Crowdsourcing for Language Resource Development: Critical Analysis of Amazon Mechanical Turk Overpowering Use

Gilles Adda, Benoît Sagot, Karën Fort, Joseph Mariani

Crowdsourcing for Language Resource Development: Critical Analysis of Amazon Mechanical Turk Overpowering Use

Gilles Adda[1] *Benoît Sagot*[2] *Karën Fort*[3,4] *Joseph Mariani*[1,5]

[1]Spoken Language Processing group, LIMSI-CNRS, Orsay, France [5]IMMI-CNRS, Orsay, France
[2]Alpage, INRIA Paris–Rocquencourt & Université Paris 7, Rocquencourt, France
[3]INIST-CNRS, Vandoeuvre-lès-Nancy, France [4]LIPN, Université Paris Nord, Villetaneuse, France
benoit.sagot@inria.fr, karen.fort@inist.fr, {gilles.adda,joseph.mariani}@limsi.fr

Abstract

This article is a position paper about crowdsourced microworking systems and especially Amazon Mechanical Turk, the use of which has been steadily growing in language processing in the past few years. According to the mainstream opinion expressed in the articles of the domain, this type of on-line working platforms allows to develop very quickly all sorts of quality language resources, for a very low price, by people doing that as a hobby or wanting some extra cash. We shall demonstrate here that the situation is far from being that ideal, be it from the point of view of quality, price, workers' status or ethics and bring back to mind already existing or proposed alternatives. Our goal here is threefold: 1 - to inform researchers, so that they can make their own choices with all the elements of the reflection in mind, 2- to ask for help from funding agencies and scientific associations, and develop alternatives, 3- to propose practical and organizational solutions in order to improve new language resources development, while limiting the risks of ethical and legal issues without letting go price or quality.

keywords: Amazon Mechanical Turk, language resources

1. Introduction

Developing annotated corpora, as well as other language resources, involves such high costs that many researchers are looking for alternative, cost-reducing solutions. Among others, crowdsourcing, microworking[1] systems which enable elementary tasks to be performed by a huge number of on-line people, are possible alternatives. Nowadays, Amazon Mechanical Turk (MTurk) is the most popular of these systems, especially in the Speech & Language community. Since its introduction in 2005, there has been a steady growth of MTurk use in building or validating language resources (Fort et al., 2011). Costs are drastically reduced due to available sparse time of human language experts on-line. But MTurk raises, among others, ethical and quality issues which have been minimized until now, and we will investigate them in this paper. However, because we are aware that the development costs of corpora often stand in the way of language research and technologies, especially for Less-Resourced Languages (LRL), we are also sensible of some visible advantages of the crowdsourcing. Developing a crowdsourcing system which retains some of the main quality of MTurk (rapidity, diversity, access to non-expert judgment) and gets rid of the ethical and labor laws issues is (theoretically) possible, but this solution will require some delay (in the best case scenario) and the help of our scientific associations (ISCA, ACL, ELRA) and of the national and international funding agencies. Therefore, we will propose existing alternatives aiming at producing high quality resources at a reduced cost, while deliberately keeping ethics above cost savings.

[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 via the web and done by many people (paid or not).

2. MTurk: legends and truth

2.1. MTurk, a hobby for the Turkers?

In order to evaluate the ethics of MTurk, we need to qualify the activity of Turkers while they are participating in MTurk. Is it a voluntary work, as the one in Wikipedia? Looking at the MTurk site or at Turker blogs, where the monetary retribution is a major issue, the answer is clearly no. Maybe the activity could be described as a hobby, where the monetary retribution can be considered as a bonus, as some articles suggested it? Studies in social sciences (Ross et al., 2010; Ipeirotis, 2010), using surveys submitted within MTurk, give us some insight[2] into Turkers' socio-economic facts (country, age, ...) or the way they use MTurk (number of tasks per week, total income in MTurk, ...), and how they qualify their activity. 91% of the Turkers mentioned their desire to make money (Silberman et al., 2010), even if the observed wage is very low;[3] if 60% of the Turkers think that MTurk is a fairly profitable way of spending free time and getting some cash, only 30% mentioned their interest for the tasks, and 20% (5% of the Indian Turkers) say that they are using MTurk to kill time. Finally, 20% (30% of the Indian Turkers) declare that they are using MTurk to make basic ends meet, and about the same proportion that MTurk is their primary source of income.

