The Language Demographics of Amazon Mechanical Turk

Abstract

We present a large scale study of the languages spoken by bilingual workers on Mechanical Turk (MTurk). We establish a methodology for determining the language skills of anonymous crowd workers that is more robust than simple surveying. We validate workers' self-reported language skill claims by measuring their ability to correctly translate words, and by geolocating workers to see if they reside in countries where the languages are likely to be spoken. Rather than posting a one-off survey, we posted paid tasks consisting of 1,000 assignments to translate a total of 10,000 words in each of 100 languages. Our study ran for several months, and was highly visible on the MTurk crowdsourcing platform, increasing the chances that bilingual workers would complete it. Our study was useful both to create bilingual dictionaries and to act as census of the bilingual speakers on MTurk. We use this data to recommend languages with the largest speaker populations as good candidates for other researchers who want to develop crowdsourced, multilingual technologies.

1 Overview

Crowdsourcing is a promising new mechanism for collecting data for natural language processing research. Access to a fast, cheap, and flexible workforce allows us to collect new types of data, potentially enabling new language technologies. Because crowdsourcing platforms like Amazon Mechanical Turk (MTurk) give researchers access to a worldwide workforce, one obvious application of crowdsourcing is the creation of multilingual technologies. With an increasing number of active crowd workers located outside of the United States, there is even the potential to reach fluent speakers of lower resource languages. In this paper, we investigate the feasibility of hiring language informants on MTurk by conducting the first large-scale demographic study of the languages spoken by workers on the platform.

There are several complicating factors when trying to take a census of workers on MTurk. The workers' identities are anonymized, and Amazon provides no information about their countries of origin or their language abilities. Posting a simple survey to have workers report this information may be inadequate, since (a) many workers may never see the survey, (b) many opt not to do one-off surveys since potential payment is low, and (c) validating the answers of respondents is not straightforward.

Our study establishes a methodology for determining the language demographics of anonymous crowd workers that is more robust than simple surveying. We ask workers what languages they speak and what country they live in, and validate their claims by measuring their ability to correctly translate words and by recording their geolocation. To increase the visibility and the desirability of our tasks, we post 1,000 assignments in each of 100 languages. These tasks each consist of translating 10 foreign words into English. Two of the 10 words have known translations, allowing us to validate that the workers' translations are accurate. We construct bilingual dictionaries with up to 10,000 entries, with the majority of entries being new.

Surveying thousands of workers allows us to analyze current speaker populations for 100 languages. The data also allows us to answer questions like: How quickly is work completed in a given language? Are crowdsourced translations reliably good? How often do workers misrepresent their language abilities to obtain financial rewards?

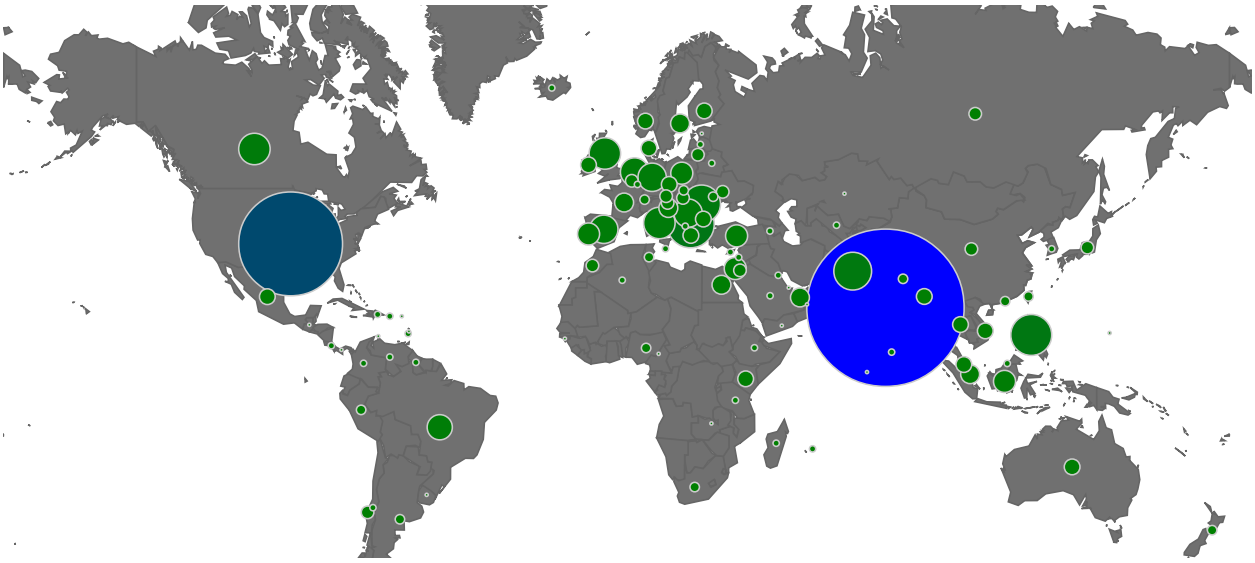


Figure 1: The number of workers per country. This map was generated based on geolocating the IP address of 4,983 workers in our study. Omitted are 60 workers who were located in more than one country during the study, and 238 workers who could not be geolocated. The size of the circles represents the number of workers from each country. The two largest are India (1,998 workers) and the United States (866). To calibrate the sizes: the Philippines has 142 workers, Egypt has 25, Russia has 10, and Sri Lanka has 4.

