Ethical Issues in Corpus Linguistics And Annotation: Pay Per Hit Does Not Affect Effective Hourly Rate For Linguistic Resource Development On Amazon Mechanical Turk

K Bretonnel Cohen, Karën Fort, Gilles Adda, Sophia Zhou, Dimeji Farri

Ethical Issues in Corpus Linguistics And Annotation:
Pay Per Hit Does Not Affect Effective Hourly Rate For Linguistic Resource Development On Amazon Mechanical Turk

K. Bretonnel Cohen[1], Karën Fort[2], Gilles Adda[3], Sophia Zhou[4], and Dimeji Farri[4]
[1] Biomedical Text Mining Group, Computational Bioscience Program, U. Colorado School of Medicine
[2] Equipe Sens Texte Informatique Histoire, Université Paris-Sorbonne
[3] Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur
[4] Philips Research North America

Abstract

Ethical issues reported with paid crowdsourcing include unfairly low wages. It is assumed that such issues are under the control of the task requester. Can one control the amount that a worker earns by controlling the amount that one pays? 412 linguistic data development tasks were submitted to Amazon Mechanical Turk. The pay per HIT was manipulated through a range of values. We examined the relationship between the pay that is offered per HIT and the effective pay rate. There is no such relationship. Paying more per HIT does not cause workers to earn more: the higher the pay per HIT, the more time workers spend on them ($R = 0.92$). So, the effective hourly rate stays roughly the same. The finding has clear implications for language resource builders who want to behave ethically: other means must be found in order to compensate workers fairly. *The findings of this paper should not be taken as an endorsement of unfairly low pay rates for crowdsourcing workers. Rather, the intention is to point out that additional measures, such as pre-calculating and communicating to the workers an average hourly, rather than per-task, rate must be found in order to ensure an ethical rate of pay.*

Keywords: ethics, corpus linguistics, corpus annotation, Amazon Mechanical Turk, crowdsourcing

1. Introduction

Crowdsourcing has become a popular way to create data for research, in particular in natural language processing (NLP). There are a variety of approaches to crowdsourcing and many crowdsourcing taxonomies, many of which are presented in (Geiger et al., 2011). One way to distinguish between these many approaches is to consider (i) the remuneration of the participants and (ii) the transparency of the task (that is, whether or not it is obvious to the participants). This small set of features allows one to distinguish between three major types of crowdsourcing: (i) volunteer and transparent, as in the case of vested volunteers who have a personal commitment to the intended use of the data (Cohen et al., 2015); (ii) volunteer and not transparent, as in the case of games with a purpose (GWAPs), which offer an entertaining experience to the participants; and (iii) remunerated and transparent crowdsourcing, i.e. microworking. The latter is typically done via dedicated platforms such as Amazon Mechanical Turk or CrowdFlower, and raises a number of ethical issues. Some of these have been addressed in various publications, including (Fort et al., 2011).

Analysts have identified a number of ethical issues with paid crowdsourcing (Adda et al., 2013). Unfairly low wages (Ross et al., 2009; Chilton et al., 2010) are one such problem. As a significant proportion of the workers use MTurk as their primary source of income, or to make basic ends meet (Ross et al., 2010; Ipeirotis, 2010; Fort et al., 2011), this becomes an ethical issue. Those very low wages are partly induced by the pay per task model, because the worker is not aware of the hourly rate before choosing the task (Callison-Burch, 2014). Another frequently mentioned problem (Silberman et al., 2010) is the fact that requesters sometimes pay late, or even not at all.

It is widely assumed that these issues are under the control of the purchaser of crowdsourcing services. The work reported here investigates a number of assumptions about these issues and about the extent of purchaser control over them. In particular, we wondered: suppose that a purchaser of annotation services through a crowdsourcing site wants to ensure that they pay an ethical wage. Can one control the amount that a worker earns by controlling the amount that one pays? It seems obvious that one should be able to, but early experiences suggested that this might not, in fact, be the case.

