisciak and Twidale, 2015). Task features are discussed below and summarised in Table 1.

Input constraint Whilst data are often structured, mainly to allow them to be input into the system, the contributions may not necessarily be. Crowdsourcing typically constrains workers to enter a restricted range of inputs via radio buttons and dropdown lists, whereas social networks and peer production allow unconstrained text input that requires post-processing. Some tasks require annotations to be aligned to an ontology and this provides structure; however, spelling mistakes and ambiguity can cause errors. Along with unconstrained page creation, Wikipedia allows for semi-constrained input through summary boxes on each page. The choice of input constraint may be driven by a further facet of whether the answers to the task need to be *objective* or *subjective* (Organisciak and Twidale, 2015).

Input order The timing of the presentation of the tasks is dependent on the system and, generally speaking, will determine how fast a system can produce an output for a task. In the case of Wikipedia, cQA and social networks, a task is added and each worker contributes in series, i.e. each contribution is dependent on the previous contributions in the way a Wikipedia page is developed or a conversation thread flows (see Figure 1). Workers on Wikipedia can edit and overwrite the text on a page. This 'last edit wins' approach is fundamental to building the content; however, contentious subjects may cause 'edit wars' and pages may become locked to prevent future editing.

In order to increase crowdsourcing efficiency, some systems allow tasks to be completed in parallel, i.e. multiple workers annotate different tasks at different times meaning that not all tasks will be completed in the same amount of time (see Figure 2). Parallel tasks are common in microworking, GWAPs and citizen science. Expert manual annotation can be completed both in series or in parallel.

A wider, systematic view of task order would be to view the system's **procedural order** and how the worker interacts with system inputs and responses from the crowd (Organisciak and Twidale, 2015; Chamberlain and O'Reilly, 2014).

Validation Quality control of a system is a feature of most typologies of crowdsourcing and can be used to distinguish between different projects (Quinn and Bederson, 2011; Das and Vukovic, 2011); however, it creates a large and complex facet group that is beyond the scope of what is required here. In this context, it is the reviewers of the annotations supplied by the workers that is of interest.

Validation on some level occurs after annotations have been applied to the data; the issue is whether those validations are part of the process that the workers are involved in or whether it is a form of checking from the requester to ensure that a sample of the annotations are of a high enough quality. It is typically the case for requesters to check a sample of annotations with experts, microworking and citizen science. In systems such as Wikipedia, social networks and cQA, the checking and validation of all answers is done by the workers themselves. GWAP annotations are typically validated by the requester; however, an increasing proportion of games are using validation as an additional worker task to reduce the workload for the requester (Chamberlain et al., 2018).

3. Case Study: Phrase Detectives

Phrase Detectives[7] is an online citizen science game designed to collect data about English anaphoric coreference (Chamberlain et al., 2008; Poesio et al., 2013).[8]

3.1. Constrained input

The game uses two styles of constrained text annotation for players to complete the linguistic task. Initially text is presented in **Annotation Mode** (called Name the Culprit in the

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

[8] Anaphoric coreference is a type of linguistic reference where one expression depends on another referential element. An example would be the relation between the entity *'Jon'* and the pronoun *'his'* in the text '*Jon rode his bike to school.*'

Table 1: A table showing task features, including whether the input is constrained, in what order it can be entered and who checks it.

	Input constraint	Input order	Validation by
Expert annotation	Constrained	Both	Requester
Peer production: Citizen science	Constrained	Parallel	Requester
GWAP	Constrained	Parallel	Both
Microworking	Constrained	Parallel	Requester
Peer production: Wikipedia	Unconstrained	Series	Worker
Peer production: cQA	Unconstrained	Series	Worker
Social Networks	Unconstrained	Series	Worker

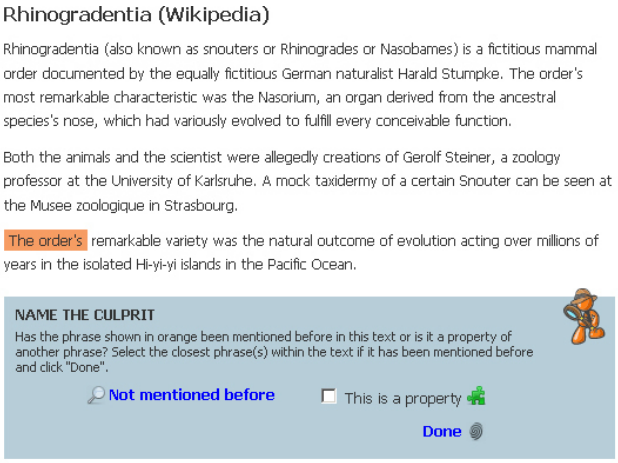

Figure 3: Constrained input (Annotation Mode) for players of *Phrase Detectives*.

