Games with a Purpose or Mechanised Labour?
A Comparative Study

Marta Sabou
Dpt. of New Media Technology
MODUL University Vienna
Marta.Sabou@modul.ac.at

Kalina Bontcheva
Dpt. of Computer Science
University of Sheffield
K.Bontcheva@dcs.shef.ac.uk

Arno Scharl
Dpt. of New Media Technology
MODUL University Vienna
Arno.Scharl@modul.ac.at

Michael Föls
Vienna University of
Economics and Business
michael.foels@wu.ac.at

ABSTRACT

Mechanised labour and games with a purpose are the two most popular human computation genres, frequently employed to support research activities in fields as diverse as natural language processing, semantic web or databases. Research projects typically rely on either one or the other of these genres, and therefore there is a general lack of understanding of how these two genres compare and whether and how they could be used together to offset their respective weaknesses. This paper addresses these open questions. It first identifies the differences between the two genres, primarily in terms of cost, speed and result quality, based on existing studies in the literature. Secondly, it reports on a comparative study which involves performing the same task through both genres and comparing the results. The study's findings demonstrate that the two genres are highly complementary, which not only makes them suitable for different types of projects, but also opens new opportunities for building cross-genre human computation solutions that exploit the strengths of both genres simultaneously.

Categories and Subject Descriptors

H.5.3 [**Group and Organization Interfaces**]: Collaborative computing; I.2.1 [**Applications and Expert Systems**]: Games; I.2.6 [**Learning**]: Knowledge Acquisition

General Terms

Measurement, Experimentation, Human Factors

Keywords

Human Computation, Games with a Purpose, Mechanised Labour, Comparison

i-KNOW'13 September 04 - 06 2013, Graz, Austria
Copyright 2013 ACM 978-1-4503-2300-0/13/09 ...$15.00

1. INTRODUCTION

Although Human Computation (HC), as a computing paradigm, has received various definitions [10], a largely accepted definition is that of von Ahn stating that HC is *"... a paradigm for utilizing human processing power to solve problems that computers cannot yet solve"* [16]. To that end, HC techniques typically engage large populations of human users and therefore are an important family of techniques for harvesting collective intelligence. Since HC typically focuses on problems that are not yet solvable by computers, it has become a useful instrument in a variety of scientific disciplines concerned with building intelligent algorithms, including natural language processing (NLP) [11], speech processing [8] or the semantic web [13]. HC methods typically help to gather training data for these algorithms, to perform tasks that are too difficult for the algorithms or to evaluate the algorithms' output [11].

The two most popular HC genres are mechanised labour and games with a purpose [10]. *Mechanised labour (MLab)* is a type of paid-for HC genre, where contributors choose to carry out small tasks (or micro-tasks) and are paid a small amount of money in return (often referred to as micro-payments). Popular platforms for mechanised labour include Amazon's Mechanical Turk (MTurk) and CrowdFlower(CF) which allow requesters to post their micro-tasks in the form of Human Intelligence Tasks (or HITs, or units) to a large population of micro-workers. *Games with a purpose(GWAP)* enable human contributors to carry out computation tasks as a side effect of playing online games [17]. An example from the area of computational biology is the Phylo game (`phylo.cs.mcgill.ca`) that disguises the problem of multiple sequence alignment as a puzzle like game [4].

In the early days of using HC in research projects for harvesting collective intelligence, adopters of such techniques focused on proving that HC methods produced comparable results to those obtainable with experts or traditionally hired and trained contributors [2, 14]. Later on, some turned their attention to comparing various HC genres with the goal of helping practitioners in choosing one of these genres [18] or for emphasizing the benefits of one genre over the other [1]. Common to these earlier studies is that their

Feature	MLab	GWAP	References
Cost			
Set-up Price	Low(+)	High(-)	[9, 15, 18]
Price per task	Low(-)	None(+)	[9, 15]
Speed			
Set-up Time	Low (+)	High(-)	[9, 15, 18]
Throughput	High(+)	Low(-)	[1]
Throughput predictability	High(+)	Low(-)	[1, 15]
Quality			
Quality	Low(-)	High(+)	[1, 18]
	High(+)	High(+)	[15]
Maintaining motivation	Easy(+)	Difficult(-)	[15]
Incentive to cheat	High(-)	(Mostly) Low (+)	[1, 18]
Task complexity	Low(-)	High(+)	[1]
Importance of task interestingness	Low(+)	High(-)	[15, 20]
Worker diversity	Low(-)	High(+)	[15]
	High(+)	Low(-)	[18]
Other			
Ethical issues	Yes(-)	(Mostly) No(+)	[3]

Table 1: Advantages and disadvantages of mechanised labour and GWAPs.

findings lack objective grounding since they are derived from literature review rather than from comparing the two genres under similar conditions. As a side effect, while they identify some (obvious) differences between these genres, they typically do (and can) not provide indications of the actual range of the differences (e.g., how much is mechanised labour cheaper/faster than a game-based approach?). Our hypothesis is that, these two HC genres are highly complementary and that this can be used to build hybrid systems that benefit from the strengths of both approaches simultaneously. However, for building such systems, a more in-depth understanding of the genres differences is necessary, over a variety of different tasks.