The methodology was as follows. In the course of our normal work on preparing linguistic resources for use in developing and testing natural language processing applications, a variety of types of tasks were submitted to Amazon Mechanical Turk. The pay per HIT (Human Intelligence Task, the basic unit of work performance on the Amazon Mechanical Turk web site) was manipulated through a range of values (never below the typical payment for a task type). The total data set contains 412 data points.

The Amazon Mechanical Turk interface provides a number of data points upon completion of a task. These include:

- Pay per HIT: this is the amount offered per HIT by the person who "requests" (in Amazon Mechanical Turk parlance) that the work be done.
- Average time per assignment: this is the average amount of time spent by a worker on a HIT.
- Effective hourly rate: this is the extrapolated amount earned per hour by a typical worker for doing the task.
- Agreement: for classification tasks, this is the agreement between workers.
- Total number of HITs completed: this is the number of HITs done at the indicated pay per HIT, effective

hourly rate, etc.

Reasonable expectations are that all other things being equal:

- Effective hourly rate should correlate positively with pay per hit.
- Average time per assignment should correlate positively with pay per hit.
- Effective hourly rate should correlate negatively with average time per assignment.
- Agreement should correlate positively with pay per hit.
- Agreement should correlate negatively with effective hourly rate.

The reader may disagree with the authors' expectations about these relationships, but the data presented here allows the testing of discordant expectations, as well. That is, whether the reader agrees with the author that effective hourly rate should correlate positively with pay per HIT, or thinks that it should correlate negatively, or doesn't think that there should be any correlation at all, the data allows testing any of those hypotheses.

1.1. Tasks

Data from a variety of task types is analyzed here. Tasks were not created specifically for this paper—these were tasks that we carried out in the course of our normal research work. The task types discussed here are:

- Information extraction: relation annotation (1 set of tasks)
- Recognizing textual entailment: language generation (3 separate sets of tasks)
- Recognizing textual entailment: classification (1 set of tasks)
- Paraphrase relations: classification (3 separate sets of tasks)

The number of workers participating in a task is variable from one set of tasks to another, as is the number of subtasks (e.g. classifying a single pair of sentences versus writing three separate sentences) and the total number of completed HITs per task.

2. Results

The data consists of 412 completed HITs. There was no attempt to balance across the various pay levels or task types—the tasks were requested in the course of the authors' normal work. Table 1 gives descriptive statistics on the number of HITs completed at each level of pay per HIT. Table 2 gives the number of HITs completed for each task type.

Figure 1 shows the effective hourly rate for the various tasks as a function of the pay per HIT. It is clear from the figure that there is no relationship between the pay that is offered and the amount that is earned. Since there is no linear relationship, we do not calculate a correlation. The data show that we cannot cause workers to earn higher wages by paying more per HIT. Regardless of whether we pay $0.05 per HIT or five times that much, the effective hourly rate hovers around the median of US $2.25. It is worth noting that the one set of tasks that shows an apparent effective hourly rate of US $12.50 per hour had only 12 completed tasks, and therefore the sample size is much smaller than for the other sets of tasks.

Table 1: Number of HITs completed at each level of pay per HIT. The 8 sets of tasks comprised 412 individual hits.

Pay per HIT	number of HITs completed
US $ 0.05	110
US $0.10	173
US $0.25	129
Total HITs completed	412