2 Background and Related Work

Amazon's Mechanical Turk (MTurk) is an online marketplace for work that gives employers and researchers access to a large, low-cost, workforce. MTurk allows employers to provide micropayments in return for workers completing microtasks. The basic units of work on MTurk are called 'Human Intelligence Tasks' (HITs). MTurk was designed to accommodate tasks that are difficult for computers, but simple for people. This facilitates research into human computation, where people can be treated as a function call (von Ahn, 2005; Little et al., 2009; Quinn and Bederson, 2011). It has application to research areas like human-computer interaction (Bigham et al., 2010; Bernstein et al., 2010), computer vision (Sorokin and Forsyth, 2008; Deng et al., 2010; Rashtchian et al., 2010), speech processing (Marge et al., 2010; Lane et al., 2010; Parent and Eskenazi, 2011; Eskenazi et al., 2013), and natural language processing (Snow et al., 2008; Callison-Burch and Dredze, 2010; Laws et al., 2011).

On MTurk, researchers who need work completed are called 'Requesters', and workers are often referred to as 'Turkers'. MTurk is a true market, meaning that Turkers are free to choose to complete the HITs which interest them, and Requesters can price their tasks competitively to try to attract workers and have their tasks done quickly (Faridani et al., 2011; Singer and Mittal, 2011). Turkers remain anonymous to Requesters, and all payment occurs through Amazon. Requesters are able to accept submitted work or reject work that does not meet their standards. Turkers are only paid if a Requester accepts their work.

Several reports examine Mechanical Turk as an economic market (Ipeirotis, 2010a; Lehdonvirta and Ernkvist, 2011). When Amazon introduced MTurk, it first offered payment only in Amazon credits, and later offered direct payment in US dollars. More recently, it has expanded to include one foreign currency, the Indian rupee. Despite its payments being limited to two currencies or Amazon credits, MTurk claims over half a million workers from 190 countries (Amazon, 2013). This suggests that its worker population should represent a diverse set of languages.

A demographic study by Ipeirotis (2010b) focused on age, gender, martial status, income levels, motivation for working on MTurk, and whether workers used it as a primary or supplemental form

of income. The study contrasted Indian and US workers. Ross et al. (2010) completed a longitudinal follow-on study. A number of other studies have informally investigated Turkers' language abilities. Munro and Tily (2011) compiled survey responses of 2,000 Turkers, revealing that four of the six most represented languages come from India (the top six being Hindi, Malayalam, Tamil, Spanish, French, and Telugu). Irvine and Klementiev (2010) had Turkers evaluate the accuracy of translations that had been automatically inducted from monolingual texts. They examined translations of 100 words in 42 low-resource languages, and reported geolocated countries for their workers (India, the US, Romania, Pakistan, Macedonia, Latvia, Bangladesh and the Philippines). Irvine and Klementiev discussed the difficulty of quality control and assessing the plausibility of workers' language skills for rare languages, which we address in this paper.

Several researchers have investigated using MTurk to build bilingual parallel corpora for machine translation, a task which stands to benefit low cost, high volume translation on demand (Germann, 2001). Ambati et al. (2010) conducted a pilot study by posting 25 sentences to MTurk for Spanish, Chinese, Hindi, Telugu, Urdu, and Haitian Creole. In a study of 2000 Urdu sentences, Zaidan and Callison-Burch (2011) presented methods for achieving professional-level translation quality from Turkers by soliciting multiple English translations of each foreign sentence. Zbib et al. (2012) used crowdsourcing to construct a 1.5 million word parallel corpus of dialect Arabic and English, training a statistical machine translation system that produced higher quality translations of dialect Arabic than a system a trained on 100 times more Modern Standard Arabic-English parallel data.

Several researchers have examined cost optimization using active learning techniques to select the most useful sentences or fragments to translate (Ambati and Vogel, 2010; Bloodgood and Callison-Burch, 2010; Ambati, 2012).

To contrast our research with previous work, the main contributions of this paper are: (1) a robust methodology for assessing the bilingual skills of anonymous workers, (2) the largest-scale census to date of language skills of workers on MTurk, and (3) a detailed analysis of the data gathered in our study.