This paper investigates genre differences, based on a shared task, namely that of defining semantic relations between concept pairs: e.g., establishing that $is\ a\ sub-category\ of$ holds between *coal* and $fossil\ fuel$. This task underlies many important semantic web problems such as ontology learning or matching. The paper starts with a discussion of the differences between the two genres collected from state-of-the-art literature (Section 2). These differences are then confirmed through a direct, comparison-based approach by running the same task through both genres. Concretely, the Climate Quiz GWAP is used to collect pair relations (Section 3) then a similar interface for the same task is built on the CrowdFlower mechanised labour platform (Section 4). Section 5 compares the results in terms of the criteria described in Section 2. Our aim is not only to confirm the genre differences, but also to report on the typical proportion of the differences between them. The findings reported here complement those of [15], a recent work that compares the performance of games against mechanised labour. The key differences with respect to that work are: (1) the type of task: instance classification in [15] vs. relation detection in this case and (2) a more detailed comparison of the two genres, which in this paper is also aligned with previous observations from the literature. Conclusions and an outlook to future work are presented in Section 6.

2. GAMES OR MECHANISED LABOUR?

The question of how games and mechanised labour compare has received limited coverage to date. The few papers on this topic include [18] which adopts a survey-based approach, as well as [1] which highlights the successes and limitations of games for language resource creation, as compared to characteristics of mechanised labour. Thaler et al. [15] compare the two HC genres by applying them to solve a shared task. Their approach, therefore, differs from [1, 18] by providing experimentally-grounded conclusions on the performance of the two genres. The different aspects of HC covered by these studies are combined and summarised in this section to get a complete picture of the differences between the two HC genres. The focus of the analysis is on the key dimensions of knowledge creation projects: cost, speed and data quality, as discussed next and summarized in Table 1.

Costs. Projects based on mechanised labour have low initial setup costs, since they reuse the platform's job creation and monitoring tools. They also allow performing tasks for very small amounts of money, however, since typically multiple judgments must be collected for each task for quality assurance purposes, the acquisition price for large resources can still be significant. In contrast, games tend to have high up-front costs to implement their user and management interfaces, but then allow gathering data virtually for free [9, 15]. Poesio and colleagues [9] take a close look at the cost reductions enabled by HC genres in the case of linguistic resources, on the scale of 1M tokens. They estimate that, compared to the cost of expert-based annotation (estimated as $1,000,000), the cost of 1M annotated tokens could be indeed reduced to less than 50% by using MTurk (i.e., $380,000 - $430,000) and to around 20% of the expert- based price (i.e., $217,927) when using GWAPs, such as their own PhraseDetectives game. Therefore, mechanised labour is more cost effective for quick and affordable acquisition of small-scale datasets, while GWAPs can make larger content creation projects more affordable, thanks to their very low ongoing maintenance costs.

Speed. HC projects use throughput (the amount of data created per human hour) to measure the speed of data creation. Chamberlain and colleagues [1] report throughputs of 450 and 648 tasks per hour for the two annotation GWAPs they describe, however, these speeds remain far behind the almost real-time completion of tasks on MTurk. Thaler et al. have also shown that the time needed to run the same experiment with the OntoPronto game was double to that needed with MTurk [15]. Indeed, paid-for HC has the advantage of a faster and more predictable completion time, since projects tap into an already existing, large labour pool. In contrast, completion times of GWAPs are often slower and can be less predictable, as they depend on the ability to recruit, retain, and motivate a large number of players.

Quality. Opinions differ about the quality of results obtainable by the two HC genres. On the one hand, survey-based approaches found that, in general, higher quality results can be obtained with games [18], particularly in the area of word-sense disambiguation [1]. On the other hand, Thaler and colleagues have shown that the quality of results from micro-workers is similar to that obtained with their game [15]. Eckert and colleagues [2] reached a similar conclusion, when comparing the quality of mechanised labour against that of data obtained from volunteers (which are similar to game players, due to their intrinsic motivation).

There are various factors that influence the quality of data obtainable through HC. In games, maintaining a motivated player base is difficult and often requires choosing (even manually) only interesting tasks (e.g., ontologies in a domain of interest - [20]). Micro-workers, on the other hand, are motivated extrinsically by pay and will accept tasks independently of their level of interestingness, thus being suitable for a broader range of projects. Chamberlain et al. [1] observe, however, that micro-workers have difficulties in performing complex tasks such as the evaluation of summarisation systems, which might otherwise be feasible with a stable player population, that can be trained on a particular task. The extrinsic motivation of micro-workers has however downsides, namely that they are more likely to cheat to obtain an economic benefit than players who play for fun [18]. A final quality related issue is data bias. Statistics from MTurk [3] and GWAPs [9] have shown that a small number of people carry out a large number of tasks (paid HITs or hours playing), which, if the aim is to have more diverse data, from different people, might bias the results. Compared to paid-for marketplaces, GWAPs promise superior results, not only due to their intrinsically motivated players but also by making better use of sporadic, explorer-type users. In fact, recent studies show that games may provide a larger variety of contributors and can reach more individuals than MTurk [8]. Similarly, Thaler et al. [15] found that their game reached out to a larger player base (270) than MTurk micro-workers (only 16).