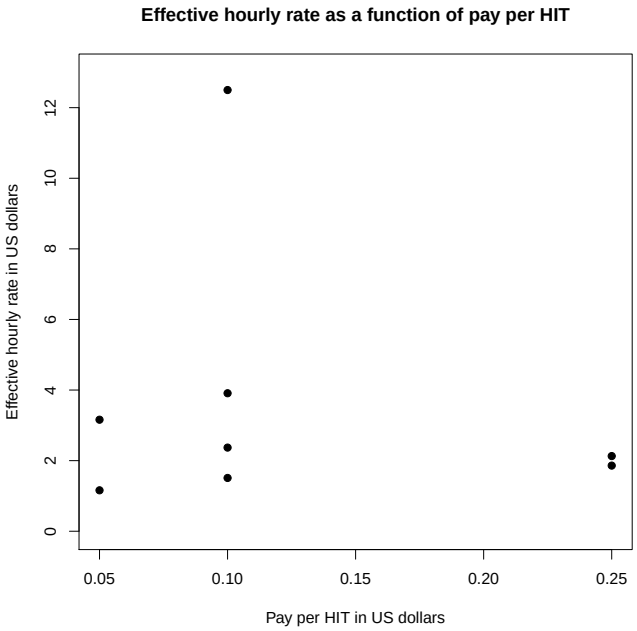

Figure 1: There is no relationship between the pay per HIT and the effective hourly rate earned by workers. Each data point on the graph represents a separate set of tasks. 412 HITs were completed in the course of these 8 sets of tasks.

Figure 2 shows the agreement for the various classification tasks as a function of the pay per HIT. Examining the inter-rater agreement on the five classification tasks as a function of pay per HIT, it does not appear that there is a relationship between the pay that is offered and the agreement that is achieved. Since there is no linear relationship, we do not calculate a correlation. The data show that we cannot get better agreement by paying more per HIT. The agreement that is achieved at a pay per HIT of $0.25 is not necessarily

Table 2: Number of HITs completed for each task type. The 8 sets of tasks comprised 412 individual hits.

Task	number of HITs completed
Information extraction (relation annotation)	56 ($0.10 per HIT)
RTE (language generation)	54 ($0.10 per HIT)
RTE (classification)	86 (25 x $0.25, 61 x $0.10 per HIT)
Paraphrase relations (classification)	270 (56 x $0.05, 56 x $0.10, 104 x $0.25 per HIT)
Total HITs completed	412

any higher than the agreement that is achieved at a pay per HIT of $0.05.

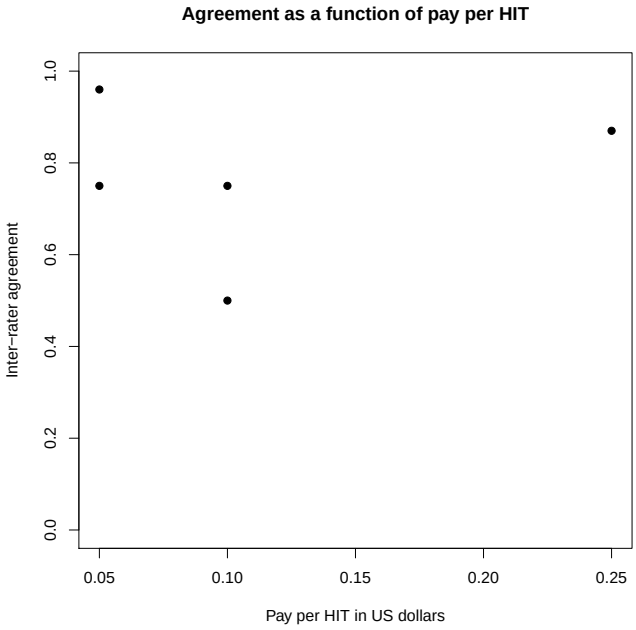

Figure 2: There is no relationship between the pay per HIT and the agreement achieved on a classification task. Each data point on the graph represents a separate set of tasks. 326 HITs were completed in the course of these 5 sets of tasks.

We cannot achieve higher agreement by paying more. Other than these two findings, the expectations listed in the Introduction were supported.