Looking at the tasks which are performed within MTurk is another way to qualify the Turkers' activity. Innovative kinds of tasks can be found which can be seen as creative hobby activities. However, many tasks correspond to activities which used to be performed by salaried employees, and therefore are working activities; for these tasks, MTurk could be assimilated to offshoring on the Web to decrease production costs.

[2]For instance, we learn that Indian Turkers were 5% in 2008, 36% in December 2009 (Ross et al., 2010), 50% in May 2010 (http://blog.crowdflower.com/2010/05/amazon-mechanical-turk-survey/) and have produced over 60% of the activity in MTurk (Biewald, 2010).

[3]$1.25/hr according to (Ross et al., 2009) $1.38/hr according to (Chilton et al., 2010)

For years, speech corpora transcription (and translation) tasks were being performed by employees of agencies like LDC or ELDA: these are jobs. The 20% of the most active Turkers who spend more than 15 hours per week in MTurk (Adda and Mariani, 2010), and produce 80% of the activity, can be called laborers when performing these tasks.

It is difficult to be conclusive about the nature of the Turkers' activity. Many different types of tasks are proposed within MTurk and the Turkers' motivations are heterogeneous. Nevertheless, those 20% of the Turkers for whom MTurk is a primary income, and those Turkers who perform tasks which are actually performed by employees, produce an activity in MTurk corresponding to a real labor.

Qualifying the MTurk activity as labor raises issues about the setup of MTurk. The very low wages (below $2 an hour (Ross et al., 2009; Ipeirotis, 2010; Chilton et al., 2010)) are a first point. A further point concerns Amazon's choice of hiding any explicit relationship between Turkers and Requesters, even the basic workplace right of unionization is denied and Turkers have no recourse to any channels for redress against employers' wrongdoing, including the fact that they have no official guarantee of payment for properly performed work. Some regulation between Requesters and Turkers exists through Turkers' Blogs or Forums[4], or the use of Turkopticon[5] which is a tool designed to help Turkers to report bad Requesters; all these solutions are unofficial and nothing protects explicitely the Turkers, especially the new ones who are mostly unaware of these tools.

2.2. Does MTurk drastically reduce costs?

Most articles dealing with MTurk and resource production indicate low costs as the primary motivation. Given the observed salaries (for instance $0.005 to transcribe a 5-second speech segment (Novotney and Callison-Burch, 2010)), the cost may indeed be very low. However, the overall cost is not to be limited to the mere salary: the time needed to develop the interface, and to shackle the spammer problem is not negligible (Callison-Burch and Dredze, 2010); validation (Kaisser and Lowe, 2008) and correction costs (Xu and Klakow, 2010) to ensure minimal quality are also to be considered. Furthermore, some tasks may become more expensive than expected. This may occur for instance, if the required Turkers' competence is hard to find: to transcribe Korean speech (Novotney and Callison-Burch, 2010), the wages were increased from $5 to $35 per hour.

2.3. MTurk allows for building resources of equivalent quality?

Many technical papers have reported that at least for transcription and translation, the quality is sufficient to train and evaluate statistical translation/transcription systems (Callison-Burch and Dredze, 2010; Marge et al., 2010). However, some of these papers bring to light quality problems[6].

2.3.1. Limitations due to the lack of expertise

Turkers being non experts, the requester has to decompose complex tasks into simpler tasks (HIT, Human Intelligence Task), to help performing them. By doing so, s/he can be led to make choices that can bias the results. An example of this type of bias is analyzed in (Cook and Stevenson, 2010), where the authors acknowledge the fact that proposing only one sentence per lexical evolution type (amelioration and pejoration) influences the results.