Ethical and legal issues related to HC are an increasingly hot topic. The use of mechanised labour (MTurk in particular) raises a number of worker right issues, such as: low wages (below $2 per hour), lack of worker rights, and legal implications of using it for longer-term projects [3].

The findings above lead to the conclude that there is a significant complementarity between the two genres, along all key dimensions (cost, speed, quality) and that this fact could be leveraged for building hybrid HC systems that exploit the benefits of both genres simultaneously. For example, complex, interesting tasks could be performed by a dedicated, well-trained player base (on a longer term and virtually for free), while more "boring" tasks that would reduce the motivation of players might be more suitable for execution by intrinsically motivated micro-workers, for a small amount of money. Starting from these hypotheses, this paper aims to quantify these genre differences through a comparative study that involves performing the same task with the Climate Quiz game on the one hand and through a similar mechanised labour interface, on the other.

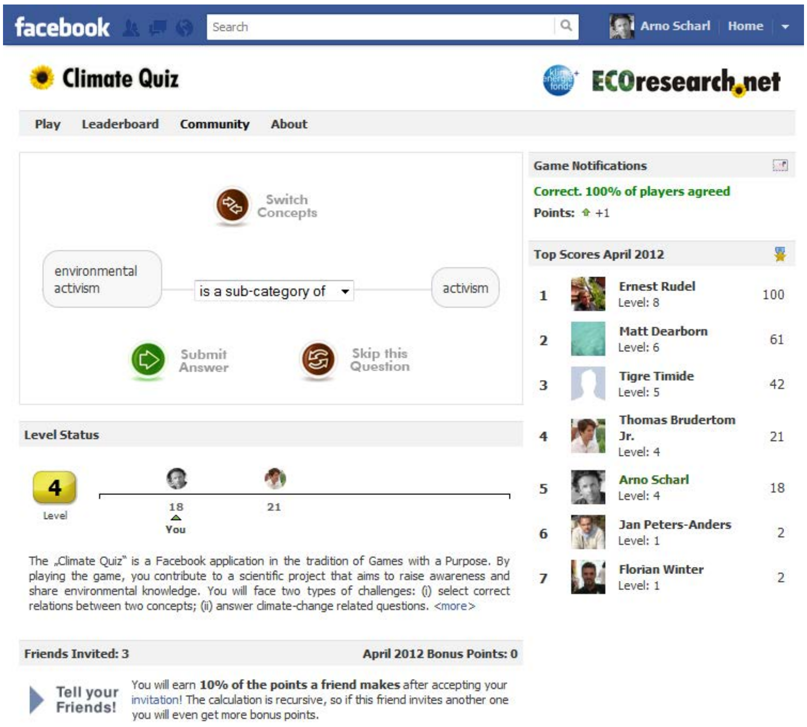
3. CLIMATE QUIZ

Climate Quiz (`apps.facebook.com/climate-quiz/`) is a game with a purpose deployed over the Facebook social networking platform. It is focused on acquiring factual knowledge in the domain of climate change (see detailed description in [12]). The game is coupled with an ontology learning algorithm, as follows [19]. The ontology learning algorithm extracts terms from unstructured and structured data sources. The term pairs that are most likely related based on the algorithm's input data sources are subsequently sent to Climate Quiz, where players assign relations to each pair. These relations are fed back into the algorithm which uses them to refine the learned ontology and to derive new term pairs that should be connected.

As depicted in Figure 1, Climate Quiz asks players to evaluate whether two concepts presented by the system are related (e.g. $environmental\ activism, activism$), and which label is the most appropriate to describe this relation (e.g. $is\ a\ sub-category\ of$). Players can assign one of eight relations, three of which are generic ($is\ a\ sub-category\ of$, $is\ identical\ to$, $is\ the\ opposite\ of$), whereas five are domain-specific ($opposes$, $supports$, $threatens$, $influences$, $works\ on\ /with$). Two further relations, *other* and *is not related to* were added for cases not covered by the previous eight relations. The game's interface allows players to switch the position of the two concepts or to skip ambiguous pairs.

Participants earn one point for each matching answer, but can also lose points if their opinion differs from the majority of players (therefore "random" relation selection is discouraged). If in doubt, the system awards a point in order not to discourage players - if the first user selects relation A, for example, and the second user selects B, both receive a point since a majority solution has yet to be determined. If the first two players have answered A, however, the answer of a third player who does not agree with them will be considered wrong. Participants are given immediate feedback about each answer in terms of the percentage of players who agreed/disagreed with their decision as well as the majority voted relation if the player's answer differs from it (top right corner). This feedback constitutes a continuous player training mechanism during the game and increases transparency by explaining how points are awarded.