2.1. Why doesn't effective hourly rate increase as a function of pay per HIT?

Examining the time spent per HIT as a function of pay per HIT, we see why the effective hourly rate does not go up as the pay per HIT increases. Figure 3 shows the average time per task as a function of the pay per HIT. There *is* a linear relationship between the pay per HIT and the average time per assignment: as the pay per HIT goes up, the average time per assignment goes up. That is, the more the workers are paid, the more time they spend on each individual HIT. The correlation between them is very strong, at R = 0.92. Thus, even though the pay per HIT increases, the effective hourly rate stays about the same.

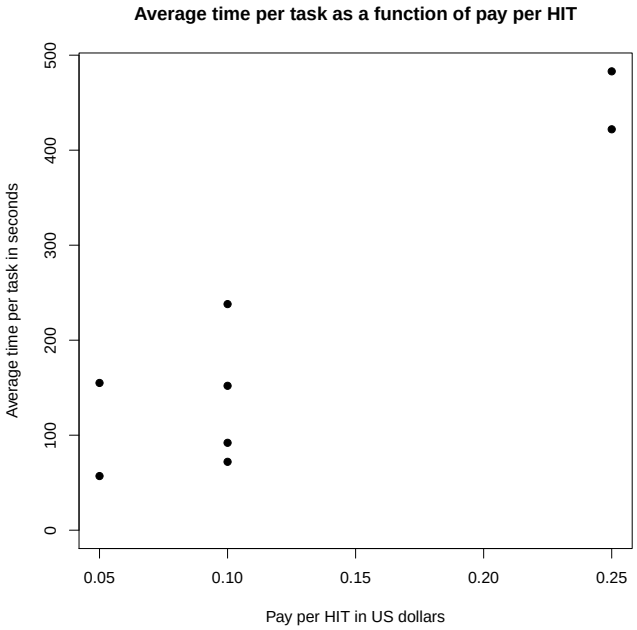

Figure 3: There is a linear relationship between the pay per HIT and the average time per task, R = 0.92. Each data point on the graph represents a separate set of tasks. 412 HITs were completed in the course of these 8 sets of tasks.

3. Discussion and Conclusions

3.1. Discussion

Although to the best of the authors' knowledge, the specific issue examined in this paper has not been studied before, there is a considerable amount of relevant work on the subject of crowdsourcing methods in general and crowdsourcing for linguistic resource creation in particular. (Callison-Burch and Dredze, 2010) give an overview of the results of a workshop on the use of Amazon Mechanical Turk to create data sets for natural language processing, held under the auspices of the Association for Computation Linguistics. The paper describes the results of 24 attempts to create language resources with Amazon Mechanical Turk, and gives some recommended practices for using the platform, including trying the task yourself and then having someone outside of the field try it, in order to assess the "doability" of the task and to estimate the time per HIT, in order to allow you to offer fair remuneration. (Sabou et al., 2012) point out some of the salutary effects of crowdsourcing linguistic resource construction, including diversification of

the task types, languages, resource types, and linguistic phenomena. In counterpoint, (Sagot et al., 2011) present a wide-ranging critique of for-pay crowdsourcing for language resource development in general, including observations consistent with the idea that crowdsourcing might not be as inexpensive as is widely assumed when one takes into account the costs of developing the interface, validating the data, and post-Turking processing; and the impossibility of determining with certainty the native language of Turkers. (Snow et al., 2008) measured the agreement between Turkers and expert annotators for five tasks, including recognizing textual entailment (the task type for 140 of the 412 HITs that were the source of the data in this paper). They found high agreement rates for all five task types. (Adda et al., Undated) also give a list of best practices, many of which deal with the ethical issues involved in crowdsourcing. These include taking into account the amount of time necessary to accomplish the task, including an estimated *hourly* wage in the work request (in addition to the pay per task that is automatically included by Amazon), defining in advance objective measures for deciding when work will be rejected (that is, not reimbursed) and making those measures known to potential Turkers, giving immediate feedback, and not requesting tasks anonymously.