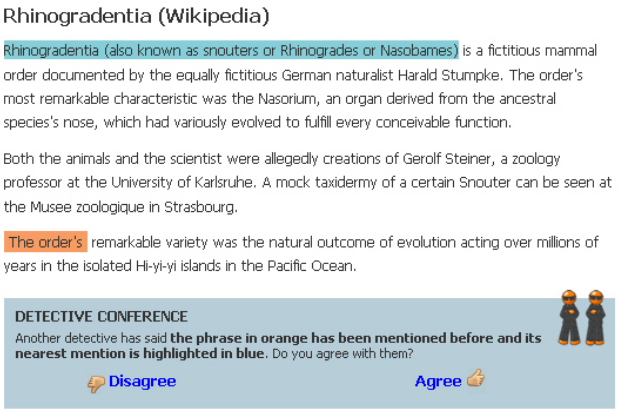

Figure 4: Constrained input (Validation Mode) for players of *Phrase Detectives*.

game, see Figure 3). This is a traditional annotation method in which the player makes an **interpretation** (annotation decision) about a highlighted **markable** (section of text). Markables are identified using pre-processing and are a defined set of options within the context of text shown to the player. Players can select multiple markable antecedents if they believe the anaphor is plural. Players can also select options without selecting a markable, e.g., to indicate the markable has not been mentioned before in the text. Al-

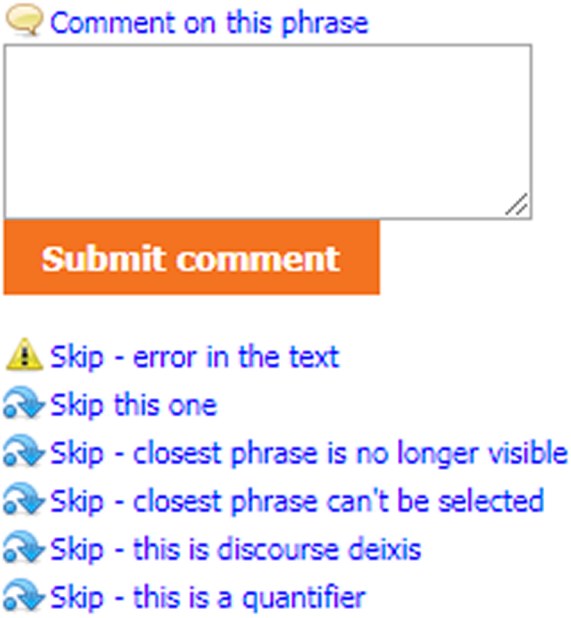

Figure 5: Unconstrained input options during Annotation Mode for players of *Phrase Detectives*.

though the number of possible interpretations players could enter is very large, in practice players converge on sensible interpretations for the task.

If different players enter different interpretations for a markable then each interpretation is presented to more players in a constrained, binary task **Validation Mode** (called Detectives Conference in the game, see Figure 4). The players in Validation Mode have to agree or disagree with the interpretation. If they disagree, their decision is recorded and they are then presented with Annotation Mode for the same markable.

This method of data collection was originally designed into the game to reduce collusion between the players during a gameplay (von Ahn and Dabbish, 2008), whilst rewarding players who made the effort to put in good quality solutions to the task.

3.2. Unconstrained input

During early prototyping of the game it became clear that players were encountering tasks they could not complete with the set of constrained inputs on offer. The most com-

mon at the time was to indicate that the pre-processing of markables contained an error, either in the boundary of the tokens or that the markable was not a noun phrase. For this reason an unconstrained input option was added to Annotation Mode (also accessible from Validation Mode by disagreeing with the interpretation) to allow players to indicate that something was wrong or what they couldn't express with the limited set of options available in the game (see Figure 5).