A consensus about a pair is reached when the most popular relation between its terms has 4 more votes than the second most popular relation. If a consensus is not reached after a



Correct. 100% of players agreed

Points: +1

1	Ernest Rudel Level: 8	100
2	Matt Dearborn Level: 6	61
3	Tigre Timide Level: 5	42
4	Thomas Brudertom Jr. Level: 4	21
5	Arno Scharl Level: 4	18
6	Jan Peters-Anders Level: 1	2
7	Florian Winter Level: 1	2

The „Climate Quiz" is a Facebook application in the tradition of Games with a Purpose. By playing the game, you contribute to a scientific project that aims to raise awareness and share environmental knowledge. You will face two types of challenges: (i) select correct relations between two concepts; (ii) answer climate-change related questions. <more>

Friends Invited: 3 April 2012 Bonus Points: 0

You will earn **10% of the points a friend makes** after accepting your invitation! The calculation is recursive, so if this friend invites another one you will even get more bonus points.

Figure 1: Climate Quiz user interface.

maximum of 10 individual judgments are collected, then the pair is discarded as unresolvable. Between April and October 2012, 1,213 concept pairs have been assessed through Climate Quiz, and a consensus was reached for 424 out of these. The remaining 789 pairs are still being shown in the game, as they have yet to reach a consensus, even though they have gathered 3,464 individual judgments already. A comparison of the game results against a gold standard data set of 147 pairs, showed an agreement level of 72%.

4. A MECHANISED LABOUR VERSION OF CLIMATE QUIZ

In order to allow the comparative analysis of the two HC genres, a mechanised labour version of Climate Quiz was created on the CrowdFlower (CF) platform.

Task Design. The mechanised labour task's interface was created with CF's Builder tool in such a way that it resembled the interface of Climate Quiz as much as allowed by the design facilities of the platform (see Figure 2). Similarly to Climate Quiz, each task (or unit in CF terminology) presented the workers with two terms and asked them to select the correct relation between them by choosing between the 8 possible relations or the *other/is not related to* relations. The relations were presented always in the same order, and following the order in which they were presented in Climate Quiz. Unlike Climate Quiz, the CF interface did not offer a "skip" option, as such judgments would not have contributed towards a decision and would have offered an easy way to cheat (i.e., a worker could select skip for all tasks and still receive payments, since CF payments cannot be made dependent on the value of an answer). Implementing a term switching functionality was not possible with the interface creation facilities of CF. Alternative workarounds such as including both the direct and indirect relations as a single option (e.g. $is\ a\ sub/super\ category\ of$, $influences/is\ influenced\ by$) or having different options for the direct and indirect relations would have resulted in more options, higher task complexity and more differences to the GWAP interface. The fact that the skip functionality was missing from the CF interface was taken into account during the results' evaluation.

While various techniques exist to filter out invalid responses after the completion of a mechanised labour task, it is preferable to prevent cheating in the first place. Task interface design plays a key role here. It has been shown experimentally (e.g. [5, 6]), that extending task interfaces with explicitly verifiable questions forces workers to process the content of the task and also signals to them that their answers are being scrutinized (e.g., asking the workers to type in a word from the processed document or the number of references that a Wikipedia article has). This seemingly simple technique had a significant positive effect on the quality of the collected data [5, 6]. Therefore, the task also included two verification questions that could only be answered correctly if the workers actually read the two terms. One question requested workers to type in the second letter of the first term, while the other required them to provide the number of words that made up the second term. Since the goal of these questions is to force workers to read the terms, it was important to place them before the relation selection part of the task, even with the risk of placing the relation selection "further" away from the compared terms.

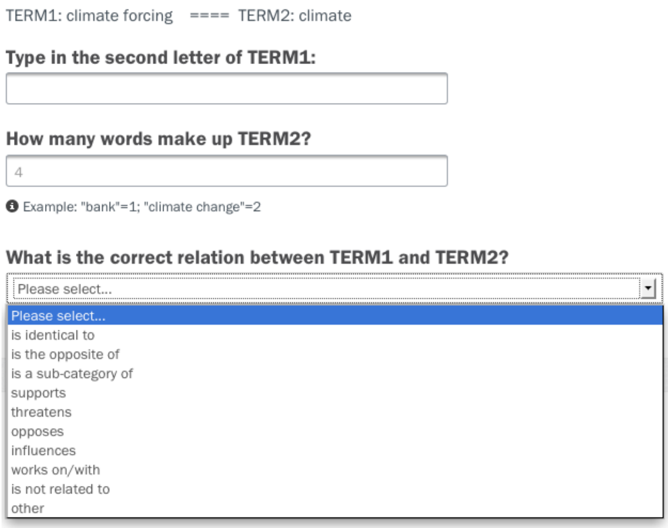
Additionally, the following methods were used to ensure the quality of the results. Firstly, detailed instructions of how to perform the task were provided, including many examples of correct and incorrect relations between terms. These were part of the task interface (not shown in Figure 2) and therefore always accessible to workers. Secondly, the task input data (147 pairs) was augmented with the recommended 5% of gold units, that is 8 gold units. CrowdFlower uses these gold units to train the workers on the go, but also to detect low-performing players early in the data acquisition process and to exclude their work automatically from the final result without payment. For example, for our jobs, the system automatically eliminated all judgments of players who failed on 4 gold units.