Even more problematic is the fact that the quality produced with MTurk on complex tasks is not satisfactory. This is for example the case in (Bhardwaj et al., 2010), in which the authors demonstrate that, for their task of word-sense disambiguation, a small number of well-trained annotators produces much better results than a larger group (the number being supposed to counterbalance non-expertise) of Turkers. From this point of view, their results contradict those presented in (Snow et al., 2008) on a task that is similar, though much simpler.The same difficulty arises in (Gillick and Liu, 2010), in which it is demonstrated that non expert evaluation of summarization systems is "risky", as the Turkers are not able to obtain results comparable to that of experts. More generally, this quality issue can be found in numerous articles in which the authors had to validate Turkers' results using specialists (PhD students in (Kaisser and Lowe, 2008)) or use a rather complex post-processing (Xu and Klakow, 2010). Finally, the quality of the work from non experts varies considerably (Tratz and Hovy, 2010).

Moreover, there is currently a "snowball" effect going on, that leads to overestimate the resources quality mentioned in articles: some researchers praise MTurk (Xu and Klakow, 2010), citing research that did use the system, but would not have given usable results without a more or less heavy post-processing (Kaisser and Lowe, 2008). A simplistic conclusion to that could be that MTurk should only be used for simple tasks, however, besides the fact that MTurk itself induces important limitations (see next section), it is interesting to notice that, in some simple cases, Natural Language Processing tools already provide better results than the Turkers (Wais et al., 2010).

2.3.2. Limitations due to MTurk itself

In (Tratz and Hovy, 2010), the authors note that the limits of the user interface constitute the "first and most important drawback of MTurk". The authors also regret that it is impossible to be 100% sure that the Turkers participating in the task are real native English speakers. If pre-tests can be designed to address, at least partly, this issue, they represent an added cost and it will still be very easy to cheat (Callison-Burch and Dredze, 2010). Of course, you can always organize various protections (Callison-Burch and Dredze, 2010), but here again, this requires time and therefore represents an additional cost that only few requesters are ready to pay for. For example, in (Xu and Klakow, 2010), the authors identified spammers but did not succeed in eliminating them.

Finally, the impact of task payment should not be neglected, as it induces as logical behavior to place the number of performed tasks above the quality, regardless of payment. In (Kochhar et al., 2010) the authors thus reached the conclusion that an hourly payment was better (with some verification and time justification procedures).

[4] for instance `mechanicalturk.typepad.com` or `turkers.proboards.com`

[5] `turkopticon.differenceengines.com`

[6] some of the problems reported, such as the interface problems, are not specific to MTurk, but are generic to many crowdsourcing systems.

3. Existing or suggested alternatives

MTurk is not the only way to achieve fast development of high quality resources at a low cost. First, and despite the lack of systematic studies, existing automatic tools seem to perform as well as (non-expert) Turkers, if not better, on certain tasks (Wais et al., 2010). Second, exploiting as much as possible existing resources can be a cheap alternative to MTurk. Finally, MTurk is not the only crowdsourcing and microworking platform.

3.1. Unsupervized and semisupervized techniques for low-cost language resource development

Unsupervised machine learning techniques have been studied in the Speech & Language community for quite a long time, for numerous and sometimes complex tasks, including tokenization, POS tagging (Goldwater and Griffiths, 2007), parsing (Hänig, 2010) or document classification. Although such techniques produce results that are below state-of-the-art supervised or symbolic techniques, which both require resources that are costly to develop, it is unclear whether they produce results that are below what can be expected from MTurk, especially for complex tasks such as parsing. Moreover, unsupervised techniques can be improved at a reasonable cost by optimizing the construction and use of a limited amount of additional information (annotations, external resources). This constitutes the **semi-supervised learning** paradigm (Abney, 2007). Such approaches for developing language resources rely on two (complementary) principles:

- Training models on a limited amount of annotated data and use the result for producing more annotation. For example, using one model, one can select within the automatically annotated data those that have a high confidence level, and consider that as additional training data (*self-training*, (Yarowsky, 1995)). Using two different models allows to using the high-confidence annotations of one model for augmenting the training corpus for the other, thus decreasing systematic biases (*co-training*, (Blum and Mitchell, 1998)). If one accepts to produce a limited amount of manual annotations not only in advance but also while developing the tools, one can request the manual annotation of carefully chosen data, i.e., data for which knowing the expected output of the system improves as much as possible the system's accuracy (*active learning* (Cohn et al., 1995)).

- Using data containing annotations that are less informative, complete and/or disambiguated than the target annotations. Examples thereof include a morphological lexicon (i.e., an ambiguous POS-annotation) for POS tagging (Smith and Eisner, 2005), a morphological description for morphological lexicon induction (Sagot, 2005) or a partly bracketed corpus for full parsers (Watson et al., 2007).

3.2. Reusing existing resources

Even less costly is the **use of existing data** for creating new language resources. An example is the named-entity recognition (NER) task. MTurk has been used for developing NER tools, in particular for specific domains such as medical corpora (Yetisgen-Yildiz et al., 2010), twitter (Finin et al., 2010) or e-mails (Lawson et al., 2010). However, converting Wikipedia into a large-scale named-entity-annotated resource leads to building high-quality NER tools (Nothman et al., 2008), including when evaluated on other types of corpora (Balasuriya et al., 2009). Apart from Wikipedia (and the related DBpedia), other wiki projects (e.g., wiktionaries) and freely-available resources (lexicons, corpora) are valuable sources of information.

3.3. Collaborative or crowdsourced development beyond MTurk

All these alternatives require a fair amount of expert work. Other approaches do exist that reduce this requirement to a low level, and in particular collaborative and game-based techniques, as well as other crowdsourcing platforms than MTurk, which try and avoid at least in part its pitfalls. **Collaborative approaches** for language resource development rely on the strategy set up by the Wikipedia and other Wikimedia projects, as well as other wikis such as Semantic Wikis (Freebase, OntoWiki...). Anyone can contribute linguistic information (annotation, lexical data...), but usually contributors are motivated because they are to some extent themselves experts. The quality control is usually done mutually by contributors themselves, sometimes by means of online discussions, often leading to high quality results. One of the first collaborative platforms for language resource development was the semantic annotation tool Serengeti (Stürenberg et al., 2007), currently used within the AnaWiki project.[7] However, such approaches remain more suitable for developing medium-scale high-quality resources. For the fast development of large-scale resources, another strategy is to attract a large number of non-experts thanks to online games, that fall in the family of so-called **games with a purpose (**GWAP**)**. This idea was initiated by the ESP online game (von Ahn, 2006) for image tagging. Its success led researchers to develop such games for various tasks, including language-related ones. A well-known example is *PhraseDetective* (Chamberlain et al., 2008) for annotating anaphoric links, a reputedly complex task, which lead the authors to include a training step before allowing players to actually provide new annotations. However, the boundary between GWAPs and crowdsourcing is not clear-cut. It is not the case that MTurk remunerates a work whereas other approaches are purely "for fun". Indeed, even contributing to Wikipedia is a job, though a voluntary unpaid job. GWAP and MTurk cannot be distinguished either by the fact that MTurk gives a remuneration, as some GWAPs do propose non-monetary rewards (e.g., Amazon vouchers for PhraseDetective). Finally, collaborative and GWAP-based techniques are not the only "ethical alternatives", since ethical crowdsourcing platforms do exist. For gathering language data, in particular for less-resourced languages (LRL), **crowdsourcing platforms apart from MTurk** seem to be particularly appropriate, as shown for example by speech corpus acquisition experiments using dedicated applications run on mobile phones (Hughes et al., 2010). An example of an ethical crowdsourcing platform is Samasource, an NGO that allows really poor people to be correctly trained and paid for specific tasks (e.g., translating SMS in Creole after the earthquake in Haiti for helping victims and international rescuers to communicate).[8]

[7] http://www.anawiki.org

[8] http://www.samasource.org/haiti/

3.4. Optimizing the cost of manual anotation: pre-annotation and dedicated interfaces

When using approaches that rely on expert annotation, this annotation can be sped up and sometimes even improved by automatic annotation tools used as **pre-annotators**. For instance, (Fort and Sagot, 2010) have shown that for POS tagging, a low-quality and non-costly pre-annotation tool can significantly improve manual annotation speed; 50 fully manually POS-annotated sentences are enough for training a pre-annotation tools that reduces manual work as much as a state-of-the-art POS tagger, allowing to developing a 10,000-sentence standard-size corpus in $\sim$100 hours of expert work. On the other hand, on such a task, one could question the ability of anonymous Turkers to correctly follow detailed and complex annotation guidelines.

Obviously, the above-mentioned remarks by (Tratz and Hovy, 2010) about the limitations of MTurk interfaces apply more generally. Past projects aiming at developing syntactically and semantically annotated corpora have shown that both the speed and quality of the annotation is strongly influenced by the annotation interface itself (Erk et al., 2003). This provides another source of improvements for annotation efficiency and quality. Put together, it might well be the case that even costly expert work can be used in optimized ways that lead to high-quality resources at a reasonable cost, even compared with that of MTurk.

4. Conclusion and perspectives

We tried to demonstrate here that MTurk is no panacea and that other solutions exist allowing to reduce the development costs of quality language resources, while respecting those working on the resources and their skills.

We would like, as a conclusion, to go beyond the present facts and insist on the longer term consequences of this trend. Under the pressure of this type of low-cost systems, funding agencies could become more reluctant to finance language resources development projects at "normal" costs. The MTurk cost would then become a *de facto* standard and we would have no other choice as for the development method.

We saw, in section 3.3., that a microworking system can generate paid tasks while preserving ethics. This can even represent a chance for people who cannot participate in the usual labor market, due to their remoteness, their handicap, etc, but it implies a strict legal framework to ensure that the system does not violate their rights as workers. This is why we propose that the concerned associations, like ACL[9] for natural language processing, ISCA[10] for speech and ELRA[11] for Language Resources take care of this problem and push to the development of the needed tools to assure quality and ethics to language resources.

5. Acknowledgments

This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation, as well as part of the French ANR project EDylex (ANR-09-CORD-008) and of the Network of Excellence "Multilingual Europe Technology Alliance (META-NET)", co-funded by the 7th Framework Programme of the European Commission through the contract T4ME (grant agreement no.: 249119).

[9] http://www.aclweb.org/

[10] http://www.isca-speech.org/

[11] http://www.elra.info/

6. References

Steven Abney. 2007. *Semisupervised Learning for Computational Linguistics*. Chapman & Hall/CRC, 1ère edition.

Gilles Adda 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.

Dominic Balasuriya, Nicky Ringland, Joel Nothman, Tara Murphy, and James R. Curran. 2009. Named entity recognition in wikipedia. In *People's Web '09: Proceedings of the 2009 Workshop on The People's Web Meets NLP*, pages 10–18, Suntec, Singapore.

Vikas Bhardwaj, 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, Suède.

Lukas Biewald. 2010. Better crowdsourcing through automated methods for quality control. *SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation*, January.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *COLT: Proceedings of the Workshop on Computational Learning Theory*. Morgan Kaufmann Publishers.

Chris Callison-Burch 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*, Los Angeles, California.

J. Chamberlain, 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.

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

David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. 1995. Active learning with statistical models. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7, pages 705–712. The MIT Press.

Paul Cook and Suzanne Stevenson. 2010. Automatically identifying changes in the semantic orientation of words. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may.

Katrin Erk, Andrea Kowalski, and Sebastian Pado. 2003. The salsa annotation tool. In Denys Duchier and Geert-Jan M. Kruijff, editors, *Proceedings of the Workshop on Prospects and Advances in the Syntax/Semantics Interface*, Nancy, France.

Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, Los Angeles, California.

Karën Fort and Benoît Sagot. 2010. Influence of Pre-

annotation on POS-tagged Corpus Development. In *Proc. of the Fourth ACL Linguistic Annotation Workshop*, Uppsala, Suède.

Karën Fort, Gilles Adda, and Kevin Bretonnel Cohen. 2011. Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics (editorial)*, 37(2).

Dan Gillick 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, Los Angeles, California.

Sharon Goldwater and Thomas Griffiths. 2007. A fully bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of ACL*, Prague, Czech Republic.

Christian Hänig. 2010. Improvements in unsupervised co-occurrence based parsing. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, pages 1–8, Uppsala, Sweden.

Thad Hughes, Kaisuke Nakajima, Linne Ha, Atul Vasu, Pedro Moreno, and Mike LeBeau. 2010. Building transcribed speech corpora quickly and cheaply for many languages. In *Proceedings of Interspeech*, pages 1914–1917, Makuhari, Chiba, Japon, Septembre.

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

Michael Kaisser and John B. Lowe. 2008. Creating a research collection of question answer sentence pairs with amazon's mechanical turk. In *Proceedings of the International Language Resources and Evaluation (LREC-2008)*.

S. Kochhar, S. Mazzocchi, and P. Paritosh. 2010. The anatomy of a large-scale human computation engine. In *Proceedings of Human Computation Workshop at the 16th ACM SIKDD Conference on Knowledge Discovery and Data Mining, KDD 2010*, Washington D.C.

Nolan Lawson, Kevin Eustice, Mike Perkowitz, and Meliha Yetisgen-Yildiz. 2010. Annotating large email datasets for named entity recognition with mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 71–79, Los Angeles, California.

Matthew Marge, 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.

Joel Nothman, James R. Curran, and Tara Murphy. 2008. Transforming Wikipedia into Named Entity Training Data. In *Proceedings of the Australian Language Technology Workshop*.

Scott Novotney 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, Los Angeles, California.

Joel Ross, 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.

Joel Ross, 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, New York, NY, USA. ACM.

Benoît Sagot. 2005. Automatic acquisition of a Slovak lexicon from a raw corpus. In *Lecture Notes in Artificial Intelligence 3658 ((c) Springer-Verlag), Proceedings of TSD'05*, pages 156–163, Karlovy Vary, Czech Republic.

M. Six Silberman, Joel Ross, Lilly Irani, and Bill Tomlinson. 2010. Sellers' problems in human computation markets. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 18–21.

N. Smith and J. Eisner. 2005. Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 354–362, nn Arbor, Michigan, USA.

Rion Snow, 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*.

Maik Stürenberg, Daniela Goecke, Nils Die-wald, Irene Cramer, and Alexander Mehler. 2007. Web-based annotation of anaphoric relations and lexical chains. In *ACL Workshop on Linguistic Annotation Workshop (LAW)*, Prague, Czech Republic.

Stephen Tratz 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, Suède, July.

Luis von Ahn. 2006. Games with a purpose. *IEEE Computer Magazine*, pages 96–98.

Paul Wais, Shivaram Lingamneni, Duncan Cook, Jason Fennell, Benjamin Goldenberg, Daniel Lubarov, David Marin, and Hari Simons. 2010. Towards building a high-qualityworkforce with mechanical turk. In *Proceedings of Computational Social Science and the Wisdom of Crowds (NIPS)*, December.

Rebecca Watson, Ted Briscoe, and John Carroll. 2007. Semi-supervised training of a statistical parser from unlabeled partially-bracketed data. In *Proceedings of the 10th International Conference on Parsing Technologies*, IWPT '07, Prague, Czech Republic.

Fang Xu and Dietrich Klakow. 2010. Paragraph acquisition and selection for list question using amazon's mechanical turk. In *Proceedings of the International Language Resources and Evaluation (LREC-2010)*, pages 2340–2345, La Valette, Malte, May.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, Cambridge, MA.

Meliha Yetisgen-Yildiz, Imre Solti, Fei Xia, and Scott Russell Halgrim. 2010. Preliminary experience with amazon's mechanical turk for annotating medical named entities. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 180–183, Los Angeles, California.