English	689	Tamil	253	Malayalam	219
Hindi	149	Spanish	131	Telugu	87
Chinese	86	Romanian	85	Portuguese	82
Arabic	74	Kannada	72	German	66
French	63	Polish	61	Urdu	56
Tagalog	54	Marathi	48	Russian	44
Italian	43	Bengali	41	Gujarati	39
Hebrew	38	Dutch	37	Turkish	35
Vietnamese	34	Macedonian	31	Cebuano	29
Swedish	26	Bulgarian	25	Swahili	23
Hungarian	23	Catalan	22	Thai	22
Lithuanian	21	Punjabi	21	Others	≤ 20

Table 1: Self-reported native language of 3,216 bilingual Turkers. Not shown are 49 languages with $\leq$20 speakers. We omit 1,801 Turkers who did not report their native language, 243 who reported 2 native languages, and 83 with $\geq$3 native languages.

3 Experimental Design

The central task in this study was to investigate Mechanical Turk's bilingual population. We accomplished this through self-reported surveys combined with a HIT to translate individual words for 100 languages. We evaluate the accuracy of the workers' translations against known translations. In cases where these were not exact matches, we used a second pass monolingual HIT, which asked English speakers to evaluate if a worker-provided translation was a synonym of the known translation.

Demographic questionnaire At the start of each HIT, Turkers were asked to complete a brief survey about their language abilities. The survey asked the following questions:

- Is [language] your native language?
- How many years have you spoken [language]?
- Is English your native language?
- How many years have you spoken English?
- What country do you live in?

We automatically collected each worker's current location by geolocating their IP address. A total of 5,281 unique workers completed our HITs. Of these, 3,625 provided answers to our survey questions, and we were able to geolocate 5,043. Figure 1 plots the location of workers across 106 countries. Table

1 gives the most common self-reported native languages.

Selection of languages We drew our data from the different language versions of Wikipedia. We selected the 100 languages with the largest number of articles,[1] (Table 2). For each language, we chose the 1,000 most viewed articles over a 1 year period,[2] and extracted the 10,000 most frequent words from them. The resulting vocabularies served as the input to our translation HIT.

500K+ ARTICLES: German (de), English (en), Spanish (es), French (fr), Italian (it), Japanese (ja), Dutch (nl), Polish (pl), Portuguese (pt), Russian (ru)
100K-500K ARTICLES: Arabic (ar), Bulgarian (bg), Catalan (ca), Czech (cs), Danish (da), Esperanto (eo), Basque (eu), Persian (fa), Finnish (fi), Hebrew (he), Hindi (hi), Croatian (hr), Hungarian (hu), Indonesian (id), Korean (ko), Lithuanian (lt), Malay (ms), Norwegian (Bokmal) (no), Romanian (ro), Slovak (sk), Slovenian (sl), Serbian (sr), Swedish (sv), Turkish (tr), Ukrainian (uk), Vietnamese (vi), Waray-Waray (war), Chinese (zh)
10K-100K ARTICLES: Afrikaans (af) Amharic (am) Asturian (ast) Azerbaijani (az) Belarusian (be) Bengali (bn) Bishnupriya Manipuri (bpy) Breton (br) Bosnian (bs) Cebuano (ceb) Welsh (cy) Zazaki (diq) Greek (el) West Frisian (fy) Irish (ga) Galician (gl) Gujarati (gu) Haitian (ht) Armenian (hy) Icelandic (is) Javanese (jv) Georgian (ka) Kannada (kn) Kurdish (ku) Luxembourgish (lb) Latvian (lv) Malagasy (mg) Macedonian (mk) Malayalam (ml) Marathi (mr) Neapolitan (nap) Low Saxon (nds) Nepali (ne) Newar / Nepal Bhasa (new) Norwegian (Nynorsk) (nn) Piedmontese (pms) Sicilian (scn) Serbo-Croatian (sh) Albanian (sq) Sundanese (su) Swahili (sw) Tamil (ta) Telugu (te) Thai (th) Tagalog (tl) Urdu (ur) Yoruba (yo)
<10K ARTICLES: Central Bicolano (bcl) Tibetan (bo) Ilokano (ilo) Punjabi (pa) Kapampangan (pam) Pashto (ps) Sindhi (sd) Somali (so) Uzbek (uz) Wolof (wo)

Table 2: A list of the languages that were used in our study, grouped by the number of Wikipedia articles in the language. Each language's code is given in parentheses. These language codes are used in other figures throughout this paper.