The fact that ethical issues exist concerning the use of for-pay crowdsourcing comes up repeatedly in these papers. It is typical for those papers that recommend best practices for crowdsourcing recommend paying a fair rate. This does not seem like a controversial recommendation. However, the data presented here suggest that it might be more difficult to figure out how to do so than it appears at first glance—simply offering a higher pay rate per task does not result in a higher effective rate of pay.

3.2. Conclusions

We examined the relationship between the pay that is offered for each task on a crowdsourcing platform and the amount that a worker earns for performing that test. The data from eight sets of tasks comprising 412 HITs is consistent with the surprising finding that there is no relationship between them. Paying more per HIT does not cause workers to earn more per HIT: the higher the rate of pay, the more time workers spend on individual HITs. So, the effective hourly rate stays roughly the same: workers do not earn more regardless of how much we pay per HIT. This finding is consistent across a variety of NLP application data types (information extraction, recognizing textual entailment, and paraphrasing) and resource-building task types (classification and language generation). The finding has serious implications for language resource builders who want to behave ethically in their treatment of workers: other means besides higher pay per HIT must be found in order to compensate workers fairly. *The findings of this paper should not be taken as an endorsement of unfairly low pay rates for crowdsourcing workers. Rather, the intention is to point out that additional measures, such as pre-calculating and communicating to the workers an average hourly, rather than per-task, rate must be found in order to ensure an ethical rate of pay.*

4. Bibliographical References

Adda, G., Mariani, J.-J., Besacier, L., and Gelas, H. (2013). Economic and ethical background of crowdsourcing for speech. In *Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment*, pages 303–334. Wiley.

Adda, G., Mariani, J. J., and Besacier, L. (Undated). Analyse économique, juridique et éthique du crowdsourcing pour le TAL.

Callison-Burch, C. and Dredze, M. (2010). Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12. Association for Computational Linguistics.

Callison-Burch, C. (2014). Crowd-workers: Aggregating information across Turkers to help them find higher paying work. In *The Second AAAI Conference on Human Computation and Crowdsourcing (HCOMP-2014)*, November.

Chilton, L. B., Horton, J. J., Miller, R. C., and Azenkot, S. (2010). Task search in a human computation market. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 1–9.

Cohen, K. B., Pestian, J., and Fort, K. (2015). Annotateurs volontaires investis et éthique de l'annotation de lettres de suicidés. In *ETeRNAL (Ethique et Traitement Automatique des Langues)*.

Fort, K., Adda, G., and Cohen, K. B. (2011). Amazon Mechanical Turk: gold mine or coal mine? *Computational Linguistics (editorial)*, 37(2).

Geiger, D., Seedorf, S., Schulze, T., Nickerson, R. C., and Schader, M. (2011). Managing the crowd: Towards a taxonomy of crowdsourcing processes. In *AMCIS 2011 Proceedings*.

Ipeirotis, P. (2010). Analyzing the Amazon Mechanical Turk marketplace. CeDER Working Papers, http://hdl.handle.net/2451/29801, September. CeDER-10-04.

Ross, J., Zaldivar, A., Irani, L., and Tomlinson, B. (2009). Who are the Turkers? worker demographics in Amazon Mechanical Turk. Social Code Report 2009-01.

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., and Tomlinson, B. (2010). Who are the crowdworkers?: shifting demographics in Mechanical Turk. In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*, CHI EA '10, New York, NY, USA. ACM.

Sabou, M., Bontcheva, K., and Scharl, A. (2012). Crowdsourcing research opportunities: lessons from natural language processing. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*, page 17. ACM.

Sagot, B., Fort, K., Adda, G., Mariani, J., and Lang, B. (2011). Un turc mécanique pour les ressources linguistiques: critique de la myriadisation du travail parcellisé. In *TALN'2011-Traitement Automatique des Langues Naturelles*.

Silberman, M. S., Ross, J., Irani, L., and Tomlinson, B.

(2010). Sellers' problems in human computation markets. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 18–21.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics.