Amazon Mechanical Turk: Gold Mine or Coal Mine?

Karën Fort, Gilles Adda, Kevin Bretonnel Cohen

Last Words

Amazon Mechanical Turk: Gold Mine or Coal Mine?

Karën Fort*
INIST-CNRS/LIPN

Gilles Adda**
LIMSI/CNRS

K. Bretonnel Cohen†
U. Colorado School of Medicine/U. Colorado at Boulder

Recently heard at a tutorial in our field: "It cost me less than one hundred bucks to annotate this using Amazon Mechanical Turk!" Assertions like this are increasingly common, but we believe they should not be stated so proudly; they ignore the ethical consequences of using MTurk (Amazon Mechanical Turk) as a source of labour.

Manually annotating corpora or manually developing any other linguistic resource, such as a set of judgments about system outputs, represents such a high cost that many researchers are looking for alternative solutions to the standard approach. MTurk is becoming a popular one. However, as in any scientific endeavor involving humans, there is an unspoken ethical dimension involved in resource construction and system evaluation, and this is especially true of MTurk.

We would like here to raise some questions about the use of MTurk. To do so, we will define precisely what MTurk is and what it is not, highlighting the issues raised by the system. We hope that this will point out opportunities for our community to deliberately value ethics above cost savings.

What is MTurk? What is it not?

MTurk is an on-line crowdsourcing, microworking[1] system which enables elementary tasks to be performed by a huge number of on-line people. Ideally, these tasks are meant to be solved by computers, but they still remain out of computational reach (for instance, the translation of an English sentence into Urdu). MTurk is composed of two populations: the Requesters, who launch the tasks to be completed, and the Turkers, who complete these tasks. Requesters create the so-called "HITs" (Human Intelligence Tasks), which are elementary components of complex tasks. The art of the requesters is to split complex tasks into basic steps and to fix a reward, usually very low (for instance

* INIST-CNRS/LIPN, 2 allée de Brabois, F-54500 Vandoeuvre-lès-Nancy, France. E-mail: karen.fort@inist.fr.
** LIMSI/CNRS, Rue John von Neumann, Université Paris-Sud F-91403 ORSAY, France. Email: gilles.adda@limsi.fr.
† Center for Computational Pharmacology, U. Colorado School of Medicine; Dept. of Linguistics, U. Colorado at Boulder. Email: kevin.cohen@gmail.com.

1 Microworking refers to the fact that tasks are cut into small pieces and their execution is paid for. Crowdsourcing refers to the fact that the job is outsourced on the web and done by many people (paid or not).

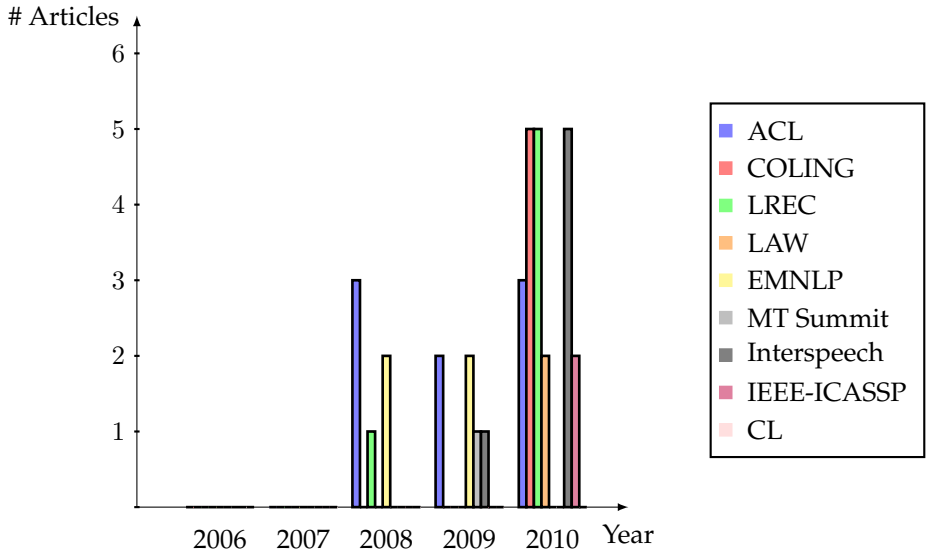

Figure 1
Evolution of MTurk usage in NLP publications

5 cents US to translate a sentence). Using the MTurk paradigm, language resources can be produced at a fraction (1/10th at least) of the usual cost (Callison-Burch and Dredze 2010).

MTurk should therefore not be considered as a game with a purpose, unlike Phrase Detectives (Chamberlain, Poesio, and Kruschwitz 2008) in which the gain is not emphasized (only the best contributors gain a prize) or, for French, JeuxDeMots ("Play On Words") (Lafourcade 2007), which does not offer any prize.

MTurk is not a game or a social network, it is an unregulated labor marketplace: a system which deliberately does not pay fair wages, does not pay due taxes, and provides no protections for workers.

Why are we concerned?

Since its introduction in 2005, there has been a steadily growing use of MTurk in building or validating NLP resources, and most of the main scientific conferences in our field include papers involving MTurk. Figure 1 was created by automatically searching the proceedings of some of the main speech and language processing conferences, as well as some smaller events specializing in linguistic resources, using the quoted phrase “Mechanical Turk“. We then manually checked the retrieved articles, source by source, to identify those which really make use of MTurk, ignoring those which simply talk about it. (For example, in the LREC 2010 proceedings, eight articles talk about MTurk, but only five used it, and in 2008, out of two papers citing MTurk, only one used it.) The present journal, *Computational Linguistics* (CL), appears in the bar chart with a zero count, as none of the articles published in it so far mention MTurk. All of the other sources contained at least one article per year using MTurk. The total number of publications varies from year to year, since e.g. conferences may accept different

numbers of papers each year, and some conferences, such as LREC, occur only every two years.

We performed another, less detailed, search, this time in the whole ACL Anthology (not source by source), using the same quoted phrase "Mechanical Turk" on November 5, 2010. We examined the hits manually, and out of the 124 resulting hits, 86 were papers in which the authors actually used MTurk as part of their research methodology. Interestingly, we noticed that at least one paper that we know to have used MTurk, namely (Biadsy, Hirschberg, and Filatova 2008), was not returned by the search. The published version of this paper does not explicitly mention MTurk, but the corresponding presentation at the conference indicated that MTurk was used. This is some evidence that use of MTurk may be under-reported. It should be noted that these results include a specialized workshop, the NAACL-HLT 2010 Workshop on Amazon Mechanical Turk (35 papers), the existence of which is, in itself, strong evidence of the importance of the use of MTurk in the domain.

A vast majority of papers present small to medium size experiments where the authors have been able to produce linguistic resources or perform evaluations at a very low cost; at least for transcription and translation, the quality is sufficient to train and evaluate statistical translation/transcription systems (Callison-Burch and Dredze 2010; Marge, Banerjee, and Rudnicky 2010). However, some of these papers bring to light language resource quality problems. For example, Tratz and Hovy (2010) note that the user interface limitations constitute "The first and most significant drawback" of MTurk, as, in their context of annotating noun compound relations using a large taxonomy, "it is impossible to force each Turker to label every data point without putting all the terms onto a single web page, which is highly impractical for a large taxonomy. Some Turkers may label every compound, but most do not." They also note that "while we requested that Turkers only work on our task if English was their first language, we had no method of enforcing this." Finally, they note that "Turker annotation quality varies considerably". Another important point is made in (Bhardwaj et al. 2010), where it is shown that, for their task of word sense disambiguation, a small number of trained annotators are superior to a larger number of untrained Turkers. On that point, their results contradict that of (Snow et al. 2008), whose task was much simpler (the number of senses per word was 3 for the latter, versus 9.5 for the former). The difficulty of having Turkers perform complex tasks also appears in (Gillick and Liu 2010), an article from the proceedings of the NAACL-HLT 2010 Workshop on Amazon Mechanical Turk, in which non-expert evaluation of summarization systems is proved to be "not able to recover system rankings derived from experts". Even more interestingly, Wais et al. (2010) show that standard machine learning techniques (in their case, a Naïve Bayes classifier) can outperfom the Turkers on a categorization task (classifying businesses into Automotive, Health, Real Estate, etc). Therefore, in some cases, NLP tools already do better than MTurk. Finally, as we said earlier, the vast majority of papers present only small or medium size experiments. This can be explained by the fact that, at least according to (Ipeirotis 2010a), submitting large jobs in MTurk results in low quality and unpredictable completion time.

Who are the Turkers?

Many people conceive of MTurk as a transposition of Grid Computing to humans, thus making it possible to benefit from humans' "spare cycles" to develop a virtual computer of unlimited power. The assumption is that there is no inconvenience for humans (as it is not a real work), and the power comes from the myriad. *This a fiction*.

Let us look first at how many Turkers are performing the HITs. This is a quite difficult task, because Amazon does not give access to many figures about them. We know that over 500k people are registered as Turkers in the MTurk system. But how many Turkers are really performing HITs? To evaluate this, we combined two different sources of information. First, we have access to some surveys about the demographics of the Turkers (Ipeirotis 2010b; Ross et al. 2009, 2010). These surveys may have a bias over the real population of Turkers, as some Turkers may be reluctant to respond to surveys. Because the results of these surveys are quite consistent, and the surveys are usually easy to complete, not particularly boring, and paid above the usual rate, we may assume that this bias is minor, and accept what they say as a good picture of the population of Turkers. In these surveys we see a lot of interesting things. For instance, there is a growing number of people from India: there were below 10% in 2008, above 33% in early 2010, and they represented about 50% of the Turkers in May[2]. Even if these surveys show that the populations from India and the US are quite different, we may take as an approximation that they have about the same reasons to perform HITs in Mturk, and produce about the same activity. We looked at how many HITs the 1,000 Turkers who completed the survey in (Ipeirotis 2010b) claim to perform each week: between 138,654 and 395,106 HITs per week.[3] The second source of information comes from the Mechanical Turk Tracker[4]: according to it, 700,000 HITs are performed each week. But the tracker system neither keeps track of the HITs which are completed in less than one hour, nor is able to quantify the fact that the same HIT can be completed by multiple workers and in fact should be, according to regular users like Callison-Burch and Dredze (2010). Asking the authors of (Ipeirotis 2010b), and the creator of the Mechanical Turk Tracker (who are in fact the same person), they (he) suggested that we should multiply the number given by the tracker by 1.7×5 to take into account these two factors[5], resulting in the (conjectural) total number of 5,950,000 HITs. Taking the two data points[6], we are able to hypothesize that the real number of Turkers is between 15,059 and 42,912. However, from the surveys, we have access to another figure: 80% of the HITs are performed by the 20% most active Turkers (Deneme 2009), who spend more than 15 hours per week in the MTurk system(Adda and Mariani 2010), which is consistent with the Pareto principle which says that 80% of the effects come from 20% of the causes. We may therefore say that 80% of the HITs are performed by 3,011 to 8,582 Turkers. These figures represent 0.6 to 1.7% of the registered Turkers, which in turn is in accord with with the "90-9-1" rule[7] valid in the Internet culture.

Another important question is if activity in MTurk should be considered as labor or something else (hobby, volunteer work,...). The observed mean hourly wages for performing jobs in the MTurk system is below $2 ($1.25 according to (Ross et al. 2009)). Because they accept such low rewards, a common assumption is that Turkers are US students or stay-at-home mothers who have plenty of leisure time and are happy to fill

2 http://blog.crowdflower.com/2010/05/amazon-mechanical-turk-survey/

3 The two figures come from the fact that each Turker gave a range of activity rather than an average number of HITs.

4 this system keeps tracks of all the HITs posted on Mturk, each hour `http://mturk-tracker.com`.

5 Personal communication in the comments of `http://behind-the-enemy-lines.blogspot.com/2010/03/new-demographics-of-mechanical-turk.html`, reporting that the tracker is missing ~ 70% of the posted HITs, which are posted and completed within less than one hour, and a 5x factor for the unobserved HIT redundancy.

6 1,000 Turkers perform between 138,654 and 395,106 HITs per week, and the total number of HITs in the Mturk system is about 5.95M HITs per week.

7 `http://en.wikipedia.org/wiki/1\%25_rule_(Internet_culture)`

their recreation time by making some extra money. According to recent studies in the social sciences (Ipeirotis 2010b; Ross et al. 2010), it is quite true that a majority (60%) of Turkers think that MTurk is a fruitful way to spend free time and getting some cash; but they are only 20% (5% of the India Turkers) who say that they use it to kill time. And these studies also show that 20% (30% of the India people) declare that they use MTurk "to make basic ends meet". From these answers, we find that money is an important motivation for a majority of the Turkers (20% use MTurk as their primary source of income, and 50% as their secondary source of income), and leisure is important for only a minority (30%). We cannot conclude from these studies that the activity in MTurk should be considered as labor for *all* the Turkers, but at least for the minority (20%) for whom MTurk represents a primary source of income.[8] Moreover, using the survey in (Ipeirotis 2010b), we find that this minority is performing more that 1/3rd of all the HITs.

What are the issues with MTurk?

