Crowdsourcing for Speech: Economic, Legal and Ethical analysis

Gilles Adda, Joseph Mariani, Laurent Besacier, Hadrien Gelas

Crowdsourcing for Speech: Economic, Legal and Ethical analysis

Gilles Adda[1], Joseph J. Mariani[1,2], Laurent Besacier[3], Hadrien Gelas[3,4]
(1) LIMSI-CNRS, (2) IMMI-CNRS, (3) LIG-CNRS, (4) DDL-CNRS

September 19, 2014

1 Introduction

With respect to spoken language resource production, Crowdsourcing – the process of distributing tasks to an open, unspecified population via the internet – offers a wide range of opportunities: populations with specific skills are potentially instantaneously accessible somewhere on the globe for any spoken language. As is the case for most newly introduced high-tech services, crowdsourcing raises both hopes and doubts, certainties and questions. A general analysis of Crowdsourcing for Speech processing could be found in (Eskenazi et al., 2013). This article will focus on ethical, legal and economic issues of crowdsourcing in general (Zittrain, 2008a) and of crowdsourcing services such as Amazon Mechanical Turk (Fort et al., 2011; Adda et al., 2011), a major platform for multilingual language resources (LR) production. These issues include labor vs leisure, spare time vs working time, labor organization and protection, payment and rewards etc. Given the multifaceted aspects of the subject, separating the wheat from the chaff might require an entire book, likely in a sociological or political science book series. This is clearly not the objective of the present contribution. However, given both the emerging role of crowdsourcing services as scientific tools and the ethical demands of science and research, a few issues of particular importance will be examined in order for researchers to sharpen their analysis and judgment. Some, such as the legal problems, are offputting, and others are extraordinarily complex, as is the case of the economic models, but all are facets of crowdsourcing.

Crowdsourcing is a neologism designed to summarize a complex process within a single word. To examine how ethics and economy are intertwined in crowdsourcing, the concept will be dissected and a short review of the different crowdsourcing services will be presented. We will describe the major ethical and economic issues raised by representative crowdsourcing and microworking services, with a focus on Amazon Mechanical Turk (MTurk), the main crowdsourcing, microworking service used nowadays by researchers in speech. In the context of this article, Microworking refers to the division of tasks into multiple parts and Crowdsourcing refers to the fact that the job is outsourced via the

web and done by many people (paid or not). In particular, the issue of compensation (monetary or otherwise) for the completed tasks will be addressed, as will be the ethical and legal problems raised when considering this work as labor in the legal sense. This is particularly relevant when the tasks are in competition with activities performed by salaried employees. The proposed debate has to be considered in relation to both the economic models of the various crowdsourcing services and the task to be performed. The use of crowdsourcing for under-resourced languages will be presented as a case study to exemplify the different issues exposed beforehand. Finally, this contribution aims to propose some specific solutions for researchers who wish to use crowdsourcing in an ethical way. Some general solutions to the problem of ethical crowdsourced linguistic resources will be outlined.

2 The Crowdsourcing fauna

2.1 The crowdsourcing services landscape

The Crowdsourcing concept arose from the evidence that some tasks could be completed by Internet users, thus relying on the advantages of Internet, namely instantaneous access to a huge number of people all over the world. Internet users may be compensated for their contribution, and this compensation can be monetary or not, depending on the crowdsourcing system and the tasks performed. A first phase of crowdsourcing development relied on the specific competences of some internet users. WIKIPEDIA[1] is one of the earliest and probably one of the most famous representatives of crowdsourcing systems relying on volunteer work. Besides WIKIPEDIA, there were many projects of collaborative science, such as THE GALAXY ZOO[2]. The first paid crowdsourcing systems involved users with special (professional) abilities such as programming, with TOPCODER[3] or designing, with 99DESIGNS.[4] In the speech domain, early attempts at collaborative annotation, such as (Draxler, 1997) should be highlighted.

More recently the concept of *Human computing*, in which the only required abilities are to be a human and to have some spare time, has appeared. This is a transposition of the Grid Computing concept to humans. The idea is to harness advantage of humans' 'spare cycles' in order to develop a virtual computer of unlimited power, as the population potentially involved is no longer limited to a subset of Internet users with some special skills, but instead includes *any* Internet user. According to this concept, each user, like a processor in a grid, is assigned a basic sub-task and only has access to the minimal information required to perform his/her sub-task. If the task is straightforward and easy to explain (for instance, good quality orthographic transcription of monologues

[1] http://www.wikipedia.org/
[2] http://www.galaxyzoo.org/
[3] http://www.topcoder.com/
[4] http://99designs.com/

as in (Parent and Eskenazi, 2010)), then defining sub-tasks consists simply of splitting the data into small pieces. If the task is complex, it could be divided into successive, easier tasks, which in turn could be cut into small, simple elementary sub-tasks; for instance (Parent and Eskenazi, 2010) have adopted a corrective strategy with successive teams of workers.

This type of crowdsourcing services, called *Microworking*, can be further classified depending on whether or not monetary compensation is provided. In this article, monetary rewards refer only to rewards in cash. Many sites of cloud labor provide monetary compensations, such as MTurk or CLICKWORKER,[5] but GWAPs (*Games with a purpose*) which make also use of the concept of microworking, usually do not offer monetary compensations. The GWAP is another strategy of attracting large numbers of non-experts: through online games. It was initiated by the ESP online game (von Ahn, 2006) for image tagging. Many projects of collaborative science are set up as a GWAP, for instance FOLDIT[6] in the domain of protein folding. GWAPs provide entertaining or stimulating activities that are interesting enough to attract people willing to perform volunteer work, sometimes with non-monetary rewards (e.g., SWAG BUCKS).[7]. In the speech domain, collaborative experiments have been set up, such as (Gruenstein et al., 2009) about collecting and transcribing with an online educational game, or (Draxler and Steffen, 2005) about the recording of 1000 adolescent speakers. Finally, some microworking systems are in an ambiguous situation. For instance, RECAPTCHA[8] uses CAPTCHAs of words that optical character recognition (OCR) software failed to read. RECAPTCHA aims to contribute to the digitization of difficult texts in the Google book project. RECAPTCHA does not offer compensation for the work done. However, this work cannot be considered as 'voluntary work' as the user fills CAPTCHAs to get access to a service and not to help Google. This latest form of crowdsourcing, usually unbeknownst to the user, is described as 'epiphenomenal' (Zittrain, 2008a).

Amazon Mechanical Turk, introduced in 2005, is a precursor to and a leader of the myriad of paid microworking systems that exist today.

The boundary between GWAPs and other microworking crowdsourcing systems is not precise. It is not the case that microworking systems propose tedious tasks whereas other approaches are purely 'for fun': entertaining tasks do exist on MTurk (see for instance The Sheep Market.[9] GWAP and MTurk cannot be distinguished by the fact that MTurk provides remuneration, as some GWAPs do propose non-monetary rewards (e.g., Amazon vouchers for PHRASEDETECTIVE (Chamberlain et al., 2008)). Finally, collaborative and GWAP-based techniques are not the only 'ethical alternatives', since ethical crowdsourcing platforms such as SAMASOURCE[10] do exist.

To classify paid crowdsourcing services, (Frei, 2009) proposed four categories

[5] http://www.clickworker.com/
[6] http://fold.it/
[7] http://www.swagbucks.com/
[8] http://www.google.com/recaptcha/learnmore
[9] www.thesheepmarket.com/
[10] http://samasource.org/

based on their complexity. At the lower end of the scale, there is the category of *Micro tasks*, which includes MTurk. Here the tasks are small and easy, require no skills, and offer very low rewards. Next comes the category of *Macro tasks* with low pay and a substantial number of propositions, as in Micro tasks, which however require some specific skills (e.g. writing a product reviews). Then come the *Simple projects*, such as basic website design. Simple projects involve higher pay and fewer propositions while requiring more skills and time. At the highest level of the ranking scale are the *Complex tasks*, those which require specialized skills and a significant time commitment (for instance the tasks available in InnoCentive.[11] The two latter categories resemble tasks which can be encountered in the 'real' world, unlike the first two categories. For instance, there is no direct communication between requesters and workers for the first two categories, while for the latter, communication is required. Other interesting taxonomies exist such as the one presented in (Quinn and Bederson, 2011) which uses six distinguishing factors to classify the human computation systems. Moreover, Frei's taxonomy of crowdsourcing services has been rendered oversimplified by the appearance of Macro tasks or Simple project services built upon Micro tasks, such as CrowdForge[12] (Kittur et al., 2011) or TurKit[13] (Little et al., 2010). Frei's taxonomy is, however, still useful for defining some targeted solutions, including fair compensation for the work done (see section 5.1).

2.2 Who are the workers?

The backbone of the crowdsourcing system, workers constitute a population with rapidly evolving characteristics. This section will give some sociological details with recent facts and figures.

Country of origin The country of origin is not a selection criterion for most crowdsourcing services. For instance, oDesk's[14] active workers are coming from (in decreasing order) the Philippines, India, the United States, Ukraine, Russia, Pakistan, Bangladesh, noting especially the Philippines' workers who seem to work 24/7 (Ipeirotis, 2012b). Some services such as MTurk, do impose restrictions: MTurk limits monetary remuneration (cash incentives) to workers with a valid US bank account (payment in dollars) or to workers from India (payment in rupees). Recently, requesters tend to *a priori* reject Indian workers as they are more likely to be spammers or be less proficient in certain tasks involving language use; this change has led some Indian workers to lie about their location (Ipeirotis, 2011a). MTurk also requires that requesters provide a US billing address and a credit card, debit card, Amazon Payments account or US bank account in order to publish tasks. Some crowdsourcing services exclusively use

[11] http://www.innocentive.com/
[12] http://smus.com/crowdforge/
[13] http://groups.csail.mit.edu/uid/turkit/
[14] https://www.odesk.com/

underprivileged workers such as MobileWorks[15] or SamaSource.[16] MobileWorks has a team of workers from India and Pakistan ready to receive jobs via their mobile phone or computer and claims to be 'socially responsible,' suggesting that its workers are paid a fair wage to encourage higher quality work; SamaSource is a nonprofit organization which establishes contracts with enterprise customers or other crowdsourcing services (such as CrowdFlower)[17] in order to provide crowdsourcing microworking services to people living in poverty around the world.

MTurk seems to be quite particular given its bimodal distribution of workers: the ones from India and the ones from US. It is difficult to obtain accurate figures concerning these workers, given their anonymity in MTurk. There is some evidence that the exact number of workers actually working in MTurk is much smaller than the official figure of 500,000 registered workers in 2011 (see section 3.3).

Relying on surveys submitted within MTurk, studies in social sciences (Ross et al., 2010; Ipeirotis, 2010a) may provide some insight into workers' socio-economic profiles (country, age, ...), the way they use MTurk (number of tasks per week, total income in MTurk, ...), and how they qualify their activity. For instance, these studies enable one to estimate the number of Indian workers in MTurk: Indian workers represented 5% in 2008, 36% in December 2009 (Ross et al., 2010), 50% in May 2010[18] and have generated over 60% of the activity in MTurk (Biewald, 2010).

Sociological facts As for determining the number of workers, it is difficult to present an exact picture of who the workers in crowdsourcing services are. The studies in social sciences mentioned above revealed that, in MTurk, 91% of the workers expressed their desire to make money (Silberman et al., 2010), even if the observed wage was very low: \$1.25/hr according to (Ross et al., 2009) \$1.38/hr according to (Chilton et al., 2010). If 60% of the workers think that MTurk is a fairly profitable way of spending free time and earning cash, only 30% mentioned their interest in the tasks, and 20% (only 5% of the Indian workers) said that they were using MTurk to kill time. Finally, 20% (30% of the Indian workers) declared that they were using MTurk to make basic ends meet, and about the same proportion stated that MTurk was their primary source of income. Furthermore, the 20% of the most active workers who spend more than 15 hours per week with MTurk (Adda and Mariani, 2010) produce 80% of the overall activity.

The population and the motivations of the workers are heterogeneous. Nevertheless, those 20% of the workers for whom crowdsourcing is a primary income generate an activity that should be considered as a labor, even if the actual labor laws (see section 3.2) are unable to clearly qualify this activity as such. Moreover, there is a huge difference between good workers who have direct and

[15] http://www.mobileworks.com/
[16] http://samasource.org/
[17] http://crowdflower.com/
[18] http://blog.crowdflower.com/2010/05/amazon-mechanical-turk-survey/

regular connections with some requesters and know how to maneuver between the tasks to avoid scams, and naive workers for whom the crowdsourcing platforms blatantly lack a robust regulatory framework (see section 3.3). There is also a difference between workers who desperately need the money and those who do not: people who need money will also undertake low-paying tasks, as there is an insufficient number of high-paying tasks to fill a working day for those looking to daily earn a maximum of cash incentives (Adda and Mariani, 2010).

2.3 Ethics and Economics in crowdsourcing: How to proceed?

Economic and ethical problems in crowdsourcing are related (among others) to the type of crowdsourcing services, the nature of the task and of the workers' activity, as well as to the place where this activity is located. Given this complex situation, connecting all these parameters poses a very difficult multivariate problem.

In the following section 3, an overview of the economic model of crowdsourcing services will be presented, together with a summary of the situation regarding labor laws. Based on the insight gained from the case study of transcribing speech of under-resourced languages, some possible solutions may be envisioned for speech science, and more generally language sciences, in order to deal with crowdsourcing services in a more ethical and efficient way.

3 Economic and Ethical Issues

Beyond the previously mentioned opportunities, the rapid growth of crowdsourcing services introduces many problems, some of which are philosophical or ethical, as those mentioned in (Zittrain, 2008a), but also legal (Felstiner, 2011), while others are economic (Ipeirotis, 2010b). This section will present an overview of all these problems.

The main ethical and economic problems concern the worker, his/her relation with the task, the requester and the crowdsourcing service. Technically, workers are usually independent contractors. They are not subject to minimum wage or overtime protection. Ethical problems may arise in two situations: if the task is comparable to a human experiment, or if it corresponds to real labor. In both cases, researchers have specific ethical obligations. For instance, speech corpora transcription tasks were being performed for years by employees of agencies like LDC[19] or ELDA[20] while the collection of speech data was carried out on a more volunteer basis. For tasks which were performed by salaried employees, crowd labor could be viewed as offshoring on the Web.

If the task corresponds to a human experiment, some experiments that would be legal for a private organization would not be approved by a university In-

[19]Linguistic Data Consortium, `http://www.ldc.upenn.edu/`
[20]Evaluations and Language resources Distribution Agency, `http://www.elda.org/`

stitutional Review Board following the US National Research Act of 1974. For instance, to obtain the authorization of the Virginia Commonwealth University IRB (Virginia Commonwealth University, 2009), there are the following guidelines about payment: 'Compensation for Research Participants: Payment for participation in research may not be offered to the subject as a means of coercive persuasion. Rather, it should be a form of recognition for the investment of the subject's time, loss of wages, or other inconvenience incurred.' Considering the statement in section 2.2 that a significative fraction of the workers who are involved in microworking platforms seem to have no alternative way of earning, many experiments that are using these platforms are hardly compliant with most IRBs. Moreover, IRBs usually require that the study participants sign some charter or agreement to ensure their informed consent. Fortunately, the collection of anonymous speech or annotations is not considered to be a human experiment, and therefore does not really fall under the scope of IRB regulation. Nevertheless, it is always a good practice to explain the whys and wherefores of the study to the participants, and, when possible, to obtain a signed agreement from them.

If the task corresponds to labor (as do for instance most of the tasks involving transcription or translation), the main question is the hourly wages paid in the crowdsourcing platforms, which are significantly lower than the minimum wage in many countries ($7.25/hour in the US, 9.22 €/hour in France), and which may entail some ethical and economical problems. Defining a useful minimum hourly wage in crowdsourcing services is quite difficult (see section 5.1). This minimum hourly wage should take into account the advantages of these tasks for the workers, such as self-assignment, the lack of time and money spent on commuting, and the fact that the crowd is not located in a single country, but defining a minimum hourly wage could help in addressing some of the problems uncovered in this article.

This question of labor within crowdsourcing services is situated in a more general framework. For instance in (Albright, 2009), it is noticed that the outsourcing of some jobs on crowdsourcing services is hardly avoidable: 'Recognize that it is happening,' advised Carl Esposti, founder of `crowdsourcing.org`, which tracks the industry. 'It's happening and an absolute inevitability that a new market for work is being created on both the supply and demand sides. An individual may not like this and may not want to participate, but they have no choice.' Esposti, from `crowdsourcing.org`, advises IT professionals who are currently employed to pay attention to this new business model and to the impact it could have on their careers. He suggests that individuals should try to determine whether crowdsourcing will be constructive or destructive for their organizations and how it might relate to their own particular jobs. He noted, for example, that companies do not crowdsource entire functions: rather they crowdsource work that can be broken up into manageable tasks. The more a person's job is activity-based, therefore, the more his or her job could be at risk.

3.1 What are the problems for the workers?

There are numbers of issues reported by the workers involved in crowdsourcing services. Some are general to all crowdsourcing platforms, while others are specific, but MTurk seems to manifest most of the problems and thus merits a closer examination.

Very low wages (below $2 an hour (Ross et al., 2009; Ipeirotis, 2010a; Chilton et al., 2010) in MTurk) are a first point to be addressed. In section 2.2 it is mentioned that 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, 2010a). Below are some statements, from different sources, (Ipeirotis, 2008), Turkopticon (*supra*), Turker Nation,[21] illustrating the economic situation of some MTurk workers in the US:

> 'I realize I have a choice to work or not work on AMT, but that means I would also not need to make the choice to eat or not eat, pay bills or not pay bills, etc.'
>
> 'How do you make ends meet on a dollar an hour? You don't. All you do is add to what you make with your regular job and hope it is enough to make a difference.'
>
> 'I don't know about where you live, but around here even McDonald's and Walmart are NOT hiring. I have a degree in accounting and cannot find a real job, so to keep myself off of the street I work 60 hours or more a week here on mTurk just to make $150–$200. That is far below minimum wage, but it makes the difference between making my rent and living in a tent.'
>
> 'I am currently unemployed and for some reason absolutely can not find a job. Every job I apply for either turns me down or I don't hear from them at all. I have been doing online surveys, freelance writing, and MTurk to try to make the most money I can. I don't make much but when you literally have no savings and no income you take what you can get.'
>
> 'No available jobs in my area, have applied to over 40 jobs no calls so far been 3 months. Do it to pay my bills which includes rent and diapers for my kids until I find work again.'

One may question whether or not the workers are free to choose or not this way of making money, especially considering the actual level of total or partial unemployment in US and in Europe, and the living standards in Third World countries.

But this conclusion must be tempered by other statements, from other workers, which illustrate the fact that the situation is not black or white:

> 'I have agoraphobia which doesn't let me work outside the home. By turking, I feel that I can at least help out in some way with the bills and stuff.'

[21] `http://turkernation.com/`

'I have a high need for feedback and seeing my HITs get approved supplies me with that satisfaction.'

'Mturk has given me a sense of conviction that 'I can'. I have started to believe in myself and the journey has been so enriching. I just love this place and when we get paid for something we love – nothing like it. Thanks to mturk.'

'Mechanical Turk work is not only for money. This is an experience of the worldwide working methods.There are different kinds of hits. Every hit on this turk is challenge to our knowledge. So I like this job very much.'

'I am a retired teacher who finds the more academic hits stimulating.'

Another frequently mentioned problem (Silberman et al., 2010) is the fact that requesters pay late. In MTurk, there is an 'auto-approval' delay in the permission of payment when requesters neglect to approve the task. It is very common for the requesters to choose the maximum delay, which is thirty days. This means that until the end of the delay, which is not visible to workers, the worker does not know if his work will be approved and paid or not.

To pay late is a problem, but to not pay at all could be a real issue for the workers. For example, many reported experiments dealing with speech crowdsourcing have implemented automatic filters in order to reject completed tasks which seem inadequate. To block a worker or to reject many tasks may result in the worker being banned from the crowdsourcing service, which could have real consequences, especially for workers for whom this money is essential. As misunderstandings of the guidelines or errors in automatic procedures are always possible, automatic procedures should be handled with great precaution.

A further point concerns the choice of anonymity by many crowdsourcing vendors in order to protect the workers and the requesters from email spamming or from incorrect use of personal information. But anonymity hides any explicit relationship between workers, and between workers and requesters. Even the basic workplace right of unionization is denied and workers have no recourse to any channels for redress against employers' wrongdoings, including the fact that they have no official guarantee of payment for properly performed work. They may complain to the site that the requester did not behave correctly, but without any guarantee.

Some regulation between requesters and workers exists through workers' Blogs or Forums, such as the Mechanical Turk Blog[22] or Turker Nation,[23] or through the use of Turkopticon,[24] a tool designed to help workers report bad requesters. All these solutions, however, are unofficial, and nothing formally protects the workers, especially the new ones who are mostly unaware of these tools.

[22] `mechanicalturk.typepad.com`
[23] `http://turkernation.com/`
[24] `turkopticon.differenceengines.com`

Given the anonymity, another concern is the nature of the task itself and the fact that the real task is cut into small pieces and presented to the workers in a way that may obscure the purpose of the work. An extreme case was provided by Zittrain (Zittrain, 2008b) who mentioned the problem of matching photos of people, which could be used by an oppressive regime to identify demonstrators. But more common cases include the task of solving a captcha which will give a spammer access to a protected site, and the task of 'testing' if ads are working on a website, which will generate fake clicks but real money. Concerning speech science, it would be a good practice (see section 3) to explain the purpose of the whole task to the workers in order to allow them the option of not participating in a study they do not agree with, such as an experiment in the domain of biometrics or a study funded by the army.

3.2 Crowdsourcing and labor laws

Given the ethical problems listed in previous sections, one may wonder why the law is not applied to the regulation of crowd labor. Some authors (Felstiner, 2011; Wolfson and Lease, 2011) looked at the possible extension of US labor laws to the crowdsourcing workplace. They found quite difficult to decide with precision how the laws could be applied to crowdsourcing. The first reason is the heterogeneity of the crowdsourcing: section 2.1 already mentioned that crowdsourcing platforms could be very different, going from micro task platforms such as MTurk, to complex task platforms such as InnoCentive[25] or oDesk.[26] The second and main reason is the inappropriateness of existing laws for dealing with the Internet, and especially with *work* on the Internet. In order to limit the complexity, (Felstiner, 2011; Wolfson and Lease, 2011) mainly looked at the application of United States state and federal laws to MTurk, as it is one of the most used crowdsourcing platforms.

Worker status in participation agreement. The main goal is to determine the exact status of workers: many crowdsourcing platforms (the vendors) include in their terms of use a statement that defines the workers as independent contractors. Workers are supposed to have accepted this term with the 'click-wrap' participation agreement; for instance, MTurk Participation Agreement contains the statement: 'As a Provider (worker), you are performing Services for a Requester in your personal capacity as an independent contractor and not as an employee of the Requester.' The Amazon terms of use upon registration state that workers are not allowed 'to use robots, scripts, or other automated methods to complete the Services,' and that they should furnish the requester with 'any information reasonably requested' and agree with not being entitled to any employee benefits or eligible for worker's compensation if injured. Amazon can cancel a worker's account at any time. When this happens, the worker loses all the earnings left in his Amazon account. As independent contractors,

[25] `http://www.innocentive.com/`
[26] `https://www.odesk.com/`

they have no protection of any sort and should arrange for their own insurance, pay self-employment taxes, etc. By clickwraping participation agreements, requesters and workers of many crowdsourcing platforms are supposed to have accepted this contractual agreement. Clickwrap participation agreements, however, present two pitfalls:

- many requesters and workers do not really read the agreement and do not have a 'clear' view of the contents of this contract;
- the agreement has been drawn up by only one partner (the vendor), which calls into question the negotiated nature of the contract.

As stated by (Felstiner, 2011): 'The vendors, in binding both workers and firms to their clickwrap, have, in essence, prospectively *filled in* the content of the worker-firm contract.' What is very clear in the participation agreement is that its terms uniformly disclaim *any* vendor responsibility.

But even though workers agreed with a click on the clickwrap participation agreement to classify themselves as 'independent contractors', it is not a decisive determination of their status. Firstly, this status is not always clear, even in participation agreements; for instance, in MTurk Participation Agreement: 'Repeated and frequent performance of Services by the same Provider on your behalf could result in reclassification of independent contractor employment status.' Secondly, the courts have already ruled that when the work is essentially done in the capacity of an employee, putting an 'independent contractor' label on the worker does not exempt him or her from the protection of the act.[27]

Employee or independent contractor: status of the crowd worker under FLSA. To decide if crowdsourcing workers are statutory employees or independent contractors, (Felstiner, 2011; Wolfson and Lease, 2011) use the Fair National Standard Acts (FLSA), which, given that the parties are 'employers' and 'employees', defines a federal minimum wage and overtime protection, and the National Labor Relation Act (NLRA). Courts have developed a series of tests to decide if someone is an employee under the FLSA or NLRA. In order to decide if, in the relations between the vendor, requesters and workers, some elements could be qualified as employer–employee relations, (Felstiner, 2011; Wolfson and Lease, 2011) look at the applicability of these tests on the crowdsourcing case. Under FLSA, courts use a multi-factor 'economic reality' test with seven factors. No single factor is determinative, thus all the factors are examined, with a different weight:

(i) *How integral the work is to the employer's business.* There is a large variety of requesters, some relying entirely upon crowd labor, others using crowdsourcing only sparsely. This factor is not decisive to determine a worker's status.

[27] see for instance *Supreme Court in Tony and Susan Alamo Foundation v. Secretary of Labor, 471 U.S. 290 (1985)*

(ii) *The duration of relationship between worker and employer.* Some workers work repeatedly with the same requester, as in oDesk, but the relationship between requesters and workers could not be qualified as permanent. The sole long-term relationship is between these parties and the crowdsourcing service.

(iii) *If the worker had to invest in equipment or material himself to do the work.* This factor is often decisive, but the definition of 'equipment' for the work on the Internet is vague. Is it the computer, which is basic equipment, or the specific web platform developed by the crowdsourcing service? This question could be debated, but courts have tended to be neutral on similar cases about tele-working.

(iv) *How much control the employer has over the worker.* The control of the requester over the worker is not direct, but (through the participation agreement) the requesters have a high level of control over how the work is done. In comparison with a contractual relation, the use of this control is not negotiated.

(v) *The worker's opportunity for profit and loss.* Crowdsourcing vendors did not structure their services such that workers may build and grow a business, and worker's opportunities for profit and loss are quite limited.

(vi) *How much skill and competition there is in the market for this type of work.* As in the preceeding factor, crowdsourcing vendors leave very little room for initiative, judgment and foresight. For microworking services, it is clear that almost anyone of any skill level may perform the proposed tasks.

(vii) *If the worker is an independent business organization.* For microworking services, it would be surprising, especially given the very low observed compensation, for a worker to build an 'independent business organization' devoted entirely to this activity. This could be different for complex tasks, such as the ones proposed in 99designs,[28] a crowdsourced design contest marketplace.

Concerning crowd labor, the last three factors weigh in favor of an employee status, while the first four do not decisively accord either employee or independent contractor status. Therefore, it is not clear if a crowd worker could be classified as employee under FLSA. But there is uncertainty, which means that potential requesters must be aware of the possibility of regulation.

Many constraints listed in the license agreements of many crowdsourcing vendors are worded in order to *not* match the FLSA or NLRA tests' factors, and more generally to eliminate any explicit relationship of subordination. We may think that this is designed to limit the risk of a reclassification of the status of crowdsourcing workers as employees.

[28] `http://99designs.com/`

Crowdsourcing vendor as joint employer. Workers in crowdsourcing services act as temporary employees, hired through a temporary staffing agency (here the vendor). In this case, the workers can be regarded as employees of the *vendor* instead of the requester. Usually, the vendor's participation agreement tries to reject this possibility, but if workers can show that the economic reality reflects an employee–employer relationship between the vendor and the pool of workers, courts could declare the vendor a *joint employer*. (Felstiner, 2011) argues that the vendors, and especially MTurk, could have difficulty escaping responsibility for the work rights of their workers as joint employers. Furthermore, any national or international regulation of the labor laws will be easier to apply to the vendor, who is relatively easy to identify and to locate. Regulation on the myriad of requesters, using only the informations provided by the requester to the vendor, may be difficult to apply efficiently.

Crowd workers' status in other countries. What about other countries? In France, the *Code du Travail* does not give an exact definition of who is a salaried employee. Instead, this status is accorded based on the jurisprudence, which lists three mandatory factors for deciding if a person is a salaried employee, linked to an employment contract:

1. a relationship of subordination with the employer;
2. a monetary compensation;
3. completion of a task.

The relation between the point 4 of FLSA and the point 1 of the above list is clear: the subordination, which is conclusive in France for deciding if a worker is an employee.[29] More generally, in many countries an individual will be considered an independent contractor if he or she independently carries out the job and if there is neither subordination nor exclusivity in the relationship between the parties. But as in the US case, laws tend worldwide to elevate substance over form when examining the parties' actual relationship. Therefore, as some of the relations between vendors and workers or between requesters and workers could be defined as an employee–employer relationship, the outcome would be uncertain if any individual workers, or national or international labor agency, were to take legal action against either crowdsourcing platforms or requesters.

Crowdsourcing and labor laws: a needed regulation. Considering the existing labor laws, it is quite difficult to qualify the workers' status in crowdsourcing services as employees. At the moment, crowd labor is in a 'gray area', because the current regulation is not adequate. It is likely that if the crowdsourcing market is still growing in terms of the number of workers and the size of the market, the national and international legislatures will take into account this innovation of the concept of work, and will amend the labor laws to regulate

[29] even if, in order to determine the status of a worker, all factors such as the points 1 and 3 of FLSA are taken into account in the French labor law jurisprudence

the market (Felstiner, 2011; Wolfson and Lease, 2011). Moreover, if crowd labor is growing at the expense of existing industries and jobs, instead of creating new activities with new types of workers, this will create a social pressure to address the deficiencies in the labor laws concerning crowd labor. The definition of a new type of temporary employment contract designed for crowdwork with monetary rewards could be beneficial to both the workers and the requesters. This contract, which could be drawn up between the vendor and the worker (see section 3.2), should help to regulate the crowdwork and to assure a stability and a clear legal framework for the requesters.

3.3 Which economic model is sustainable for crowdsourcing?

Pragmatically, making ethics a priority is feasible if the law enforces it and if the economic situation is compliant with it. Economics is indeed a major concern for obtaining the conditions for ethics in the real world. Highlighting some of the driving forces of its economic model will support our understanding of the crowdsourcing.

Low vs. high reward The frequent assumption that the low rewards are a result of the classical law of supply-and-demand (large numbers of workers means more supply of labor and therefore lower acceptable salaries) is false. First of all, this assumption relies on the belief that the number of workers is huge, while (Fort et al., 2011) observe that there are not too many active workers: (Fort et al., 2011) looked at the number of tasks effectively completed by the 1,000 MTurk workers (Ipeirotis, 2010a), and compared it to the total number of tasks completed according to the Mechanical Turk Tracker.[30] Taking into account the different factors, they found the number of effective workers to be below 50k, and the number of active workers below 10k. These figures are very far from the official figure of 500K registered workers. While only valid for MTurk, this ratio between the number of registered and active users/workers is compatible with other observations about the activity on the Internet, such as the "90–9–1" rule (Arthur, 2006), and the number of active workers close to 2% observed in (Ipeirotis, 2012a). This calculation could explain the difficulty in finding workers with certain abilities, such as understanding a specific language (Novotney and Callison-Burch, 2010), or speaking an under-resourced language: in section 4, the expected number of Swahili speakers available on MTurk is lower than could be expected from the number of people speaking Swahili in the US (3 instead of 32).

Many explanations could be provided to the fact that the rewards are so low. The first is that the low reward is a result of the requesters' view of the relation between quality and reward: many articles (see for instance (Marge et al., 2010)) observe that there is no correlation between reward and final quality. The reason is that increasing the price is believed to attract spammers (i.e. workers

[30] `http://mturk-tracker.com`

who cheat and not really perform the job, using robots or answering randomly instead). Spammers are numerous for instance in the MTurk system (Ipeirotis, 2010b) due to a worker reputation system that makes it easy for a spammer to build a new account with 100% approval rate (Ipeirotis, 2010c). This is 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 a period of 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 crowdsourcing system is stable in terms of the number of workers, because workers leaving the system are replaced by new workers unaware of this situation (70% of the workers use MTurk for less than 6 months (Ross et al., 2009)). A second explanation given in (Bederson and Quinn, 2011) uses the theory of moral hazard in economy (Holmstrom, 1979), which explains that, given that the requesters do not incur the full cost of their actions because of the asymmetry of the relation, the anonymity, and the fact that they could reject the work done,... the cost for the other party (the workers) increases. In turn, workers tend to generate 'just good enough' work or even cheat.

This lack of a well designed reputation system is a stumbling block for many microworking services. For instance, Amazon's attitude towards reputational issues is passive: Amazon, as other crowdsourcing service vendors, maintains its position as a neutral clearinghouse for labor, in which all other responsibility falls of the two consenting parties (see section 3.2).

As highlighted by (Ipeirotis, 2010b), without major developments, especially in financial rewards and reputation, flaws and a faulty economic model call into question medium term viability of microworking services such as MTurk.

Microworking crowdsourcing systems with low rewards should have difficulty keeping the 'good' workers, because of the process described above. If work quality is an important aspect of the task (for transcription or annotation of speech for instance), this model is not adequate: relying on low-quality workers is not cost-effective, even if the standard redundancy method is used to improve the quality. For instance (Ipeirotis, 2011c) shows that employing 3 high-quality Masters workers (the elite group of workers created in MTurk), and paying them 20% more than the usual price, results in the same quality as using 31 workers of 70% accuracy. It is far more beneficial to get around the system and retain good workers by paying them higher (but still modest) wages. In another example, (Chen and Dolan, 2011) note that given the incentive to maximize their rewards, workers often cheat; to solve this problem, (Chen and Dolan, 2011) used a 2-tiered payment system to reward workers who submit good descriptions. This is a way to select the most qualified workers and to grant them with a bonus if

they perform well. As the nature of the task becomes harder, (Chen and Dolan, 2011) report the benefits of longstanding relationships with the workers instead of the anonymous relation as proposed by MTurk: less quality control, ability to train the workers to improve their competence, ability to correlate reward and quality (fair rewards result in worker loyalty).

It is indeed one of the reason many crowdsourcing services such as MTurk do not adhere to fair wages: a number of good workers operate within a separate framework built by some long-standing requesters who give higher financial rewards. These good workers are not available for other newly-arrived requesters who offer the standard (very low) price.

Task vs time reward As a requester, you can see the effective hourly rate along with the average completion time. But MTurk workers do not have direct access to this information. In particular, they do not see the hourly rate, which is fundamental information for judging if the money received will be fair compensation for the work done and the time spent. A rational action of an experienced worker is to choose a large set of tasks, to use one unit task to test its real difficulty, and to determine his effective hourly rate. But the workers are not all experienced or rational. This method of payment provokes behavior which is not always compatible with quality work, as the worker is not aware of the hourly rate before choosing the task. The same behavior may be observed in online games: a gamer is keeping track of his absolute score or level, and not of the time needed to obtain them. Similarly, the worker looks at the *absolute* level of funding rather than at hourly rate: 'Today, I will work until I have made $10.', which is certainly not the best way to optimize the overall reward. Moreover (Kochhar et al., 2010) reached the conclusion that an hourly payment was better (with some verification and time justification procedures), as task payment logically encourages one to place the number of performed tasks above the quality, regardless of payment. Our experiments described in section 4 corroborate these observations.

Furthermore, piecework retribution is strictly regulated in developed countries in order to prevent a wage lower than the legal hourly minimum: for instance, piecework retribution, similar to other forms of variable remuneration, is possible in France only if it results in a wage above the legal minimum. But determining an hourly rate is difficult in any remote workplace (see section 5.1), as only all worked hours should be compensated; practical solutions should be learned from the telework/telecommuting case.

Which is the economic model? As pointed out in (Ipeirotis, 2011c), one may have the feeling that the visible flaws in some crowdsourcing services such as MTurk concerning reputation system, anonymity, and very low rewards, are deliberate. This hypothesis has been put forward because these flaws induce an undue advantage for the first-comers. They were able, using their own remuneration and reputation system, to catch the good workers and to subsequently keep them because the newcomers offering high rewards are overwhelmed by

spammers and thus disappear or reduce their rewards. On the other hand, the low rewards proposed by the newcomers keep the remuneration at a sufficiently low level in order to present a very competitive cost. A visible effect of this is the growing number of specialized services which serve as interfaces between requesters and microworking platforms: CastingWords,[31] SpeakerText,[32] Serv.io Translate[33]... The development of ethical crowdsourcing services enforces amendments and improvements to this model.

4 Under-resourced languages: a case study

It is difficult to discuss the ethical and economic aspects of crowdsourcing without experiencing the concept oneself. A case study is presented here, in a domain where crowdsourcing seems to be a particularly hot topic: the processing of under-resourced languages. For these languages, data collection and annotation (for instance speech transcription) is a particularly difficult problem and crowdsourcing is a very attractive tool, especially for connecting speech technology developers and language experts. Moreover, since many under-resourced languages are spoken in developing countries, the potential workers (native speakers of the under-resourced language considered) tend to be the ones (mentioned in section 2.2) who are more likely to rely on crowdsourcing for income, as do the Indian workers in the surveys from (Ipeirotis, 2010a; Ross et al., 2010). In this section, the transcription of a speech corpora of two under-resourced languages from Africa using crowdsourcing is evaluated, and the main results of this experiment as well as the lessons learned are presented.

4.1 Under resourced languages definition and issues

The term under-resourced languages introduced by (Berment, 2004) refers to a language characterized by some (if not all) of the following aspects: lack of a unique writing system or stable orthography; limited presence on the web; lack of linguistic expertise; lack of electronic resources for NLP (natural language processing) such as monolingual corpora, bilingual electronic dictionaries, transcribed speech data, etc. Developing a NLP system (e.g. a speech recognition system) for such a language requires techniques that go far beyond a basic re-training of the models. Indeed, processing a new language often leads to new challenges (special phonological systems, word segmentation problems, unwritten language, etc.). For its part, the lack of resources requires, innovative data collection methodologies (crowdsourcing being one of them) or models in which information is shared between languages (e.g. multilingual acoustic models (Schultz and Kirchhoff, 2006; Le and Besacier, 2009)). In addition, some social and cultural aspects related to the context of the targeted language bring

[31] `http://castingwords.com/`, speech transcription
[32] `http://www.speakertext.com/`, video transcription
[33] `http://www.serv.io/translation`, translation

additional problems: languages with many dialects in different regions; code-switching or code-mixing phenomena (switching from one language to another within the discourse); massive presence of non-native speakers (in vehicular languages such as Swahili).

4.2 Collecting annotated speech for African languages using Crowdsourcing

Recently MTurk has been studied as a means of reducing the cost of manual speech transcription. Most of the studies conducted on the use of MTurk for speech transcription have been done for the English language, which is one of the most well-resourced languages. The studies on English, including (Snow et al., 2008; McGraw et al., 2009), showed that MTurk can be used to cheaply create data for natural language processing applications. However, apart from a research conducted recently by (Novotney and Callison-Burch, 2010) on Korean, Hindi and Tamil, MTurk has not yet been studied as a means to acquire useful data for under-resourced languages. As for as these languages are concerned, it is all the more important to collect data using highly ethical standards, as doing so usually involves people from developing countries who may suffer from extremely low standards of living.

The use of MTurk for speech transcription has been studied in the hopes of developing Automatic Speech Recognition (ASR) for two under-resourced African languages without combining transcription outputs. The experimental setup, including the subject languages, is described in subsection 4.3. Subsection 4.4 presents the result of the experiment, and a discussion is provided in subsection 4.5.

4.3 Experiment Description

4.3.1 Languages

Amharic is a member of the Ethio-Semitic languages, which belong to the Semitic branch of the Afro-Asiatic super-family. It is related to Hebrew, Arabic, and Syrian. According to the 1998 census, it is spoken by over 17 million people as a first language and by over 5 million as a second language throughout different regions of Ethiopia. The language is also spoken in other countries such as Egypt, Israel and the United States. Amharic has its own writing system which is a syllabary. It is possible to transcribe Amharic speech using either isolated phoneme symbols or concatenated CV (Consonant Vowel) syllabary symbols.

Swahili is a Bantu language often used as a vehicular language in a wide area of East Africa. In addition to being the national language of Kenya and Tanzania, it is spoken in different parts of the Democratic Republic of Congo, Mozambique, Somalia, Uganda, Rwanda and Burundi. Most estimations claim over 50 million speakers (with only less than 5 million native speakers). Structurally, Swahili is often considered to be an agglutinative language (Marten, 2006). Despite being non-tonal, it displays other typical Bantu features, such

as noun class and agreement systems and complex verbal morphology. It was written with an Arabic-based orthography before it adopted the Roman script (standardized since 1930).

4.3.2 Corpora

Both Amharic and Swahili audio corpora were collected following the same protocol. Texts were first extracted from news websites and then segmented by sentence. Recordings were made by native speakers reading sentence-by-sentence with the possibility to re-record mispronounced sentences. The whole Amharic speech corpus (Abate et al., 2005) contains 20 hours of training speech collected from 100 speakers who read a total of 10,850 sentences (28,666 tokens). The Swahili corpus used in this study corresponds to three and a half hours read by 5 speakers (3 male and 2 female). The sentences read by speakers serve as our gold standards and will be used to evaluate the transcriptions obtained by MTurk.

4.3.3 Transcription Task

For the transcription task, all 1183 of the audio files between 3 and 7 seconds (mean length 4.8 sec and total one and a half hours) were selected from the Swahili corpus. The same number of files were selected from the Amharic corpus (mean length 5.9 sec). These files were published (a task for a file) on MTurk. To avoid cheaters, task descriptions and instructions were given in the respective languages (Amharic and Swahili). The Swahili transcriptions did not require a special keyboard, but for the Amharic transcriptions, workers were given the address of an online Unicode-based Amharic virtual keyboard,[34] as the workers would not necessarily have access to Amharic keyboards.

4.4 Results

4.4.1 Analysis of the workers' contributions

After manual approval using the MTurk web interface, some experiments were conducted to evaluate a posteriori different automatic approval methods. Table 1 shows the proportion of approved and rejected tasks for both approval methods (manual and automatic). The higher rate of rejected tasks for Amharic can be explained by the much longer period of time during which the task was made available to workers. The tasks rejected with the manual process contained empty transcriptions, copies of instructions, nonsensical text and tasks completed by workers without any knowledge of the language. This manual approval process is time consuming, thus an experiment involving automatic approval methods was also conducted. This was done a posteriori, and no worker was rejected using such an automatic procedure. As can be seen in Table 1, it is possible to obtain results equivalent to those of the manual approval with the following task filtering:

[34] `www.lexilogos.com/keyboard/amharic.htm`

	♯ workers		♯ tasks			
	AMH	SWH	AMH		SWH	
	(Man&Auto)		Man	Auto	Man	Auto
APP	12	3	589	584	1183	1185[35]
REJ	171	31	492	497	250	248
TOT	177[36]	34	1081		1433	

Table 1: Submitted tasks approval

(i) empty and short (shorter than 4 words) transcriptions;

(ii) transcriptions using non-Amharic writing system, including copy of URLs (for Amharic);

(iii) transcriptions containing bigrams of instructions and descriptions from the tasks;

(iv) transcriptions that are outside the distribution space set by $Avg + 3 * Stdv(log2(ppl))$ (where *ppl* is the perplexity assigned by a language model developed with a different text).

The detailed completion rate per day was analyzed for both languages. Among the 1183 sentences requested, 54% of the Amharic tasks were approved in 73 days. On the other hand, Swahili was completed after 12 days, thus showing that there is a substantial variety in the rate of completion among different languages. This result is important since it shows that the Amharic transcription could not be achieved using MTurk with this set-up.

One hypothesis for such a result could simply be the effective population having access to MTurk. A recent survey (Ipeirotis, 2010a) shows that 47% of the workers were from the United States, 34% from India and the last 19% were divided among 66 other non-detailed countries. However, U.S.ENGLISH[37] shows that Swahili speakers are less numerous than Amharic speakers in the United States.[38]

Moreover, Table 1 shows that workers doing coherent work were more numerous for Amharic than for Swahili (12 and 3, respectively). A more probable explanation would thus be the input burden for Amharic language, considering the necessity to use an external virtual keyboard and to copy/paste from another web page. The difficulty to perform this task while managing and listening to the audio file may have complicated the task and therefore discouraged workers.

[35] 7 AMH transcriptions and 4 SWH transcriptions that were approved manually were rejected automatically while 2 AMH and 2 SWH transcriptions that were rejected manually were approved automatically.

[36] It is the number of all the workers who submitted one or more Amharic tasks. It is not, therefore, the sum of the number of rejected and approved workers because there are workers who submitted some rejected tasks and some approved ones.

[37] `www.usefoundation.org/view/29`

[38] less than 40,000 Swahili speakers against more than 80,000 Amharic speakers

Nevertheless, the tasks' transcription productivity indicates similar mean worker productivities (15 and 17xRT for Amharic and Swahili, respectively). These numbers are close to the ones cited in (Novotney and Callison-Burch, 2010) for transcriptions of English (estimated at 12xRT). These numbers must, however, be considered with caution, since they do not include the time corresponding to the manual approval process or to the development of an ad-hoc automatic approval procedure.

4.4.2 Evaluation of workers' transcriptions quality

To evaluate workers' transcriptions (TRK) quality, the accuracy of the manually approved tasks was calculated based on our reference transcriptions (REF). As both Amharic and Swahili are morphologically rich languages, it was found relevant to calculate error rate at word-level (WER), syllable-level (SER) and character-level (CER). Furthermore, real usefulness of such transcriptions must be evaluated in an Automatic Speech Recognition system. Some misspellings or differences of segmentation (which can be quite frequent in morphologically rich languages) will indeed not necessarily impact system performance but will still inflate WER (Novotney and Callison-Burch, 2010). The CER is less affected and is therefore more reflective of the transcription quality than the WER. The reference transcriptions are the sentences read during corpora recordings, and reading errors may have occurred.

Table 2 presents error rates for each language depending on the computed level accuracy (five of the approved Amharic transcriptions and four of the Swahili ones were found unusable and were disregarded). As expected, WER is relatively high (16.0% for Amharic and 27.7% for Swahili) while CER is lower. It seems to approach disagreement among expert transcribers even if it was not possible to explicitely calculate such disagreement (because data was transcribed by only one worker without overlap). In the literature,[39] it was found that the word level disagreement for a non-agglutinative language with a well-normalized writing system ranges from 2 to 4% WER. The gap between WER and SER may also be a good weight indicator of the different segmentation errors resulting from the rich morphology.

	Amharic			Swahili		
Level	♯ Snt	♯ Unit	ER (%)	♯ Snt	♯ Unit	ER (%)
Wrd	584	4988	16.0	1179	10,998	27.7
Syl	584	21,148	4.8	1179	31,233	10.8
Chr	584	42,422	3.3	1179	63,171	6.1

Table 2: Error Rate (ER) of workers transcriptions

The low results for Swahili are clarified by providing per-worker ER. Among the three workers who completed approved tasks, two have similar disagreement

[39] `www.itl.nist.gov/iad/mig/tests/rt`

Frq	REF	TRK	Frq	REF	TRK
15	serikali	serekali	6	nini	kwanini
13	kuwa	kwa	6	sababu	kwasababu
12	rais	raisi	6	suala	swala
11	hao	hawa	6	ufisadi	ofisadi
11	maiti	maiiti	5	dhidi	didi
9	ndio	ndiyo	5	fainali	finali
7	mkazi	mkasi	5	jaji	jadgi

Table 3: Most frequent confusion pairs for Swahili

with REF: 19.8% and 20.3% WER, 3.8% and 4.6% CER. The last worker has a higher ER (28.5% WER and 6.3% CER) and was the most productive, performing 90.2% of the tasks. Looking more closely at error analysis, one could suggest that this worker is a second-language speaker with no difficulty listening and transcribing but with some variation in writing (see details below).

4.4.3 Error analysis

Table 3 shows the most frequent confusion pairs for Swahili between REF transcriptions and TRK transcriptions. Most of the errors can be grouped into five categories that can also be found in Amharic.

- Incorrect morphological segmentation: see words *nini, sababu*, both preceded by *kwa* in REF.
- Common spelling variations of words such as *serikali* and *rais* (sometimes even found in newspapers article).
- Misspellings due to English influence in loanwords like *fainali* and *jaji* (meaning 'final' and 'judge').
- Misspellings based on pronunciation (see words *kuwa, ndio, suala*).
- Misspellings due to personal orthographic convention, which can be seen in words *maiti, mkazi, ufisadi, dhidi*.

Errors in the last two categories were all made by the same worker (the most productive one, having a high WER). Our assumption that this worker is a second-language speaker relies on the errors' frequency and regularity. One interesting conclusion of this analysis is that an accurate check of the origin of the workers (native/non native) is not easy to implement. For instance, in this latter case, qualification tasks would not have been efficient enough to detect this non native speaker.

In an experiment not reported here, an ASR system was trained using both REF and TRK transcriptions and nearly similar performances for both languages were observed. This suggests, therefore, that non-expert transcriptions

using crowdsourcing can be accurate enough for ASR. It also highlights the fact that even if most of the transcriptions are made by second-language speakers, it will not particularly affect ASR performances. This result is not particularly surprising: it demonstrates that ASR acoustic model training is rather robust to transcription errors. This result is in line with other works published on unsupervised and lightly supervised training where the machine (instead of the workers) transcribes speech data that will later be integrated into the training set (Wessel and Ney, 2005).

4.5 Discussion and lessons learned

In this section, the use of Amazon's Mechanical Turk speech transcription for the development of acoustic models for two under-resourced African languages was investigated. The main results are the following:

- For a simple task (transcribing speech data), it is possible to collect usable data for ASR systems training; however, all languages are not equal in completion rate. The languages of this study clearly had a lower completion rate than English.
- Among the targeted languages, Amharic's task was incomplete after a period of 73 days; this may be due to a higher task difficulty (use of a virtual keyboard to handle Amharic scripts). This questions the use of Amazon's Mechanical Turk for less elementary tasks that require more of a worker's time or expertise.
- Analysis of the Swahili transcriptions shows that it is necessary to verify the workers' expertise (native / non native). However, designing a qualification test to filter out non-native workers is not straightforward. Furthermore, the acoustic model training is rather robust to workers transcription errors. The use of MTurk in this context can be seen as another form of the lightly supervised scenario where machines are replaced by workers.

MTurk has proved to be a valuable tool for NLP domains, and some recommended practices were already proposed in (Callison-Burch and Dredze, 2010), mainly concerning how to be productive with MTurk. However, one should be careful about the way in which the data are collected or the experiments are conducted in order to prevent any legal or ethical controversies. Due to the characteristics of MTurk discussed earlier in this article, it was decided, after that experiment, to work directly with a Kenyan institute[40] to collaboratively transcribe 12 hours of our web broadcast news corpus. In order to reduce the repetitive and time-consuming transcription task, a collaborative transcription process was considered, based on the use of automatic pre-annotations (pre-transcriptions) to increase productivity gains. Details on this procedure can be

[40] http://www.taji-institute.com/

found in (Gelas et al., 2012). At $103 per transcribed hour, such collaboration is significantly more expensive than using MTurk ($37 per transcribed hour), but in this situation both employer and employee benefit from a more equitable relationship between the two. The price was set by the workers and corresponds to the task as well as to the reality of the local labor market (setting the fair price for work is important in such a context). This resulted in a positive and complete involvement of the workers; direct communication was a major benefit compared to MTurk. It allowed for both direct feedback on the experiment and a sufficient margin for adaptation. Such a direct collaboration is just one example of what can be done in order to carry out research along the highest ethical standards.

5 Towards ethically produced Language Resources

The preceeding sections have illustrated the different economic, ethical, and legal problems of crowdsourcing. They are numerous and serious and, when paired with experiments such as the one described in section 4, may lead to the adoption of a reserved stance on crowd labor use in speech science. However, given its huge potential, crowdsourcing will continue to develop even if some do not wish to participate. Solutions will be proposed for some, if not all, of the problems listed in this article in order to enable speech researchers or agencies to make use of crowdsourcing in an ethical way.

These solutions could be individual, namely guidelines for good practices. For instance, (Wolfson and Lease, 2011) provides some useful advice about the legal concerns that could be summarized in few points:

Be Mindful of the Law National and local legislatures and agencies may create new laws and administrative rules to protect the crowdsourcing workers and preserve the local labor. Anyone involved in crowdsourcing should consider all the potential legal ramifications, and weigh the costs and benefits.

Use Contracts to Clearly Define Your Relationships The relationship between requester and worker, as defined by the clickwrap participation agreement provided by the crowdsourcing vendor, is not clear, and some crowdsourcing agreements may not stand up in court. Defining a clear contract between employer and worker could help resolve many problems in advance.

Be Open and Honest In order to prevent from possible problems, and especially to avoid legal problems, providers should be open and honest about their expectations so that workers can understand them and adjust their behavior.

Other solutions are general and involve the speech community as a whole in designing a more ethical framework.

After an outline of the various views concerning the difficult problem of the monetary compensation for the work done, a short overview of what could be the possible consequences of a 'laissez-faire' attitude, especially for the development of language resources, will be presented.

The different foreseeable individual and general solutions will be presented in the hopes of eliminating or at least reducing the ethical problems.

5.1 Defining a fair compensation for work done

There are evident solutions for the problem of payment to workers, including those pointed out by Sharon Chiarella, vice-president of Amazon Mechanical Turk (Chiarella, 2011):

- Pay well – Don't be fooled into underpaying Workers by comparing your HITs (tasks) to low priced HITs that aren't being completed.
- Pay fairly – Don't reject an Assignment unless you're SURE it is the Worker who is wrong.
- Pay quickly – If you approve or reject Assignments once a week, Workers may do a few HITs and then wait to see if they are paid before doing more. This is especially true if you're a new Requester and haven't established your reputation yet.

These recommendations are a good starting point, but they do not address all the problems highlighted in this article.

Tasks could be subdivided (see section 3) based on if they correspond to human experiments (speech acquisition for instance) or to a real labor (such as speech transcription). Furthermore, section 2.1 lists some of the existing crowdsourcing services, the utility of which depends on the task to be accomplished. It should be said that it is easier to establish fair compensation for a given task if the chosen crowdsourcing service is adequately set up for doing so. Many crowdsourcing platforms look like a huge bazaar where tasks of different complexity, requiring workers with very different skills and therefore offering very different levels of compensation for the work done, coexist in an anarchic way. In the case of under-resourced languages transcription described in section 4, a classical framework (for example using direct contact with a local university) has produced better results than the use of a microworking crowdsourcing platform.

In order to be able to establish fair compensation, it should be clear what the tasks are in the crowdsourcing platforms. But even if the task is well defined, determining fair compensation is not only a question of ethics, but also a pragmatic question of efficiency, as it was already mentioned in section 3. As it has been pointed out in (Ipeirotis, 2011b), 'Pay enough or not pay at all'. If we want to set up an efficient sustainable framework, the only two stable solutions are:

- offer a fair reward, which could be modified in response to the quality delivered,

- do not to pay anything, as it is organized in most of the collaborative science or Game With A Purpose projects.

This quite radical assumption relies on the fact that the motivations underlying collaboration in a voluntary or a retributed work are drastically different for most people. One of the most counter-intuitive result is that providing incentives for a task with no initial payment actually *reduces* performance (Gneezy and Rustichini, 2000).

In addition to the problem of quality, completion time should also be considered. (Frei, 2009) shows a clear relation among the completion times of different tasks, depending on whether or not the task is interesting or involves payment. The conclusion is twofold: first of all, one should not ask that a tedious task be completed for free; and secondly, the level of incentives is clearly correlated to the obtention of a manageable completion time.

According to these different facts, a basic taxonomy could be defined. If the task could be performed through a traditional framework, or if some special ability is desired, a good strategy would be to attract and keep the 'good' workers. In this context, use of a microworking platform such as MTurk is not useful. It is not a problem of ethics, but mainly a problem of quality and stability. A specialized service, such as CastingWords for transcription, or a crowdsourcing service which could provide a direct link with the workers (for instance oDesk) would be preferable. Designing a Game With a Purpose or building a collaborative science project is a good alternative. The task needs to be a large-scale one, as it requires significant development time, advertising, etc, and interestingly enough; the incentive is not necessary in this context. The microworking services should only be used for tasks which do not call for high quality or special abilities but do require very rapid completion.

With project-based crowdsourcing (Simple projects or Complex tasks in the categories presented in 2.1), a requester usually hires a vendor that has access to a network of skilled professionals. The vendor is then responsible for recruiting people who can help with the work. The community can represent different categories of professionals, such as IT (information technology) experts, software developers, or CAD (Computer-Aided Design) specialists, for example. Those selected to perform the work are compensated with cash prizes or other rewards or incentives. Here a clear framework could be defined, one which resembles the classical relationship between employers and employees or client and individual contractors, as soon as the relations with the workers are clear and anonymity is discarded. In this framework, fair compensation is determined based on the classical balance between the difficulty of the task, the time spent, the amount of money available to perform it, and the negotiation with the workers. The fact that the task will be done through crowdsourcing will enable some cost reduction (streamlined recruitment and dismissal, no charge to equip the workers, etc) but should not impose a lower wage on workers. The advantage for the workers, such as self-assignment and the lack of time or money spent on commuting, should compensate for the fact that they have to pay for their own insurance and equipment. In oDesk for instance, requesters are connected with a team

of skilled workers to complete a whole job. In this sense, oDesk is close to a traditional workplace: it allows requesters to distribute a task using an hourly wage; the communication between requesters and workers is direct; a requester could provide his team with relevant training and supervision; and the wages are substantially higher than the ones available on others platforms such as MTurk (between \$10 to \$25 per hour). But, paradoxically, this method which reduces many of the inherent crowdsourcing problems in regard to the workers raises the issue that workers in oDesk are much closer to being defined as employees under the FLSA (see section 3.2) than MTurk workers. oDesk thus runs a higher risk of being taken to court by some of their 'employees'.

In contrast to project-based crowdsourcing, in most microworking services (Micro and Macro Tasks in the categories presented in 2.1), and especially in MTurk, the situation is less clear, and establishing 'fair' compensation is quite difficult. The payment should not be separated from the general economic model, and in a shaky economic model it is very difficult to determine a fair compensation. For instance, in the MTurk model it is not possible to establish a clear correlation between the reward and the final quality (see for instance (Marge et al., 2010). This fact should also be seen in light of the article (Faridani et al., 2011), in which the authors show that increasing the reward decreases the demand for the task. The reason is that high rewards mean complex tasks, with higher risks for the worker.

Section 3.3 has pointed out that setting an hourly wage is better for quality and ethics. But an hourly wage could be difficult to evaluate in a task-based environment, as there are individual variations among workers and the time spent working may decrease drastically as the workers learn how to perform the task efficiently, etc. Nevertheless, an hourly-based payment should be used (whenever possible) instead of a task-based one, as it is the common law for salaried employees in the majority of countries throughout the world. Moreover, hourly payment complies with the concept of *minimum* wage commonly found in developed countries which is not fulfilled in many crowdsourcing systems such as MTurk, which pays less than \$2/hour. Minimum wage has many ethical and practical advantages, but is quite difficult to settle in the unregulated crowdsourcing system. Minimum wage should be accompanied by regulation rules concerning both parties (requester and worker); the consequences of an minimum hourly wage without other regulations will be (among others):

- It is quite difficult to verify remotely how long a worker is actually working on a task. Online regulation tools should be designed to enable this verification while respecting privacy concerns.

- It will increase the 'Market for lemon' effect (see section 3.3) by overpaying poor quality workers and spammers. The relations between requesters and workers should be symmetrical, without anonymity and with an efficient reputation system.

- Minimum wages are country-specific, while crowd labor is spread across many countries in the crowdsourcing global marketplace. Based on classi-

cal laws of supply-and-demand, defining a minimum wage will effectively orient the work towards the places with the lowest minimum wage (see for instance the growth in the number of Indian workers in MTurk). The minimum wage should be set in order to encourage the participation of workers from countries with higher minimum wages in order to preserve local work in these countries. One possible solution is to fix the reward according to the worker's country and to impose (by law, by social pressure, or through a quality label) on requesters a quota of unit tasks to be performed by workers from his own country; this quota could be adjusted based on the task to be performed (for instance if the task could not be completed in his country, because of the language involved), or on the existing 'ethical' foundation of a crowdsourcing platform (for instance SamaSource).[41]

5.2 Impact of crowdsourcing on the ecology of linguistic resources

What are the possible consequences of collecting, transcribing or annotating speech with the help of the crowdsourcing services in their current state, given the ethical, economic, and legal problems related to these services?

For some of these crowdsourcing services, the future is insecure, given the flaws in their economic model (see section 3.3). For most of them, national or international regulation of labor laws on the internet may be foreseeable if the quantity of existing jobs outsourced on the internet is sufficiently large enough to exert pressure on political decision-makers (see 3.2). Until then, relying entirely on paid crowdsourcing services for the development of speech and language resources seems hazardous.

Beyond the present facts, some other problems could be considered foreseeable longer-term consequences of the use of crowdsourcing for language resources development. The main problem is derived from the fact that many researchers present the very low cost of crowdsourcing as its main advantage. If the Language and Resource community persists in claiming that with crowdsourcing it is now possible to produce any linguistic resource or perform any manual evaluation at a very low cost, funding agencies will come to expect just that. One can predict that in assessing projects involving language resource production or evaluation, funding agencies will prefer projects which propose to produce 10 or 100 times more data for the same amount of money. Costs such as the ones proposed in MTurk 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. The very low costs (available sometimes at the price of unreliable quality) would create a *de facto* standard for the development of language resources detrimental to other development methods.

[41] `http://samasource.org/`

5.3 Defining an ethical framework: some solutions

The situation As is the case for many other implications of Information and Communication Technologies (ICT) (Mariani et al., 2009), it is worth conducting a study on the ethical dimension of crowdsourcing, with ethics here meaning 'the way to live well together', and following a precautionary principle: potentially harmful uses should be discouraged and beneficial uses should be encouraged (Rashid et al., 2009). And just as with many other consequences of ICT development, the population is faced with the problem once it has been largely deployed at the international level and has become a matter of concern even for the professionals in computer technology who created the problem (Albright, 2009). Many researchers working in language science and technology still only see the positive aspects of crowdsourcing without apprehending the negative ones, and most papers on crowdsourcing simply ignore the ethical aspects. Large professional organizations such as the Institute for Electrical and Electronics Engineers (IEEE) (Rashid et al., 2009) and the Association for Computing Machinery (ACM) (Bederson and Quinn, 2011) recently published papers warning the community about those ethical issues.

The scientific community working in the area of language resources (Adda and Mariani, 2010), as well as the one working on speech processing (Mariani, 2011) or language processing (Fort et al., 2011), identified this problem and conducted discussions on the ethical dimension of crowdsourcing through conferences, journals or forums. One researcher said she preferred using a machine to using a human crowd for evaluating a spoken dialog system, even if the human crowd may provide better results, because of the ethical problem attached to crowdsourcing (Scheffler et al., 2011). Another researcher remarked that crowdsourcing is in fact a way to identify specialists who were not known beforehand, and that this search for specialists came with its own costs.[42] The conclusion of a panel on crowdsourcing at the International World Wide Web WWW2011 conference revealed a similar orientation, stating that crowdsourcing is best for 'parallel, scalable, automatic interviews' and for quickly finding good workers, as reported by Panos Ipeirotis (Ipeirotis, 2011c). While domain independent crowdsourcing companies such as CrowdFlower or Amazon Mechanical Turk gather a taskforce of about 1 million workers, a more specialized company like TopCoder also has a community of 400,000 specialized software engineers and computer scientists from more than 200 countries who develop software following a rigorous, standards-based methodology (Rashid et al., 2009).

Let us therefore consider the positive aspects of crowdsourcing, and explore how to encourage those positive aspects while avoiding the negative ones.

Towards collaborative solutions Requesters should take into account principles of ethics when planning to use a crowdsourcing approach in the area of Language Resources (LR) and Language Technologies (LT). B.B. Bederson and A.J. Quinn (Bederson and Quinn, 2011) provide appropriate guidelines for re-

[42] Personal communication Karen Fort, April 2012

questers using a platform in the design of a crowdsourcing operation that can be summarized as follows:

Requester design guidelines

(i) Hourly pay: Price tasks based on time. The time to do tasks can be estimated in-house before posting tasks.

(ii) Pay disclosure: Disclose the expected hourly wage.

(iii) Value worker's time: Optimize tasks to use worker's time effectively.

(iv) Objective quality metrics: Decide to approve or reject work based on objective metrics that have been defined in advance and disclosed to workers.

(v) Immediate Quality feedback: Give immediate feedback to workers, showing whatever metrics are available.

(vi) Longer-term feedback: Give warnings to problematic workers.

(vii) Disclose payment terms: Disclose in advance when payment will be made.

(viii) Follow payment terms: Pay as promptly as possible, and always within the disclosed time-frame.

(ix) Provide task context: Given the risk of doing objectionable work, tasks should be described in the context of why the work is being done.

System design guidelines

(i) Limit anonymity: Anonymity of requesters enables them to reject good work with near impunity. It also enables them to post unethical or illegal tasks with no public scrutiny. Anonymity for workers enables them to engage in large-scale cheating with nearly no risk since, as with requesters, if their reputation gets damaged, they can simply create a new account.

(ii) Provide grievance process: Provide a fair means for workers to request a review of work that was rejected.

It appears from the discussions within the scientific community that it is difficult for the researcher alone to determine the ethical way to use crowdsourcing. Scientific associations, specifically in the area of Language Resources, such as the European Language Resource Association (ELRA),[43] or in areas related to speech and language processing, such as the International Speech Communication Association (ISCA),[44] or the Association for Computational Linguistics

[43] `http://www.elra.info`
[44] `http://www.isca-speech.org`

(ACL),[45] are expected to play a role in promoting and ensuring the ethical dimension of language resources production and distribution in general, and of the use of crowdsourcing in particular.

Here is a list of tasks that those associations could take into consideration to that effect:

(i) **Promote an Open Data approach to Language Resources in general and build up the Language Resources ecosystem overall**

- Promote a Data Sharing approach to the scientific community, encouraging all to share the resources they have developed for conducting research in order to allow others to verify the results, especially when the production of resources has been fully or partially supported by public funding.
- Attribute a Persistent and Unique Identifier (LRID) to a Language Resource in order to facilitate its identification, use and tracking, in cooperation with all parties that are concerned worldwide.
- Compute a Language Resource Impact Factor (LRIF) in order to recognize the merits of LR producers.
- Attach a tag to a Language Resource that will accompany that resource for life by listing the contributors who participated in various aspects of its creation and improvement (design, specification, methodology, production, transcription, translation, annotation of various natures, validation, correction, addition, etc.).
- Assign a copyright status to a LR based on *Creative Commons* (CC) categories, or the like.

(ii) **Promote the Ethical dimension of crowdsourcing**

- Make the community aware of the ethical dimension of crowdsourcing.

(iii) **Provide information in order to observe an Ethical approach for crowdsourcing**

- Identify the requirements of the community, in terms of resources (data, tools, services) and of economic (price, payment), administrative (ordering, licensing) and ethical issues attached to resource production.
- Provide advice concerning the best approach to producing a specific resource (using a crowd, a set of specialists, automatic or semi-automatic systems, or a mixture of them).
- Identify the crowdsourcing platforms that exist and provide a description for each of them, including the pricing policy, the way it deals with the ethical aspects and the constraints of use.

[45] `http://www.aclweb.org`

- Provide information about the magnitude of the efforts (time spent, including time to learn the task...) attached to various kinds of resource production and give an estimate of the corresponding salaries, with the aim of defining a fee schedule.

(iv) **Provide tools and services to facilitate and follow an Ethical approach for crowdsourcing**

- Provide a platform for producing the data:
  - Either its own platform, possibly involving a network of specialists and complying with ethics (Fair Trade principles, minimum guarantee of wages, auto-approval delay, etc.). But this approach may not be convenient for communicating with a large community of speakers,
  - Or third-party generic crowdsourcing platforms addressing a large set of non-specialists workers, either fully generic or tailored to a given purpose, after checking the ethical dimension of those platforms.
- Standardize simple tasks in order to facilitate reusability, trading commodities and true market pricing (Ipeirotis, 2012c).
- Constitute and maintain a network of specialists for many different languages all over the world.
- Pre-qualify workers through ability tests, such as those concerning their proficiency in different languages
  - Establish the means to conduct tests on worker ability through Gold Standard data and ground truth.
  - Establish the means to provide public information about the reputation of workers and requesters.
- Help to define a network of local contacts for resource-needed languages; those local contacts might be non-profit organizations that supervise the data annotation for one or several languages of a particular area and remunerate the workers in keeping with a minimum wage.

(v) **Provide recommendations and validation for Ethical approaches in crowdsourcing**

- Write and distribute a Charter for the ethical production of resources.
- Write and distribute Best Practices and guidelines for the ethical production of resources.
- Attribute an 'Ethically Produced' label to Language Resources which have been produced in an ethical way:
  - Such resources should be produced within the parameters of legal and ethical working conditions (Fair Trade principles, careful

pricing of the tasks, minimum guarantee of wages, maximum number of working hours, compliance with the tax regulations, etc.) and should come with quality insurance (in terms of the technical quality of the resource, as well as of compliance with legal regulations (Intellectual Property Rights, privacy, etc.)).

- Act so that third-party generic crowdsourcing platforms follow an ethical approach.
- Attribute an 'Ethically Resource Producer' label to such platforms.

6 Conclusion

As do many other topics, crowdsourcing can be considered from two angles.

On the positive side:

- it allows to decrease the price of producing resources,
- it may therefore increase the size of the data,
- it allows one to address a large quantity and diversity of people,
- it facilitates access to people who would be difficult to reach in other ways,
- it establishes a direct link between the employers and the workers,
- it offers a salary for those who have none,
- it bypasses intermediaries.

It therefore seems to be an especially attractive option for **Less-Resourced Languages** because it is less costly, given that investments may be difficult to procure for economically uninteresting languages for many reasons, including: there may be fewer experts on those languages who could intervene; access to native speakers of those languages who are abroad or who were part of a diaspora may present different difficulties or have to be conducted via intermediaries at a certain cost; the number of those native speakers may be low; and finding financial support is complicated by difficult economic conditions. However, as it has been shown in the related experience of using crowdsourcing for producing annotated corpus in the less-resourced languages of Swahili and Amharic, the reality may be somewhat different and, in some cases, may result in shifting back to a more traditional approach.

On the negative side:

- it doesn't guarantee a proper salary for the workers,
- it doesn't guarantee the quality of the result,
- it may ultimately result in a more significant cost than traditional approaches,

- it bypasses all legal aspects attached to social security, pensions, or union rights.

The basic assumptions on quality, price and motivation may therefore be discussed, as well as the legal and scientific policy dimensions.

Quality: The quality of the transcription of a speech corpus and/or of its translation should be enough to train ASR or MT systems, for example. And teams participating in evaluation campaigns are highly sensitive to the quality of the training and testing material. However, some quality problems may appear, especially if the task is complex. The task may then be subdivided into sub-tasks, but this increases the complexity of the organization, as it necessitates coordination and correlation. If people are primarily interested by the financial income, they may cheat in order to increase their productivity and thus their salary, and this also makes the quality checking mandatory. The initial detection of spammers is necessary and, in some cases, the task has to be duplicated or triplicated for cross-validation. Final validation and post-processing are also to be added.

Price: Salaries are usually rather low in crowdsourcing, and therefore the production cost is supposed to be low. However, the development of interfaces for non-experts, the detection of cheating, the spammer problem, and the quality issue necessitating the previously mentioned operations add extra costs. If the competence is hard to find, the salaries will also have to be higher.

Motivation: Some crowdsourcing actions, such as Wikipedia, are based on voluntary contributions, but most are conducted for money. Therefore, the workers may only be interested in the salary and not in the task. This may provoke those workers to consider efficiency first and to try to earn the maximum of money with the minimum of efforts.

Legal: The legal dimension must also be taken into account (Wolfson and Lease, 2011). The employers may not pay taxes for the employees, while the employees may also not be taxed, as the action is conducted in an international framework which may escape national regulations. Wages are usually lower than the amount paid in the employer's country (while still sometimes being higher than the usual salary in the employee's country). There is no social or health security, no guarantee of payment and no support coming from unions (which may however be replaced on the Internet by blogs and forums such as Turker Nation[46] or Turkopticon).[47] And of course there is no protection of IPR and copyright.

Science Policy: Given that the use of crowdsourcing reduces production costs, funding agencies may reduce their support for resource production and therefore impose crowdsourcing as a *de facto* standard.

[46] `http://www.turkernation.com/`
[47] `http://turkopticon.differenceengines.com`

References

S.T. Abate, W. Menzel, and B. Tafila. An amharic speech corpus for large vocabulary continuous speech recognition. In *Interspeech*, pages 67–76, Lisbon, 2005.

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

Gilles Adda, B. Sagot, K. Fort, and Joseph-Jean Mariani. Crowdsourcing for language resource development: critical analysis of amazon mechanical turk overpowering use. In *Language & Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics (L&TC 2011)*, pages 304–308, Poznan, Poland, 25/11 au 27/11 2011.

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

Peggy Albright. Is crowdsourcing an opportunity or threat? `http://www.computer.org/portal/web/buildyourcareer/crowdsourcing`, March 2009.

Charles Arthur. What is the 1% rule? The Guardian, July 2006.

Benjamin B. Bederson and Alexander J. Quinn. Web workers unite! addressing challenges of online laborers. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*, CHI EA '11, pages 97–106. ACM, 2011. ISBN 978-1-4503-0268-5. doi: http://doi.acm.org/10.1145/1979602.1979606. URL `http://doi.acm.org/10.1145/1979602.1979606`.

V. Berment. *Méthodes pour informatiser les langues et les groupes de langues "peu dotées"*. PhD thesis, Université Joseph Fourier, Grenoble, 2004.

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

Chris Callison-Burch and Mark Dredze. 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, 2010.

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

David L. Chen and William B. Dolan. Building a persistent workforce on mechanical turk for multilingual data collection. In *Proceedings of The 3rd Human Computation Workshop (HCOMP 2011)*, August 2011.

Sharon Chiarella. Cooking with Sharon. http://mechanicalturk.typepad.com/blog/2011/07/cooking-with-sharon-tip-3-manage-your-reputation.html, 2011.

Lydia B. Chilton, John J. Horton, Robert C. Miller, and Shiri Azenkot. Task search in a human computation market. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 1–9, 2010. ISBN 978-1-4503-0222-7.

Chr. Draxler. WWWTranscribe – a Modular Transcription System Based on the World Wide Web. In *Proc. Eurospeech*, pages 1691–1694, Rhodes, 1997.

Chr. Draxler and A. Steffen. Ph@ttsessionz: Recording 1000 adolescent speakers in schools in germany. In *Proc. Interspeech*, pages 1597–1600, Lisbon, 2005.

Maxine Eskenazi, Gina-Anne Levow, Helen Meng, Gabriel Parent, and David Suendermann, editors. *Crowdsourcing for Speech Processing*. Wiley, 2013. ISBN 978-1-118-35869-6, 978-1-118-54127-2.

Siamak Faridani, Björn Hartmann, and Panagiotis G. Ipeirotis. WhatâĂŹs the right price? pricing tasks for finishing on time. In *Proceedings of The 3rd Human Computation Workshop (HCOMP 2011)*, August 2011.

Alek Felstiner. Working the Crowd: Employment and Labor Law in the Crowdsourcing Industry. *Berkeley Journal of Employment and Labor Law*, 32(1), August 2011.

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

Brent Frei. Paid Crowdsourcing: Current State & Progress towards Mainstream Business Use. Smartsheet White Paper. http://www.smartsheet.com/files/haymaker/PaidCrowdsourcingSept2009-ReleaseVersion-Smartsheet.pdf, September 2009.

H. Gelas, L. Besacier, and F. Pellegrino. Developments of swahili resources for an automatic speech recognition system. In *SLTU*, 2012.

Uri Gneezy and Aldo Rustichini. Pay enough or don't pay at all. *Quarterly Journal of Economics*, 115(3):791–810, 2000. URL http://www.mitpressjournals.org/doi/abs/10.1162/003355300554917.

Er Gruenstein, Ian Mcgraw, and Andrew Sutherl. A self-transcribing speech corpus: collecting continuous speech with an online educational game. In *the Speech and Language Technology in Education (SLaTE) Workshop*, Warwickshire, England, September 2009.

Bengt Holmstrom. Moral hazard and observability. *Bell Journal of Economics*, 10(1):74–91, 1979.

Panos Ipeirotis. Why people participate on mechanical turk, now tabulated. http://www.behind-the-enemy-lines.com/2008/09/why-people-participate-on-mechanical.html, September 2008.

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

Panos Ipeirotis. A plea to amazon: Fix mechanical turk! http://behind-the-enemy-lines.blogspot.com/2010/10/plea-to-amazon-fix-mechanical-turk.html, October 2010b.

Panos Ipeirotis. Be a top mechanical turk worker: You need $5 and 5 minutes. http://www.behind-the-enemy-lines.com/2010/10/be-top-mechanical-turk-worker-you-need.html, October 2010c.

Panos Ipeirotis. Do mechanical turk workers lie about their location? http://www.behind-the-enemy-lines.com/2011/03/do-mechanical-turk-workers-lie-about.html, March 2011a.

Panos Ipeirotis. Pay enough or don't pay at all. http://www.behind-the-enemy-lines.com/2011/05/pay-enough-or-dont-pay-at-all.html, May 2011b.

Panos Ipeirotis. Does lack of reputation help the crowdsourcing industry? http://www.behind-the-enemy-lines.com/2011/11/does-lack-of-reputation-help.html, November 2011c.

Panos Ipeirotis. Mechanical turk vs odesk: My experiences. http://www.behind-the-enemy-lines.com/2012/02/mturk-vs-odesk-my-experiences.html, February 2012a.

Panos Ipeirotis. Philippines: The country that never sleeps. http://www.behind-the-enemy-lines.com/2012/04/when-is-world-working-odesk-edition-or.html, April 2012b.

Panos Ipeirotis. The need for standardization in crowdsourcing. http://www.behind-the-enemy-lines.com/2012/02/need-for-standardization-in.html, February 2012c.

Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E. Kraut. Crowdforge: crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, UIST '11, pages 43–52, 2011. ISBN 978-1-4503-0716-1.

S. Kochhar, S. Mazzocchi, and P. Paritosh. 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., 2010.

V.B. Le and L. Besacier. Automatic speech recognition for under-resourced languages: application to vietnamese language. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(8):1471–1482, 2009.

Greg Little, Lydia B. Chilton, Max Goldman, and Robert C. Miller. Turkit: human computation algorithms on mechanical turk. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, UIST '10, pages 57–66, 2010. ISBN 978-1-4503-0271-5.

Matthew Marge, Satanjeev Banerjee, and Alexander I. Rudnicky. 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 2010.

Joseph Mariani. Ethical dimension of crowdsourcing. In *Special session on Crowdsourcing, Interspeech'2011*, Firenze, August 2011.

Joseph Mariani, Jean-Michel Besnier, Jacques Bordé, Jean-Michel Cornu, Marie Farge, Jean-Gabriel Ganascia, Jean-Paul Haton, and Evelyne Serverin. Pour une éthique de la recherche en Sciences et Technologies de l'Information et de la Communication (STIC). Technical report, COMETS-CNRS, `http://www.cnrs.fr/fr/organisme/ethique/comets/avis.htm`, November 2009.

Lutz Marten. Swahili. In Keith Brown, editor, *The Encyclopedia of Languages and Linguistics, 2nd ed.*, volume 12, pages 304–308. Oxford: Elsevier, 2006.

Ian McGraw, Alexander Gruenstein, and Andrew Sutherland. A self-labeling speech corpus: Collecting spoken words with an online educational game. In *Proceedings of Interspeech*, pages 3031–3034, Brighton, UK, 6-10 September 2009.

Scott Novotney and Chris Callison-Burch. 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, 2010. ISBN 1-932432-65-5.

Gabriel Parent and Maxine Eskenazi. Toward better crowdsourced transcription: Transcription of a year of the let's go bus information system data. In *Proceedings of IEEE Workshop on Spoken Language Technology*, pages 312–317, Berkeley, California, December 2010.

Alexander J. Quinn and Benjamin B. Bederson. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI EA '11, pages 1403–1412, 2011.

A. Rashid, J. Weckers, and E. Lucas. Software engineering, ethics in a digital world. In *IEEE Computing Now*. `http://www2.computer.org/portal/web/computingnow/0709/theme/co1_softengethics`, June 2009.

Joel Ross, Andrew Zaldivar, Lilly Irani, and Bill Tomlinson. 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`, 2009.

Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 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, 2010. ACM. ISBN 978-1-60558-930-5. doi: http://doi.acm.org/10.1145/1753846.1753873.

Tatjana Scheffler, Roland Roller, and Norbert Reithinger. Speecheval: A domain-independent user simulation platform for spoken dialog system evaluation. In *IWSDS 2011*, Granada, September 2011.

Tanja Schultz and K. Kirchhoff, editors. *Multilingual Speech Processing*. Berlington, MA, USA: Academic Press, 2006.

M. Six Silberman, Joel Ross, Lilly Irani, and Bill Tomlinson. Sellers' problems in human computation markets. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 18–21, 2010. ISBN 978-1-4503-0222-7.

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

Virginia Commonwealth University. VCU Institutional Review Board Written Policies and Procedures. `http://www.research.vcu.edu/irb/wpp/flash/XVII-2.htm`, November 2009.

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

F. Wessel and H. Ney. Unsupervised training of acoustic models for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13(1):23–31, 2005.

Stephen M. Wolfson and Matthew Lease. Look before you leap: Legal pitfalls of crowdsourcing. *Proceedings of the American Society for Information Science and Technology*, 48(1):1–10, 2011. ISSN 1550-8390. doi: 10.1002/meet.2011.14504801135. URL `http://dx.doi.org/10.1002/meet.2011.14504801135`.

Jonathan Zittrain. Ubiquitous human computing. *Phil. Trans. R. Soc. A 28*, 366(1881):3813–3821, October 2008a.

Jonathan Zittrain. *The Future of the Internet–And How to Stop It*. Yale University Press, New Haven, CT, USA, 2008b. ISBN 0300124872, 9780300124873.