Translation HIT For the translation task, we asked Turkers to translate individual words. We showed each word in the context of three sentences that were drawn from Wikipedia. Turkers were allowed to mark that they were unable to translate a word. Each task contained 10 words, 8 of which were words with unknown translations, and 2 of which were quality control words with known translations. We gave special instruction for translating names of people and places, giving examples of how to handle 'Barack Obama' and 'Australia' using their interlanguage links. For languages with non-Latin alphabets, names were transliterated.

The task paid $0.15 for the translation of 10 words. Each set of 10 words was independently translated by three separate workers. 5,281 workers completed 256,604 translations assignments, totaling more than 3 million words, over a period of three and a half months.

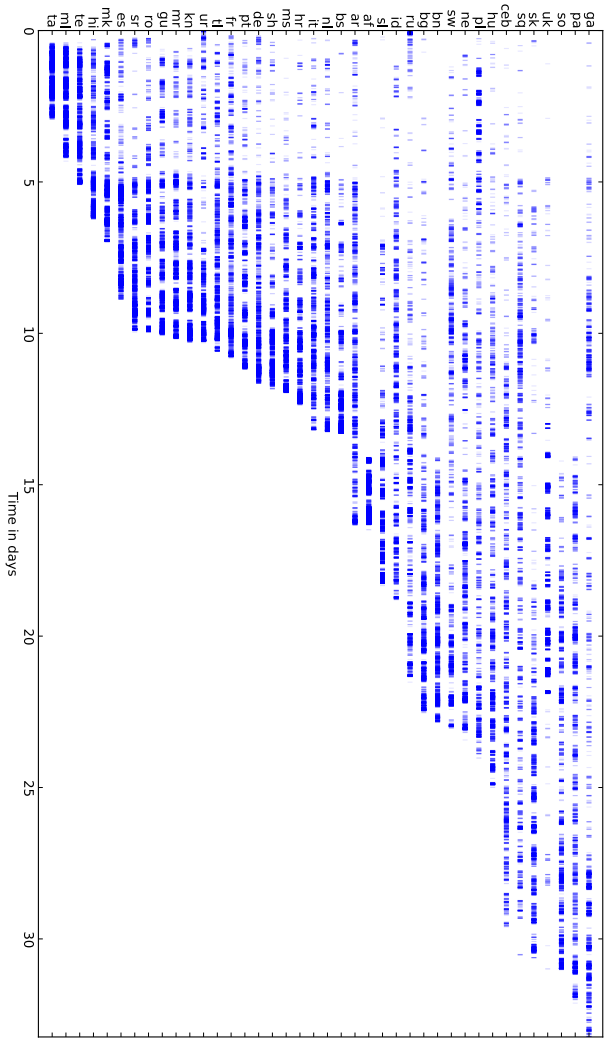


Figure 2: Days to complete the translation HITs for 40 of the languages. Tick marks represent the completion of individual assignments.

[1] http://meta.wikimedia.org/wiki/List_of_Wikipedias

[2] http://dumps.wikimedia.org/other/pagecounts-raw/

Gold standard translations A set of gold standard translations were automatically harvested from Wikipedia for every language to use as embedded controls. We used Wikipedia's inter-language links to pair titles of English articles with their corresponding foreign article's title. To get a more translatable set of pairs, we excluded any pairs where: (1) the English word was not present in the WordNet ontology (Miller, 1995), (2) either article title was longer than a single word, (3) the English wikipedia page was a subcategory of person or place, or (4) the English and the foreign titles were identical or a substring of the other.

Manual evaluation of non-identical translations We counted all translations that exactly matched the gold standard translation as correct. For non-exact matches we created a second-pass quality assurance HIT. Turkers were shown a pair of English words, one of which was a Turker's translation of the foreign word used for quality control, and the other of which was the gold-standard translation of the foreign word. Evaluators were asked whether the two words had the same meaning, and chose between three answers: 'Yes', 'No', or 'Related but not synonymous.' Examples of meaning equivalent pairs include: <*petroglyphs, rock paintings*>, <*demo, show*> and <*loam, loam: soil rich in decaying matter*>. Non-meaning equivalents included: <*assorted, minutes*>, and <*major,* URL of image>. Related items were things like <*sky, clouds*>. Misspellings like <*lactation, lactiation* > were judged to have same meaning, and were marked as misspelled. Three separate Turkers judged each pair, allowing majority votes for difficult cases.

We checked Turkers who were working on this task by embedding pairs of words which were either known to be synonyms (drawn from WordNet) or unrelated (randomly chosen from a corpus). Automating approval/rejections for the second-pass evaluation allowed the whole pipeline to be run automatically. Caching judgments meant that we ultimately needed only 20,952 synonym tasks to judge all of the submitted translations (a total of 74,572 non-matching word pairs). These were completed by an additional 1,005 workers. Each of these assignments included 10 word pairs and paid $0.10.

Full sentence translations In addition to the bilingual dictionaries, we are interested in Turkers' ability to translate complete sentences. Because the Indian subcontinent is so well represented among our workers, we chose six Indian languages for which to gather translations of full sentences. We selected Bengali, Malyalam, Hindi, Tamil, Telugu, and Urdu as good candidates for our study : these languages have large native speaker populations and are well represented on Wikipedia but they are currently not covered by any professionally translated corpora.

Each of the sentence translation HITs contained ten sequential sentences from a source-language Wikipedia article and asked the worker to enter a free-form translation for each. We collected four translations from different translators for each source sentence. Workers were paid $0.70 per HIT.

Since there is no gold-standard data available for our chosen languages, and it could not be mined in a straight-forward way as in the single word case, we were unable to embed controls into our HITs. We instead accepted or rejected translations based on a review of each worker's submissions, which included a comparison of the translations to a monotonic gloss (produced with a dictionary), and metadata such as the amount of time the worker took to complete the HIT and their geographic location.

Figure 3 contains some hand-picked examples of the sorts of translations we obtained. While the lack of a gold-standard data set prevented us from evaluating the collected translations in terms of BLEU scores, as in Zaidan and Callison-Burch (2011), we evaluate the quality of the data by using it to train an SMT system. We present results in section 5.

4 Measuring Translation Quality

For single word translations, we calculate the quality of translations on the level of individual assignments and aggregated over workers and languages. We define an assignment's quality as the proportion of controls that are correct in a given assignment, where correct means exactly correct or judged to be synonymous.

$$\text{Quality}(a_i) = \frac{1}{k_i} \sum_{j=1}^{k_i} \delta(tr_{ij} \in \texttt{syns}[g_j]) \quad (1)$$

மார்ச் 15,2007இல் ஆக்ஸ்ஃபோர்டு ஆங்கில அகராதி யில் விக்கி இடம்பற்றது.

In March 15,2007 Wiki got a place in Oxford English dictionary.
On March 15, 2007 wiki was included in the Oxford English dictionary.
ON MARCH 15, 2007, WIKI FOUND A PLACE IN THE OXFORD ENGLISH DICTIONARY
March 15, 2007 oxford english index of wiki's place.

Figure 3: An example of the variance in translation quality for the human translations of a Tamil sentence; the formatting of the translations has been preserved exactly.

where a_i is the i^{th} assignment, k_i is the number of controls in a_i, tr_{ij} is the Turker's provided translation of control word j in assignment i, g_j is the gold standard translation of control word j, $\texttt{syns}[g_j]$ is the set of words judged to be synonymous with g_j and includes g_j, and $\delta(x)$ is Kronecker's delta and takes value 1 when x is true. Most assignments had two known words embedded, so most assignments had scores of either 0, 0.5, or 1.

Since computing overall quality for a language as the average assignment quality score is biased towards a small number of highly active Turkers, we instead report language quality scores as the average per-Turker quality, where a Turker's quality is the average quality of all the assignments that she completed:

$$\text{Quality}(t_i) = \frac{\sum_{a_j \in \texttt{assigns}[i]} \text{Quality}(a_j)}{|\ \texttt{assigns}[i]\ |} \quad (2)$$

where $\texttt{assigns}[i]$ is the assignments completed by Turker i, and Quality(a) is as above.

Quality for a language is then given by

$$\text{Quality}(l_i) = \frac{\sum_{t_j \in \texttt{turkers}[i]} \text{Quality}(t_j)}{|\ \texttt{turkers}[i]\ |} \quad (3)$$

When a Turker completed assignments in more than one language, their quality was computed separately for each language. Figure 4 shows the translation quality for languages with contributions from at least 50 workers.

Use of machine translation One obvious way for dishonest workers to shortcut is to use available online translation tools. Although we followed best practices to deter copying-and-pasting into online MT systems by rendering words and sentences as images (Zaidan and Callison-Burch, 2011), this strategy does not prevent workers from typing the words into an MT system if they are able to type in the language's script.

We quantified each Turker's overlap with the Google translations so that we could identify and remove workers that appeared to be misusing Google Translate. We used Google to translate all 10,000 words for the 51 foreign languages that Google Translate covered at the time of the study. We measured the percent of workers' translations that exactly matched the translation returned from Google.

We found that while overall overlap is high (the average Turker's overlap was 41%), it is not consistently high across all Turkers. Figure 5a shows that while a small number of Turkers have a very high amount of overlap with Google (near 80%) the majority overlap less than half of the time. It seems likely that those with very high overlap are cheating, but it is also reasonable to assume that honest Turkers will overlap with Google with some regularity. We divide the workers into three groups: those with very high overlap with Google (likely cheating by using Google to translate words), those with reasonable overlap, and those with no overlap (likely cheating by other means, e.g. submitting random text).

Our controls are designed to identify workers that fall into the third group (those who are spamming or providing useless translations), but they will not effectively flag workers who are cheating with Google Translate. We therefore remove the 500 Turkers with the highest overlap with Google. This equates to removing all workers with greater than 70% overlap. Figure 5b shows that removing workers at or above the 70% threshold retains 90% of the collected translations and over 90% of the workers.

Quality scores reported throughout the paper reflect only translations from Turkers whose overlap

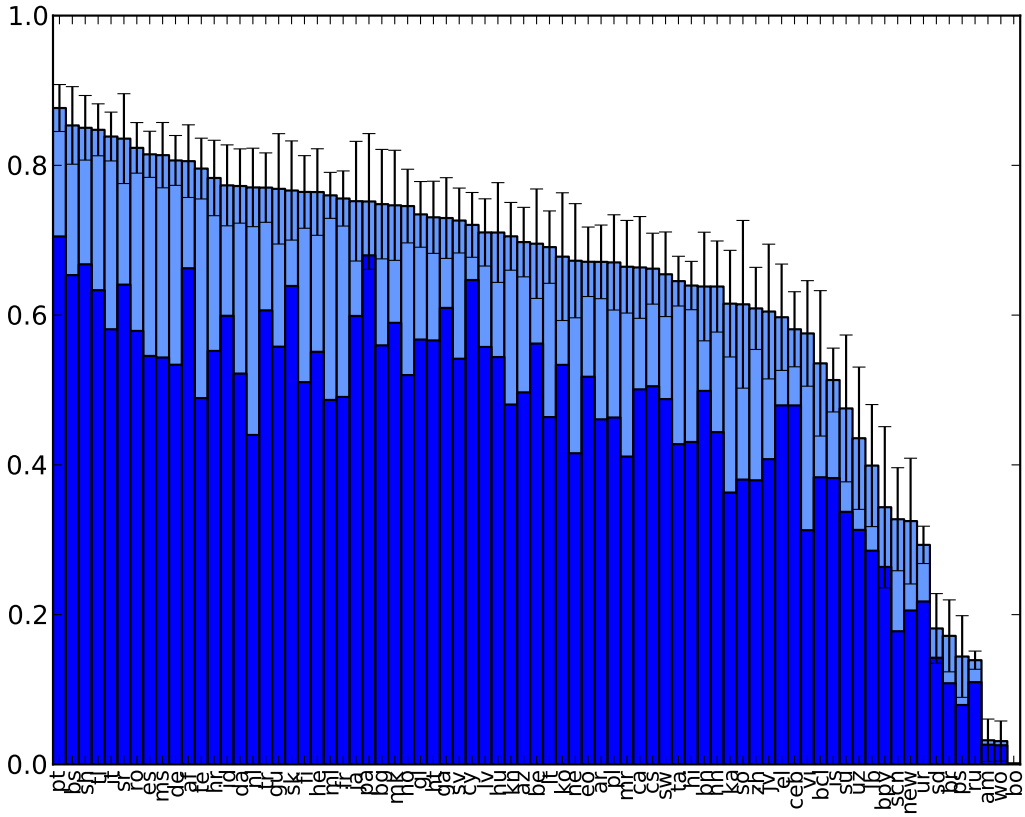


Figure 4: Translation quality for languages with at least 50 Turkers. The dark blue bars indicate the proportion of translations which exactly matched gold standard translations, and light blue indicate translations which were judged to be correct synonyms. Error bars show the 95% confidence intervals for each language.

with Google falls below this 70% threshold.

5 Data Analysis

We performed an analysis of our data to address the following questions:

- How quickly can we expect work to be completed in a particular language?
- Do workers accurately represent their language abilities? Should we constrain tasks by region?
- Are our gold standard translations valid?
- Can Turkers' translations be used to train an MT system?

Speed of completion Figure 2 gives the completion times for 40 languages. The 10 languages to finish in the shortest amount of time were: Tamil, Malayalam, Telugu, Hindi, Macedonian, Spanish, Serbian, Romanian, Gujarati, and Marathi. Seven of the ten fastest languages are from India, which is unsurprising given the geographic distribution of workers. Some languages follow the pattern of having a smattering of assignments completed early, with the rate picking up later.

Language skills and location We measured the average quality of workers who were in countries that plausibly speak a language, versus workers from countries that did not have large speaker populations of that language. We used the Ethnologue (Lewis et al., 2013) to compile the list of countries where

	Avg. Turker quality (# Ts) In region	Out of region	Primary locations of Turkers in region	Primary locations of Turkers out of region
Hindi	0.635 (296)	**0.695** (7)	India (284) UAE (5) UK (3)	Saudi Arabia (2) Russia (1) Oman (1)
Tamil	**0.653** (273) *	0.250 (2)	India (266) US (3) Canada (2)	Tunisia (1) Egypt (1)
Malayalam	0.764 (234)	**0.827** (2)	India (223) UAE (6) US (3)	Saudi Arabia (1) Maldives (1)
Spanish	0.814 (191)	**0.837** (18)	US (122) Mexico (16) Spain (14)	India (15) Newzealand (1) Brazil (1)
French	0.750 (170)	**0.816** (11)	India (62) US (45) France (23)	Greece (2) Netherlands (1) Japan (1)
Chinese	**0.601** (116)	0.552 (21)	US (75) Singapore (13) China (9)	Hongkong (6) Australia (3) Germany (2)
German	**0.825** (91)	0.768 (41)	Germany (48) US (25) Austria (7)	India (34) Netherlands (1) Greece (1)
Italian	**0.863** (90)	0.796 (42)	Italy (42) US (29) Romania (7)	India (33) Ireland (2) Spain (2)
Amharic	**0.136** (16) *	0.009 (99)	US (14) Ethiopia (2)	India (70) Georgia (9) Macedonia (5)
Kannada	0.704 (105)	NA (0)	India (105)	
Arabic	**0.741** (60) *	0.604 (45)	Egypt (19) Jordan (16) Morocco (9)	US (19) India (11) Canada (3)
Sindhi	**0.190** (96)	0.062 (9)	India (58) Pakistan (37) US (1)	Macedonia (4) Georgia (2) Indonesia (2)
Portuguese	0.869 (101)	**0.963** (3)	Brazil (44) Portugal (31) US (15)	Romania (1) Japan (1) Israel (1)
Turkish	0.759 (76)	**0.804** (27)	Turkey (38) US (18) Macedonia (8)	India (19) Pakistan (4) Taiwan (1)
Telugu	**0.799** (102)	0.500 (1)	India (98) US (3) UAE (1)	Saudi Arabia (1)
Irish	**0.742** (54)	0.709 (47)	US (39) Ireland (13) UK (2)	India (36) Romania (5) Macedonia (2)
Swedish	**0.734** (54)	0.711 (45)	US (25) Sweden (22) Finland (3)	India (23) Macedonia (6) Croatia (2)
Czech	**0.706** (45)	0.612 (50)	US (17) Czech Republic (14) Serbia (5)	Macedonia (22) India (10) UK (5)
Russian	**0.146** (67)	0.119 (27)	US (36) Moldova (7) Russia (6)	India (14) Macedonia (4) UK (3)
Breton	0.167 (3)	**0.178** (89)	US (3)	India (83) Macedonia (2) China (1)

Table 3: Translation quality when partitioning the translations into two groups, one containing translations submitted by Turkers whose location is within regions that plausibly speak the foreign language, and the other containing translations from Turkers outside those regions. In general, in-region Turkers provide higher quality translations. (*) indicates differences significant at p=0.05.

each language is spoken. Table 3 compares the average translation quality of assignments completed within the region of each language, and compares it to the quality of assignments completed outside that region.

Our workers reported speaking 95 languages natively. US workers alone reported 61 native languages. Overall, 4,297 workers were located in a region likely to speak the language from which they were translating, and 2,778 workers were located in countries considered out of region (meaning that about a third of our 5,281 Turkers completed HITs in multiple languages).

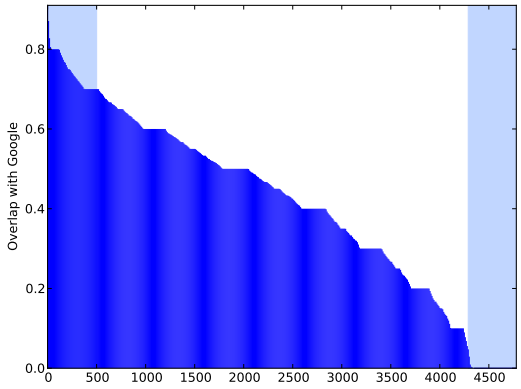
Table 3 shows the differences in translation quality when computed using in-region versus out-of-region Turkers, for the languages with the greatest number of workers. Within region workers typically produced higher quality translations. Given the number of Indian workers on Mechanical Turk, it is unsurprising that they represent majority of out-of-region workers. For the languages that had more than 75 out of region workers (Malay, Amharic, Icelandic, Sicilian, Wolof, and Breton), Indian workers represented at least 70% of the out of region workers in each language.

Validation of controls Higher overlap with Google Translate corresponds slightly to higher quality scores (Pearson ρ = 0.27), suggesting that some control translations have already been indexed by Google. We therefore investigated the validity of our use of Wikipedia inter-language links as gold standard translations.