Pilot. A trial version of the task was run with a small amount of the data (20 randomly selected pairs) to make sure that the task design and the aggregation method were appropriate. Each page (or HIT) contained 5 individual pair judgments and was worth $0.05. With 10 judgements being collected for each individual pair judgement, the cost of the pilot was around $3.

To ensure that the workers had the command of English necessary for completing the task, only workers located in the USA could access the job. Because the pilot data set was selected randomly from the set of pairs resolved by the game, it contained several ambiguous (difficult) cases. Although only the most intuitive pairs were chosen to create two gold units, their difficulty level was too high and hampered recruiting workers to perform the task. Finally, the task had to be canceled as it was too difficult: no player managed to correctly rate any of the associated gold units.

Task Settings and Execution. Based on the conclusions from the pilot, the main experiment was run over a set of 147 gold standard pairs. Since two experts agreed on a relation for these pairs, they were less ambiguous (and hopefully more amenable to be solved by micro-workers), than the randomly selected pairs of the pilot study. The experimental settings were the same as in the case of the pilot, leading to a cost per unit of $0.183 and a total cost of $26.98. Eight gold units were created from the easiest cases in order to avoid over-restricting the accessibility of the task. Nevertheless, the task and the underlying dataset proved rather difficult: the job was paused automatically by CF whenever the agreement of workers with the gold units was under a "normal" threshold. Even with the delays introduced by the job being paused (6 times), the job finished within approximately 24 hours, being active for about 8 hours overall (based on CF statistics).

Evaluation of Data Quality. The CrowdFlower results were compared to the gold standard in terms of overall agreement (i.e., precision) and specific agreements. For a relation type R and two annotators A and B, the specific agreement is $\frac{2*R_{A\&B}}{R_A+R_B}$, where $R_{A\&B}$ is the number of pairs for which both A and B agree that a relation R holds, while R_A and R_B are the number of term pairs judged as related through R by annotator A and B respectively. Table 2 sums up the results of the evaluation and compares them against those obtained with Climate Quiz.



TERM1: climate forcing ==== TERM2: climate

Type in the second letter of TERM1:

How many words make up TERM2?

4

Example: "bank"=1; "climate change"=2

What is the correct relation between TERM1 and TERM2?

Figure 2: CrowdFlower task interface.

In a first evaluation, the game specific judgment aggregation method was applied to the CrowdFlowe output (CF1 in Table 2). With 52 of the pairs unresolved (i.e., in these cases the most popular relation was less than 4 votes away from the second most popular one), the data collected with CF had a lower agreement level with the gold standard than the game results, namely 59%. CF1 lead to higher relation specific agreements for four relation types, most notably for subsumption where agreement was over 90%. From this comparison, it appears that, overall, game results are superior to those obtainable with mechanised labour.

To understand the influence of the aggregation methods, the results were also evaluated when using the aggregation method of the CF platform (CF2 in Table 2). This method computes a confidence value for each relation as a ratio between the sum of worker trust for all workers who selected that relation over the sum of trust levels of all workers who provided a relation between a pair. In turn, worker trust is computed based on the worker's performance in answering the gold units (e.g., a worker that answered correctly all gold units will have a trust value of 1). The relation with the highest confidence value is selected as the final result, while the confidence value provides an indication of the trustworthiness of that relation. For example, in the case of the pair $(oil, fossil\ fuel)$, 8 workers selected $is\ a\ sub-category\ of$ (sum of trust=7.25), one worker (trust=0.875) voted for *is identical to*, whereas one worker (trust=0.75) chose $works\ on/with$. The confidence values for the three individual relations were 0.816, 0.098 and 0.084 respectively, and therefore $is\ a\ sub-category\ of$ was selected as the final relation for this pair.

Contrary to our previous observation, these results show that, with a more sophisticated aggregation method, microworkers outperform players for most relation types, as well as overall, reaching a total agreement of 75%, as opposed to 72%. In terms of relation types, micro-workers performed best at identifying subsumption relations, with a relation specific agreement level of 91%. The worst performance was

Relation	**CQ**	**CF1**	**CF2**
is identical to	40%	67%	67%
is the opposite of	0	0	0
is a subcategory of	84%	91%	91%
supports	59%	36%	87%
threatens	95%	77%	87%
opposes	55%	33%	67%
influences	0	57%	31%
works on/with	25%	33%	33%
is not related to	71%	73%	81%
other	11%	0	0
Total	**72%**	**59%**	**75%**

Table 2: Relation specific and total agreement values for Climate Quiz (CQ) and CrowdFlower, with game (CF1) and CrowdFlower (CF2) specific aggregation.

obtained for the *other* relation as they did not assign a single *other* relation, preferring to choose one of the given relations. For both evaluations, a manual inspection of the pairs with incorrectly assigned relations revealed that none of them could have been assigned one of the eight relations if its terms were inverted. Therefore, the lack of term switching, did not have an effect on the quality of the results.