The very low wages (below $2 an hour) are a first issue, but the use of Mechanical Turk raises other ethical issues as well. The position of many prototypical Turkers would be considered ethically unacceptable in major first-world countries. Denied even the basic workplace right of collective bargaining (unionization), this community has no recourse to any channels for redress of employer wrongdoing, let alone the normal ones available to any typical worker in the United States and many other developed nations (e.g. class action lawsuits, other lawsuits, and complaints to government agencies) while simultaneously being subjected to egregious vulnerabilities, including the fact that they have no guarantee of payment for work properly performed.

Legal issues surrounding the use of MTurk have also been encountered. At least one university legal department was sufficiently concerned that Turkers working for several months would claim employee status and demand health and other benefits that they refused to allow grant funds to be expended on MTurk. A small number of universities have insisted on institutional review board approval for MTurk experiments (Institutional review boards in US universities are independent bodies that review proposed experiments for legal and ethical issues).

Is MTurk the future of linguistic resource development?

The implicit belief that the very low cost of MTurk derives from the fact that incentivizing casual hobbyists requires only minimal payment is a mirage: once you admit that a majority of Turkers are not considering MTurk as a hobby, but as a primary or a secondary source of income, and that 1/3rd of the HITs are performed by Turkers who need MTurk to make basic ends meet, you then have to admit that MTurk is, at least for them, a labor marketplace. Moreover, the frequent assumption that the low rewards are a result of the classical law of supply-and-demand (large numbers of Turkers means more supply of labor and therefore lower acceptable salaries) is false. Firstly, we do not observe that there are too many Turkers. In fact, there are not enough Turkers. This can be observed through the difficulty to find Turkers with certain abilities (for instance understanding a specific language (Novotney and Callison-Burch 2010)), and in the difficulty to perform very large HIT groups (Ipeirotis 2010a). This is not surprising,

8 And even for the 50% who are utilizing MTurk as a secondary source of income.

as we have seen that the number of active Turkers is not that big. Secondly, the low cost is a result of the requesters' view of the relation between quality and reward: many articles (see for instance (Marge, Banerjee, and Rudnicky 2010)) relate that there is no correlation between the reward and the final quality. The reason is that increasing the price is believed to attract spammers (i.e. Turkers who cheat, not really performing the job, but using robots or answering randomly), and they are numerous in the MTurk system, because of an inadequate worker reputation system.[9] We obtain here a schema which is very close to what the 2001 economics Nobel prize winner George Akerlof calls "the market for lemons", where asymmetric information in a market results in "the bad driving out the good". He takes the market for used cars as an example (Akerlof 1970), where owners of good cars (here, good workers) will not place their cars on the used car market, because of the existence of many cars in bad shape (here, the spammers), which encourage the buyer (here, the requester) to offer a low price (here, the reward) because he does not know the exact value of the car. After some time, the good workers leave the market because they are not able to earn enough money given the work done (and sometimes they are not even paid), which in turn decreases the quality. At the moment, the system is stable in terms of the number of Turkers, because good workers are replaced by naive workers.

Amazon's attitude towards reputational issues has been passive. It maintains itself to be a neutral clearinghouse for labor, in which all else is the responsibility of the two consenting parties. This attitude has led to an explosion of micro-crowdsourcing start-ups, which observed the MTurk flaws and tried to overcome them.[10] Some of them could become serious alternatives to MTurk (TheQuill 2010), like Samasource[11], which offers at least a fair wage to workers, who in turn are clearly identified on the Web site, with their resume. But others are even worse than MTurk, ethically speaking. MTurk is ethically questionable enough; as a scientific community with ethical responsibilities we should seek to minimize the existence of even less ethical alternatives to it.

What's next?

If we persist in claiming that with MTurk we are now able to produce any linguistic resource or perform any manual evaluation of output at a very low cost, funding agencies will come to expect it. It is predictable that in assessing projects involving linguistic resource production or manual evaluation of output, funding agencies will prefer projects which propose to produce 10 or 100 times more data for the same amount of money. MTurk costs will then become the standard costs, and it will be very difficult to obtain funding for a project involving linguistic resource production at any level that would allow for more traditional, non-crowdsourced resource construction methodologies. Therefore, our community's use of MTurk not only supports a work-place model that is unfair and open to abuses of a variety of sorts, but also creates a de facto standard for the development of linguistic resources that may have long-term funding consequences.

9 For more details, see `http://behind-the-enemy-lines.blogspot.com/2010/10/be-top-mechanical-turk-worker-you-need.html`

10 ... for instance, Agent Anything, Clickworker, CloudCrowd, CrowdFlower, DoMyWork, JobBoy, LiveWork, Microtask, microWorkers, MiniFreelance, MiniJobz, MinuteWorkers, MyEasyTask, MyMicroJob, OpTask, RapidWorkers, Samasource, ShortTask, SimpleWorkers, SmartSheet, ...

11 `http://www.samasource.org`

Non-exploitative methods for decreasing the cost of linguistic resource development exist. They include semi-automatic processing, better methodologies and tools, games with a purpose, as well as microworking websites (like Samasource) that guarantee workers minimum payment levels. We encourage the computational linguistics and NLP communities to keep these alternatives in mind when planning experiments. If a microworking system is considered desirable by the ACL and ISCA communities, then we also suggest that they explore the creation and use of a linguistically specialized special-purpose microworking alternative to MTurk that both ensures linguistic quality and holds itself to the highest ethical standards of employer/employee relationships. Through our work as grant evaluators and recipients, we should also encourage funding bodies to require institutional review board approval for crowdsourced experiments and to insist on adherence to fair labor practices in such work.

Acknowledgments

We would like to thank Sophie Rosset, Joseph Mariani, Panos Ipeirotis, Eduard Hovy, and Robert Dale for their suggestions and encouragements. Any remaining errors are our own.

References

Adda, Gilles and Joseph Mariani. 2010. Language resources and amazon mechanical turk: legal, ethical and other issues. In *LISLR2010, "Legal Issues for Sharing Language Resources workshop", LREC2010*, Malta, 17 May.

Akerlof, George A. 1970. The market for 'lemons': Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84(3):488–500.

Bhardwaj, Vikas, Rebecca Passonneau, Ansaf Salleb-Aouissi, and Nancy Ide. 2010. Anveshan: A tool for analysis of multiple annotators' labeling behavior. In *Proceedings of The fourth linguistic annotation workshop (LAW IV)*, Uppsala, Sweden.

Biadsy, Fadi, Julia Hirschberg, and Elena Filatova. 2008. An unsupervised approach to biography production using Wikipedia. In *Proceedings of ACL 2008*, pages 807–815. Association for Computational Linguistics.

Callison-Burch, Chris and Mark Dredze. 2010. Creating speech and language data with amazon's mechanical turk. In *CSLDAMT '10: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Morristown, NJ, USA. Association for Computational Linguistics.

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

Deneme. 2009. How many turkers are there? http://groups.csail.mit.edu/uid/deneme/?p=502, December.

Gillick, Dan and Yang Liu. 2010. Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 148–151, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ipeirotis, Panos. 2010a. Analyzing the amazon mechanical turk marketplace. CeDER Working Papers, http://hdl.handle.net/2451/29801, September. CeDER-10-04.

Ipeirotis, Panos. 2010b. Demographics of mechanical turk. CeDER Working Papers, http://hdl.handle.net/2451/29585, March. CeDER-10-01.

Lafourcade, Mathieu. 2007. Making people play for lexical acquisition. In *Proc. SNLP 2007, 7th Symposium on Natural Language Processing*, Pattaya, Thailand, 13-15 December.

Marge, Matthew, Satanjeev Banerjee, and Alexander I. Rudnicky. 2010. Using the amazon mechanical turk for transcription of spoken language. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 5270–5273, Dallas, TX, 14-19 March.

Novotney, Scott and Chris Callison-Burch. 2010. Cheap, fast and good enough: automatic speech recognition with non-expert transcription. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 207–215, Morristown, NJ, USA. Association for Computational Linguistics.

Ross, Joel, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 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, pages 2863–2872, New York, NY, USA. ACM.

Ross, Joel, Andrew Zaldivar, Lilly Irani, and Bill Tomlinson. 2009. Who are the turkers? worker demographics in amazon mechanical turk. Social Code Report 2009-01, http://www.ics.uci.edu/ jwross/pubs/SocialCode-2009-01.pdf.

Snow, Rion, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP 2008*, pages 254–263.

TheQuill. 2010. Making money on-line, part 2: Microworking. http://thequill.org/personal-finance/9-making-money-on-line-part-2-microworking-.html, 21 June.

Tratz, Stephen and Eduard Hovy. 2010. A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 678–687, Uppsala, Sweden, July. Association for Computational Linguistics.

Wais, Paul, Shivaram Lingamneni, Duncan Cook, Jason Fennell, Benjamin Goldenberg, Daniel Lubarov, David Marin, and Hari Simons. 2010. Towards building a high-quality workforce with mechanical

turk. In *Proceedings of Computational Social Science and the Wisdom of Crowds (NIPS)*, December.