For player convenience, several 'skip' buttons were shown that allow the player to quickly skip the task but also to indicate why in a single click. By clicking a skip option, a 'skip' event is created in the database; if the skip option had a reason a 'comment' event was additionally created in the database. The full range of unconstrained player responses were:

1. **Comment on this phrase** A freetext comment that when submitted does not conclude the task, i.e, the player can also add a solution or skip;
2. **Skip - error in the text** Skip the task because the markable has an error;
3. **Skip this one** Skip the task but not provide a reason why (no comment is created);
4. **Skip - closest phrase no longer visible** Skip the task because the player has seen the solution in a previous part of the text that is no longer accessible;
5. **Skip - closest phrase can't be selected** Skip the task because although the phrase the player wants to select is in the text it is not one of the predefined markables (and this also occurs when markables are embedded in larger markables, such as in the case of apposition.);
6. **Skip - this is discourse deixis** Discourse deixis is a relatively easy linguistic phenomenon for players to identify but there was no way to mark it as a solution to the task (this was added due to player requests);
7. **Skip - this is a quantifier** As above, players could easily identify solutions to tasks that were quantifiers but did not have the option to mark it as such (again, added due to player requests).

3.3. Consolidation of Unconstrained Input

The constrained inputs from the players have been analysed in several ways, initially using majority voting for a collective decision making (Chamberlain et al., 2018), then with more advanced modelling through Mention-Pair Analysis (MPA) (Poesio et al., 2019). However, these techniques did not make use of any of the unconstrained data collected from the players.

In order to make the unconstrained data into a more useful form it was consolidated semi-automatically (see Figure 6) and included in the corpora released for further research (Poesio et al., 2019). Each comment was classified initially by the player (by the type of skip they select) and then by an administrator. The administrator can then take action in relation to the comment, e.g., correcting markable boundaries

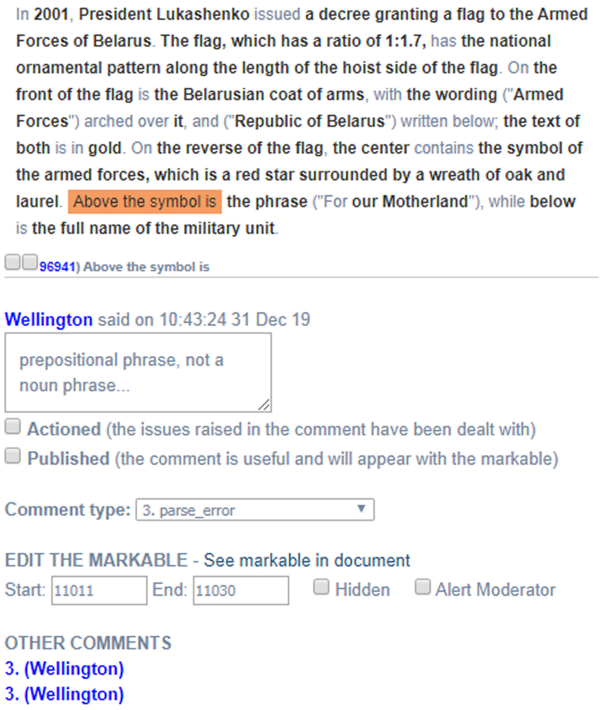

Figure 6: Admin screen in *Phrase Detectives* that allows reviewers to process the unconstrained input of players.

Table 2: A breakdown of comments received in *Phrase Detectives*, in which *Skip* relates to the type of skip made in the interface.

Classification	Skip	Comments
Not selectable	[5]	31,846
Out of context window	[4]	21,732
Parse error	[2]	15,707
Discourse deixis	[6]	328
Ambiguous		49
Non-referring		24
Nearest mention embedding		237
Bridging reference		11
Quantifier	[7]	50
Unclassified		6,899
TOTAL		76,883

(which is flagged in a checkbox) and/or publish the comment with the corpus (in fact, all comments are published in the corpus, this flag is an indication that the administrator thought the comment was useful). Links to other comments on the same markable can be seen so they can all be dealt with at the same time.

3.4. Data

As of 18 Feb 2020 there were 114,353 skips and 76,883 comments added by players of *Phrase Detectives*, in comparison to 3,179,850 annotation and 1,420,191 validation decisions, from a total of 60,965 players working on 843 documents. A breakdown of each comment type can been seen in Table 2. The ratio of skips to annotations per player is approx. 4% and comments to annotations is approx. 2%.

3.5. Uses of Unconstrained Data

The most immediate use of the skip and comment functionality in *Phrase Detectives* was to elicit feedback from the players regarding errors in the corpus and interface design problems. The skip data was incorporated as a way to determine whether players should stop being given a markable because there was something wrong with it. Comments regarding pre-processing errors, markables not being available to be selected or beyond the piece of text visible to the player account for the majority of comments from users.

The way players provided unconstrained input to the system in this way enabled the development of specific functionality for a small group of high performing players who wanted to provide more detailed solutions to the tasks. For example, these players frequently used the comment field to indicate markables where discourse deixis or quantifier was the most appropriate interpretation by commenting 'DD' and 'QQ' respectively. By creating their own annotation input (likely based on other annotation schemes) the players were providing a level of input to the system that was beyond what the interface was designed for. Based on these comments, additional skip types were added to the interface to enable these players to provide this input faster during their gameplay.

The verbatim comments allowed us to understand some interesting and ambiguous phenomena encountered in the data that could only have been understood with posthoc analysis. Issues of context, plural union and separation, bridging, naming conventions, temporal revelations, measurements, dates, and generality/specificity were all addressed using the comment functionality giving administrators a unique understanding into why player decision making diverged from consensus.

In addition to manual posthoc analysis, the skips and comments are being developed into future versions of the MPA algorithm (Poesio et al., 2019), used to detect emergent communities of players who respond to stimuli in different ways. Anaphoric resolvers that analyse complex, ambiguous datasets (like those created by *Phrase Detectives*) using neural network approaches may perform better due to the richness of multi-dimensional data at their disposal.

3.6. A Fully Unconstrained Interface?

To conclude our case study of how unconstrained input was gathered from players of *Phrase Detectives*, we report on two efforts that were made to create interfaces that were entirely unconstrained (due to the platform limitations, rather than design requirements).

An attempt was made to emulate the anaphoric coreference task in *Phrase Detectives* using microworking; however, this proved to be very difficult as the users were restricted to entering an imprecise text notation, for example having to write *DO line 2 "the door"* for a highlighted markable or using two inputs to select the class of relation and where the antecedent is (see Figure 7).

In the hope of leveraging the social networking platform Facebook's community of users, an unconstrained version of the task was presented through a user group called Anaphor from your Elbow, a contraction of the question *Do you know your anaphor from your elbow?*, (see Figure 8) where an image of the language task was posted and the users commented on the image as to where the antecedent was in the text (in the same way as above).

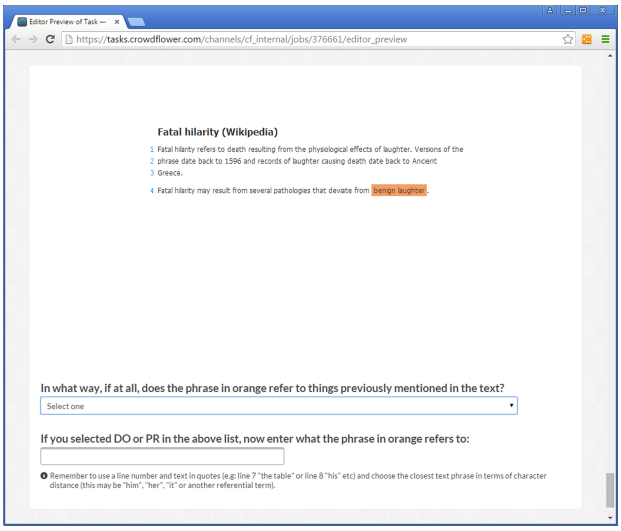

Figure 7: Screenshot of the anaphoric coreference task presented in Crowdflower.

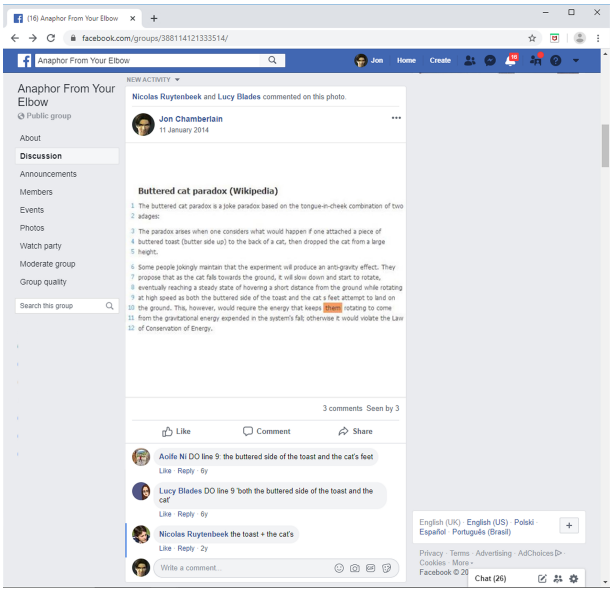

Figure 8: Screenshot of the Anaphora from your Elbow Facebook group where the unconstrained anaphoric coreference task was presented.

Given the difficulties of pre-formatting the text as an image, as well as post-processing the unconstrained comments from the users, these experiments were abandoned in favour of developing the constrained game interface of *Phrase Detectives* to incorporate more unconstrained input from players.

4. Discussion

4.1. Interface Design

The design of the interface will determine how successfully the user can contribute data to a crowdsourcing system. In *Phrase Detectives* 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 interaction with the game). The pre-processing of text allows the interface to be constrained in this way, but is subject to errors in pre-processing that must also be fixed.

The interface of microworking sites is also predefined and presents limitations that constitute an important issue for some tasks, for example, in annotating noun compound relations using a large taxonomy (Tratz and Hovy, 2010). In a word sense disambiguation task, considerable redesigns were required to get satisfactory results (Hong and Baker, 2011). These examples show how difficult it is to design tasks for crowdsourcing within a predefined system. The design of social network interfaces is dictated by the owners of the platforms, rather than the requester or the community of users and crowdsourcing efforts may be in conflict with other revenue-generating activities such as advertising.

The interface 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 (Aker et al., 2012) and response times (Chamberlain and O'Reilly, 2014). Errors in the data constitute wasted effort and should be dealt with by bug testing the system rather than post-processing.

4.2. Task Difficulty

Crowdsourcing and citizen science can produce high-quality work from users, comparable to work of an expert, if communities of users can be found to do the task. The task of anaphoric coreference as used in *Phrase Detectives* is not simple and, although the majority of tasks were not hard, it is the uncommon difficult tasks that require the power of human computation. A less-constrained environment allows these difficult tasks to be solved in more organic ways compared to a fully constrained system.

There is a clear difference in quality when we look at the difficulty of the tasks in *Phrase Detectives*. Looking separately at the agreement on each class of markable annotation, we observe near-expert quality for the simple task of identifying discourse-new (DN) markables, whereas discourse-old (DO) markables are more difficult (Chamberlain et al., 2016). This demonstrates that quality is not only affected by player motivation and interface design but also by the inherent difficulty of the task. Users need to be motivated to rise to the challenge of difficult tasks and this is when financial incentives may prove to be too expensive on a large scale.

The quality of the work produced by microworking, with appropriate post-processing, seems sufficient to train and evaluate statistical translation or transcription systems (Callison-Burch and Dredze, 2010; Marge et al., 2010). However, it varies from one task to another according to the defining parameters. Unsurprisingly, workers seem to have difficulty performing complex tasks, such as the evaluation of summarisation systems (Gillick and Liu, 2010).

A task may be difficult for several reasons: the correct answer is difficult, but not impossible, to determine; the true interpretation is a difficult type of solution to determine; or that the answer is genuinely ambiguous and there is more than one plausible solution. The latter tasks can be rare, but are of the most interest to computational linguists and machine learning algorithms. In these cases the users need to have a thorough understanding of how to add their solutions and an unconstrained input option would capture data beyond what the interface may have been designed for; however, automatically processing these cases can be difficult.

4.3. Citizen Science on Social Networks

Social networking sites such as Facebook, Twitter and Instagram have all been used for conducting citizens science activities. Harnessing the collective intelligence of communities on social networks is not straightforward, but the rewards are high. If a suitable community can be found to align with the task of the requester and the data can be extracted from the network, it has shown to be a useful type of crowdsourcing approach. Aggregating the social network data in a similar way to crowdsourcing (Chamberlain, 2014) will allow the automatic extraction of knowledge and sophisticated crowd aggregation techniques (Raykar et al., 2010) can be used to gauge the confidence of data extracted from threads on a large scale.

A validation model is intuitive to users and features in some form on most social network platforms. Typically a 'like' or 'upvote' button can be found on messages and replies, allowing the community to show favour for particular solutions, and this method has been shown to be effective and efficient in experimental work (Chamberlain, 2014). Other forms of voting exist, such as full validation (like and dislike) or graded voting (using a five star vote system) allowing for more fine-grained analysis of the community's preference; however, further research is needed to assess whether this is actually a waste of human effort and a simple like button proves to be the most effective (Chamberlain et al., 2018).

In most crowdsourcing and citizen science systems users are rewarded for agreement and not punished for being disagreed with; however, other scoring models of this kind do exist (Rafelsberger and Scharl, 2009). It seems intuitive that positive behaviour be reinforced in crowdsourcing to encourage participation.

4.4. Limitations and Challenges

One drawback to offering unconstrained inputs is that users use them in different ways. There is a risk of accounts being used for malicious content, spreading advertising or for spamming. Users have different expectations that may lead to segregation into groups and data not being entered in a fashion that is expected. A significant challenge for unconstrained methods is the automatic processing of the threads (Maynard et al., 2012). There are a large quantity of unnecessary data associated with unconstrained inputs and removing this overhead is essential when processing on a large scale. The natural language processing needs to cope with ill-formed grammar and spelling, and sentences for which only context could make sense of the meaning. Additionally, the automatic processing of sentiment on poorly formed text is also challenging, with negative and compound assertions causing problems for automatic processing.

5. Conclusion

This paper explored how crowdsourcing and citizen science systems collect data and complete tasks, illustrated by a case study from the online language game-with-a-purpose *Phrase Detectives*. Understanding the interface design and task deconstruction are critical for enabling users to participate in such systems. Processing unconstrained input from users has applications within crowdsourcing and citizen science system design to allow users to express their solutions when they are beyond what the system was designed to collect. It would also enable efforts on a larger scale by analysing highly complex datasets created though social networking platforms.

6. Acknowledgements

The authors would like to thank all the players who played the game. The creation of the original game was funded by EPSRC project AnaWiki, EP/F00575X/1. The analysis of the data and preparation of this paper was funded by the DALI project, ERC Grant 695662

7. Bibliographical References

Agichtein, E., Castillo, C., Donato, D., Gionis, A., and Mishne, G. (2008). Finding high-quality content in social media. In *Proceedings of the 1st ACM International Conference on Web Search and Data Mining (WSDM'08)*, pages 183–194.

Aker, A., El-haj, M., Albakour, D., and Kruschwitz, U. (2012). Assessing crowdsourcing quality through objective tasks. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*.

Benkler, Y. and Nissenbaum, H. (2006). Commons-based peer production and virtue. *Journal of Political Philosophy*, 14(4):394–419.

Bunt, H., Alexandersson, J., Choe, J.-W., Fang, A. C., Hasida, K., Petukhova, V., Popescu-Belis, A., and Traum, D. (2012). ISO 24617-2: A semantically-based standard for dialogue annotation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, may.

Callison-Burch, C. and Dredze, M. (2010). Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010) Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT '10)*.

Chamberlain, J. and O'Reilly, C. (2014). User performance indicators in task-based data collection systems. In *Proceedings of the 2014 iConference workshop MindTheGap'14*.

Chamberlain, J., Poesio, M., and Kruschwitz, U. (2008). Phrase Detectives: A web-based collaborative annotation game. In *Proceedings of the 2008 International Conference on Semantic Systems (I-Semantics'08)*.

Chamberlain, J., Poesio, M., and Kruschwitz, U. (2016). Phrase detectives corpus 1.0 crowdsourced anaphoric coreference. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, may.

Chamberlain, J., Kruschwitz, U., and Poesio, M. (2018). Optimising crowdsourcing efficiency: Amplifying human computation with validation. *it - Information Technology*.

Chamberlain, J. (2014). Groupsourcing: Distributed problem solving using social networks. In *Proceedings of 2nd AAAI Conference on Human Computation and Crowdsourcing (HCOMP'14)*.

Clery, D. (2011). Galaxy evolution. Galaxy Zoo volunteers share pain and glory of research. *Science*, 333(6039):173–5.

Dabbish, L., Stuart, H. C., Tsay, J., and Herbsleb, J. D. (2014). Transparency and coordination in peer production. *Computing Research Repository (CoRR)*, abs/1407.0377.

Das, R. and Vukovic, M. (2011). Emerging theories and models of human computation systems: A brief survey. In *Proceedings of the 2nd International Workshop on Ubiquitous Crowdsouring (UbiCrowd'11)*, pages 1–4.

Elsholz, E., Chamberlain, J., and Kruschwitz, U. (2019). Exploring language style in chatbots to increase perceived product value and user engagement. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, CHIIR '19, page 301–305, New York, NY, USA. Association for Computing Machinery.

Gillick, D. and Liu, Y. (2010). Non-expert evaluation of summarization systems is risky. In *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010) Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT '10)*.

Gonella, P., Rivadavia, F., and Fleischmann, A. (2015). Drosera magnifica (Droseraceae): the largest New World sundew, discovered on Facebook. *Phytotaxa*, 220(3):257–267.

Hong, J. and Baker, C. F. (2011). How good is the crowd at "real" WSD? In *Proceedings of the 5th Linguistic Annotation Workshop (LAW V)*.

Howe, J. (2008). *Crowdsourcing: Why the power of the crowd is driving the future of business*. Crown Publishing Group.

Lakhani, K. R., Jeppesen, L. B., Lohse, P. A., and Panetta, J. A. (2007). The value of openness in scientific problem solving. Working Paper 07-050, Harvard Business School.

Malone, T., Laubacher, R., and Dellarocas, C. (2009). Harnessing crowds: Mapping the genome of collective intelligence. Research Paper No. 4732-09, Sloan School of Management, Massachusetts Institute of Technology, February.

Marge, M., Banerjee, S., and Rudnicky, A. I. (2010). Using the Amazon Mechanical Turk for transcription of spoken language. In *Proceedings of the 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP'10)*.

Maynard, D., Bontcheva, K., and Rout, D. (2012). Chal-

lenges in developing opinion mining tools for social media. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12) Workshop @NLP can u tag #user_generated_content*.

Novotney, S. and Callison-Burch, C. (2010). Cheap, fast and good enough: Automatic speech recognition with non-expert transcription. In *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*.

Organisciak, P. and Twidale, M. (2015). Design facets of crowdsourcing. In *Proceedings of the 2015 iConference*.

Parameswaran, M. and Whinston, A. B. (2007). Social computing: An overview. *Communications of the Association for Information Systems*, 19.

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., and Ducceschi, L. (2013). Phrase Detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Interactive Intelligent Systems*, 3(1):1–44, April.

Poesio, M., Chamberlain, J., Paun, S., Yu, J., Uma, A., and Kruschwitz, U. (2019). A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1778–1789, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Quinn, A. J. and Bederson, B. B. (2011). Human computation: A survey and taxonomy of a growing field. In *Proceedings of the 2011 SIGCHI Conference on Human Factors in Computing Systems (CHI'11)*, pages 1403–1412.

Rafelsberger, W. and Scharl, A. (2009). Games with a purpose for social networking platforms. In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L. (2010). Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322.

Schenk, E. and Guittard, C. (2011). Towards a characterization of crowdsourcing practices. *Journal of Innovation Economics*, 7:93–107.

Sidlauskas, B., Bernard, C., Bloom, D., Bronaugh, W., Clementson, M., and Vari, R. P. (2011). Ichthyologists hooked on Facebook. *Science*, 332(6029):537.

Su, Q., Pavlov, D., Chow, J.-H., and Baker, W. C. (2007). Internet-scale collection of human-reviewed data. In *Proceedings of the 16th International Conference on World Wide Web (WWW'07)*, pages 231–240.

Tratz, S. and Hovy, E. (2010). A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10)*.

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

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

Wiggins, A. and Crowston, K. (2011). From conservation to crowdsourcing: A typology of citizen science. In *Proceedings of the 44th Hawaii International Conference on System Sciences (HICSS'11)*, pages 1–10.

Yang, H. and Lai, C. (2010). Motivations of Wikipedia content contributors. *Computers in Human Behavior*, 26.