5. GENRE COMPARISON

Table 3 sums up our observations when comparing the two HC genres and compares them to the results in [15].

Cost. When computing the costs involved in setting up the two projects, and in order to allow comparison to other similar cost-focused studies [1], the wage of a research scientist implementing the projects was assumed to be $54,000 per annum. This leads to a setup cost of $9,000 corresponding to the two months development time for the Climate Quiz and to $450 for setting up the CrowdFlower interface (including running the pilot, but excluding worker fees). Even though setting up the game is more expensive than designing a CF task (about 20 times in our case), the situation is different when considering the cost per solved unit. Indeed, in the case of the game players contribute their judgments for free, whereas micro-workers are paid $0.183 per unit solved.

Speed. Our observations confirm all previously stated hypotheses. Setting up a game is much more time consuming, than designing a CrowdFlower task (e.g., requiring 20 times more time). Additionally, thanks to the large worker pool of CF (as well as the motivation of workers to finish tasks quickly, in order to increase their earnings per hour), its throughput in terms of individual judgments per hour was also superior to that of the game (243 vs. 180, almost twice as fast). Another advantage of CF is that task completion time is more predictable, than that of Climate Quiz.

Quality. Result quality highly depends on the used aggregation method: applying the Climate Quiz aggregation method leads to better results for games. However, the best results in comparison to the gold standard are obtained with the CF specific aggregation, and in this case, they are higher than those obtainable with the GWAP. This is a significant result showing that complex tasks such as selecting a relation from 10 possible alternatives can be outsourced to micro-workers. Therefore, micro-workers are capable of performing tasks of the same complexity as game players. It has to be noted here, however, that although the task structure per se was not an issue, the ambiguity of the data caused major issues for micro-workers. Indeed, while they produced good results, this behavior has been observed only on low-ambiguity data (easier cases). Indeed our pilot, which relied on a random selection of data, was too difficult and had to be cancelled. Even on the low-ambiguity data set, the crowdsourcing process was stopped several times, since periodically the disagreement with the gold standard data fell under the expected threshold. This is an indication, that, even though the task was ultimately solved, it was difficult when compared to other typical micro-task types.

Result quality depends on the motivation of the contributors. The type of motivations of players and micro-workers is a major difference between the two HC genres. Players are intrinsically motivated by the fun-factor, whereas micro-workers primarily have an extrinsic motivation relying on financial reward. This difference has the following consequences. Firstly, recruiting and maintaining players is significantly more difficult than recruiting micro-workers and requires considerable effort in (continuously) advertising the game on various channels [12]. Secondly, given the pay-based motivation of micro-workers, they are more likely to cheat than players. We cannot provide conclusive evidence on this point because: 1) we do not perform cheating detection among players; and 2) cannot judge which percentage of the 18% of automatically rejected judgments by CF was due to cheating, as opposed to the high task difficulty. Finally, because the fun factor is the players' primary motivation, the interestingness of the game tasks plays a major role. Difficult tasks quickly frustrate players and lower the time they play the game or even cause them to stop playing the game altogether. Indeed, 43% of all player actions are skips, thus indicating a low tolerance to difficult (or uninteresting tasks). Micro-workers on the other hand solve all available tasks to secure payments.

In terms of data bias, Climate Quiz has collected contributions from a larger player base than CrowdFlower (648 vs. 83 contributors). These numbers are, however, poor indicators of bias as the two projects ran for different lengths of time and Climate Quiz has been fed more data than its mechanised labour version. To better understand the potential bias, the percentage of contributions by top players/workers was computed. The 10 best players contributed to 37% of all the gathered judgments in the game, whereas this percentage was much higher in the case of CrowdFlower where the top 10 workers provided 61% of all judgments. In fact, three workers have completed all 147 pairs. These results suggest that the bias in the game results is lower, than the one of the mechanised labour task.

Comparison to other studies. Thaler et al. [15] reproduce the instance classification task supported by their OntoPronto game through MTurk. Although our experimental setup is similar to [15] , the focus is on a different task (relation detection) and therefore complements Thaler et al.'s earlier findings as summarised in Table 3 and discussed next.

Since set-up costs are not report in [15], for comparison pur-

	Study Observations		**Thaler et al. [15]**	
Feature	**CrowdFlower**	**Climate Quiz**	**MTurk**	**OntoPronto**
Cost				
Set-up Price	$450	$9,000	est. $4,500	est. $22,500
Price per unit	$0.183	$0	$0.74	$0
Speed				
Set-up Time	2 days	2 months	1 month	5 months
Throughput (judgments/H)	243	180	-	-
Throughput predictability	within hours	completion difficult to estimate	-	-
Quality				
Precision	CF1= 59%	72%	99%	97%
	CF2= 75%	72%		
Maintaining motivation	no effort to recruit micro-workers	significant effort for recruiting players	easy primarily financial	difficult - relies on sophisticated game design
Task complexity	similar	similar	similar	similar
Importance of task interestingness	micro-workers solve all tasks	players skip many tasks	-	-
Worker diversity	83	648	16	270

Table 3: Comparison of mechanised labour and games based on observations from this study and [15].

poses, these costs were estimated using $54,000 salary per annum as in the case of our experiments. As expected, although they have spent more on implementing their systems, the mechanised labour solution is cheaper than the development of the game. The ratio of the costs is 1:5 as opposed to 1:20 in our case since Thaler et al.'s system automatically manages HITs as opposed to the manual solution adopted by us. The cost per correct answer, in both cases, is $0 for games and under $0.20 when relying on mechanised labour.

Both studies confirmed that setting up a mechanised labour experiment is faster than designing and building a game. Similarly, the completion time of the actual experiments is about twice as fast as when using games. Indeed, Thaler et al. needed about 4 weeks to complete the experiment using the OntoProto and 2 weeks when using AMT.

Similarly to us, Thaler et al. report obtaining high quality data with both approaches, with the mechanised labour results being slightly better in terms of precision. Note however, that there is a significant difference in the level of obtainable precision between the work of Thaler and our own, which results from the different difficulty level of the two tasks. It can be therefore confirmed, that, even across tasks of different difficulty, both approaches are likely to provide similar quality output. As in the experiment reported here, the work of Thaler et al. shows that workers and players can solve tasks of similar complexity, that maintaining player motivation is crucial in games while rather straightforward on mechanised labour platforms, and that the lower number workers than players is likely to induce bias in mechanised labour results.

While this study confirms all findings of [15] and generalises those over tasks of different difficulty, it invalidates some of the assumptions made by earlier papers (Table 1), namely that (i) games outperform mechanised labour in terms of data quality [1, 18]; (ii) players can solve more complex tasks than workers [1]; (iii) worker diversity is higher in mechanised labour platforms than in games [18].

Mason and Watts investigated the impact of financial incentives on the performance of the crowds, concluding that paying more will not increase the quality of the data although it might shorten the acquisition time [7]. While they focused on mechanised labour platforms alone, their finding has been confirmed by our work in a cross-genre setting: indeed, free game-based results have similar quality as those that are payed for, however, completion times are always slower.

6. CONCLUSIONS

This paper examined the complementarity of the two most common HC genres, mechanised labour and GWAPs. By combining observations from previous comparative studies, it was concluded that such complementarity did exist, along the key dimensions of cost, speed and quality. The direct comparison of the two genres, using the task of relation selection between term pairs, confirmed most of the complementarities observed by previous studies, rejected three of the earlier assumptions and provided concrete quantitative data about the level of genres differences.

To leverage the complementarity of the two genres, a *cross-genre HC platform* is being designed, as part of the uComp project(`www.ucomp.eu`). The platform will utilise HC to acquire collective intelligence, and provide methodological support for defining and deploying HC tasks across genres, as well as embedding them into complex knowledge creation workflows. The novel elements come from: *(i)* building a novel, customisable and reusable HC Framework for knowledge extraction and verification; *(ii)* providing HC task deployment on social and mobile platforms, as well as across HC genres; and *(iii)* offering advanced methods for human contributor engagement and retention.

Even though this work is a step forward, future work will focus on more comparative studies, on tasks of different types and difficulty levels, in order to gain a more complete understanding of how the two genres complement each other. In particular, the task investigated here has a high difficulty, so future studies should be performed on lower difficulty tasks as well. In particular, followup studies should clarify any potential influence of the current study design. Firstly, the slight differences between the two user interfaces might have affected worker performance. While the uComp platform will allow experimentation with similar interfaces, interface quality is a differentiating feature of the two genres and it is unreasonable to assume that the same high-quality interfaces will be built for mechanised labour projects, as for games. Secondly, our comparison does not account for the fact that CF systematically removes poor quality judgements from the final results based on its training mechanism. Finally, access to the CF tasks was limited to US workers only while anyone could access Climate Quiz. Based on their Facebook profile, 55% of the players set English as their main language, so overall, the game population might have a lower level of English skills than CF workers. The uComp platform's more detailed, skill-based player/worker profiling will facilitate comparisons between similar user groups.

Future work will also focus on implementing the uComp platform. This entails, among others, work on: (1) extending our existing aggregation method and adapting it for use within the platform, as opposed to just within Climate Quiz; (2) investigating whether similar quality results can be obtained with fewer than 10 workers, (3) exploring the use of CF confidence values for improving result precisions; and (4) creating workflows that outsource tasks to the appropriate HC genre based on their characteristics (e.g., complexity).

ACKNOWLEDGMENTS

The Climate Quiz Facebook application was developed within the DIVINE (`www.weblyzard.com/divine/`) and the Triple-C (`www.modul.ac.at/nmt/triple-c`) research projects funded by FIT-IT Semantic Systems of the Austrian Research Promotion Agency and the Austrian Climate Research Program of the Austrian Climate and Energy Fund. The CrowdFlower-based methods were developed and evaluated as part of uComp (`www.ucomp.eu`), a European research effort in the CHIST-ERA program line (EP/K017896/1) and also with partial support from the UK Engineering and Physical Sciences Research Council (grant EP/I004327/1).

7. REFERENCES

[1] J. Chamberlain, K. Fort, U. Kruschwitz, M. Lafourcade, and M. Poesio. Using Games to Create Language Resources: Successes and Limitations of the Approach. In I. Gurevych and K. Jungi, editors, *The People's Web Meets NLP. Collaboratively Constructed Language Resources*. Springer, 2013. To Appear.

[2] K. Eckert, M. Niepert, C. Niemann, C. Buckner, C. Allen, and H. Stuckenschmidt. Crowdsourcing the Assembly of Concept Hierarchies. In *Proc. of the 10th Annual Joint Conference on Digital Libraries*, JCDL '10, pages 139–148. ACM, 2010.

[3] K. Fort, G. Adda, and K. Cohen. Amazon Mechanical Turk: Gold Mine or Coal Mine? *Computational Linguistics*, 37(2):413 –420, 2011.

[4] A. Kawrykow, G. Roumanis, A. Kam, D. Kwak, C. Leung, C. Wu, E. Zarour, and P. players. Phylo: A Citizen Science Approach for Improving Multiple Sequence Alignment. *PLoS ONE*, 7(3):e31362, 2012.

[5] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing User Studies with Mechanical Turk. In *Proc. of the 26th Conference on Human Factors in Computing Systems*, pages 453–456, 2008.

[6] F. Laws, C. Scheible, and H. Schütze. Active Learning with Amazon Mechanical Turk. In *Proc. of the Conf. on Empirical Methods in NLP*, pages 1546–1556, 2011.

[7] W. Mason and D. J. Watts. Financial Incentives and the "Performance of Crowds". In *Proc. of the ACM SIGKDD Workshop on Human Computation*, HCOMP '09, pages 77–85. ACM, 2009.

[8] G. Parent and M. Eskenazi. Speaking to the Crowd: Looking at Past Achievements in Using Crowdsourcing for Speech and Predicting Future Challenges. In *Proc. of INTERSPEECH*, pages 3037–3040, 2011.

[9] M. Poesio, U. Kruschwitz, J. Chamberlain, L. Robaldo, and L. Ducceschi. Phrase Detectives: Utilizing Collective Intelligence for Internet-Scale Language Resource Creation. *Transactions on Interactive Intelligent Systems*, 2012. To Appear.

[10] A. J. Quinn and B. B. Bederson. Human Computation: A Survey and Taxonomy of a Growing Field. In *Proc. of Human Factors in Computing Systems*, pages 1403–1412, 2011.

[11] M. Sabou, K. Bontcheva, and A. Scharl. Crowdsourcing Research Opportunities: Lessons from Natural Language Processing. In *Proc. of the 12th International Conference on Knowledge Management and Knowledge Technologies (iKNOW), Special Track on Research 2.0*, 2012.

[12] A. Scharl, M. Sabou, and M. Föls. Climate Quiz: a Web Application for Eliciting and Validating Knowledge from Social Networks. In *Proc. of the 18th Brazilian symposium on Multimedia and the web*, WebMedia '12, pages 189–192. ACM, 2012.

[13] K. Siorpaes and M. Hepp. Games with a Purpose for the Semantic Web. *Intelligent Systems, IEEE*, 23(3):50 –60, 2008.

[14] 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 *Proc. of EMNLP*, pages 254–263, 2008.

[15] S. Thaler, E. Simperl, and S. Wölger. An Experiment in Comparing Human-Computation Techniques. *IEEE Internet Computing*, 16(5):52–58, 2012.

[16] L. von Ahn. *Human Computation*. PhD thesis, Carnegie Mellon University, December 2005.

[17] L. von Ahn. Games With a Purpose. *Computer*, 39(6):92 –94, 2006.

[18] A. Wang, C. Hoang, and M. Y. Kan. Perspectives on Crowdsourcing Annotations for Natural Language Processing. *Language Resources and Evaluation*, 47(1), 2013.

[19] G. Wohlgenannt, A. Weichselbraun, A. Scharl, and M. Sabou. Dynamic Integration of Multiple Evidence Sources for Ontology Learning. *Journal of Information and Data Management*, 3(3):243–254, 2012.

[20] L. Wolf, M. Knuth, J. Osterhoff, and H. Sack. RISQ! Renowned Individuals Semantic Quiz - a Jeopardy like Quiz Game for Ranking Facts. In *Proc. of the 7th International Conference on Semantic Systems*, I-Semantics '11, pages 71–78. ACM, 2011.