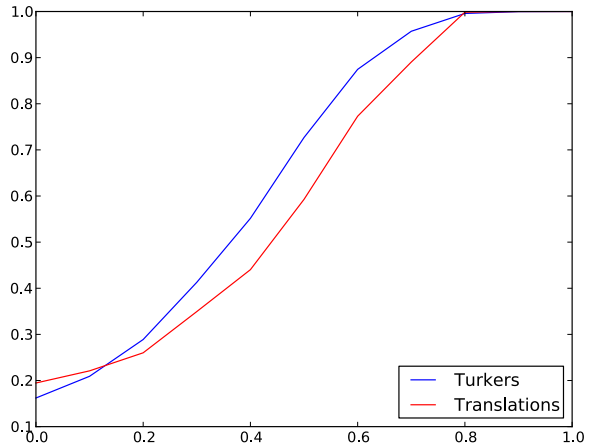
As an extrinsic measure of translation quality, we measure the proportion of *<word, translation>* pairs received from Mechanical Turk which appear in external dictionaries. We had external bilingual dictionaries for 24 of the 100 languages examined in this study.[3] We used each of these bilingual dictionaries to compute language-level quality scores, using the dictionary's translation as a gold standard translation for any of the 10,000 foreign words in our study. In some cases, the overlap of words was low, since dictionaries list root forms of words and we translated inflected forms.

We ranked the 24 languages based on their translation quality scores according to the external dictionaries, and calculated correlation when compared

[3]The bilingual dictionaries were between English and az, bg, bn, bs, cy, es, fa, hi, id, lv, ms, ne, pl, ro, ru, sk, so, sq, sr, ta, tr, uk, ur, and uz.



(a) Individual Turker overlap with Google Translate. We remove the 500 workers with the highest overlap (shaded region on the left) from our analysis, as it is reasonable to assume these workers are cheating by submitting translations from Google. We believe the workers with no overlap (shaded region on the right) are also likely to be cheating, e.g. by submitting random text, and we target these workers through our embedded quality controls.



(b) Cumulative distribution of overlap with Google translate for workers and translations. We see that eliminating all workers with >70% overlap with google translate still preserves 90% of translations and >90% of workers.

Figure 5

to the ranking produced by our gold standard control. The Pearson correlation was $\rho = 0.53$. We computed a correlation coefficient of $\rho = 0.37$ for ranking of the translation quality of individual Turkers using the two sources of reference translations.

Given the strong positive correlation, we are satisfied that the trends reported in this paper hold when using either our gold standard translations or external bilingual dictionaries.

language	full sentences	dictionaries
Bengali	732k (22k)	22k
Hindi	1,488k (40k)	22k
Malayalam	863k (32k)	23k
Tamil	916k (38k)	25k
Telugu	1,097k (46k)	21k
Urdu	1,356k (35k)	20k

Table 4: Size of data set (number of words) collected for each language. For full sentences, number of parallel sentences is shown in parentheses.

language	sentences	dictionaries
Bengali	12.03	17.29
Hindi	16.19	18.10
Malayalam	6.65	9.72
Tamil	8.08	9.66
Telugu	11.94	13.70
Urdu	19.22	21.98

Table 5: BLEU scores for translating into English using full sentence translations alone, and with the addition of single-word dictionaries. Scores are calculated using four reference translations and represent the mean of three MERT runs.

Quality improvements for SMT systems Using the full sentence translations we gathered for the six Indian languages, we were able to train a statistical machine translation system. Table 4 gives the size of the data set for each of these languages. Table 5 shows that the sentences alone produce decent translation performance, with BLEU scores as high as 19.22 for Urdu. We also see that adding the single-word dictionaries to the training set produces consistent performance gains, ranging from 1 to 5 BLEU points. Unfortunately, we do not have professionally translated data against which to compare the resulting performances, but we find the outcomes encouraging as an indication of the potential of Turkers to provide high quality translations.

6 Discussion

Mechanical Turk gives researchers access to a diverse set of bilingual workers, making it a promising resource for researchers and developers of mul-

workers	quality	speed	
many	high	fast	de es fr gu it kn ml nl pt ro sr te tl
		slow	ar ga he pa sv tr
	low	fast	hi mr ta ur
		slow	bn bo bpy ceb ne new pl ru sd zh
few	high	fast	bs hr mk ms sh
		slow	af an ast bcl be bg cs da el eu fi fy gl ht hu id ilo is ja jv kk ko lt mg nds no pam scn sk sl sq th uk uz war yo
	low	fast	–
		slow	am az br ca cy hy ka lb lv nap nn pms ps so su sw tt vi wa
none	low	slow	diq eo fa io ku qu wo

Table 6: The green box shows the languages which we have found most supported by MTurk in terms of worker population, worker ability, and speed of completion. Languages have "many" workers if we recorded more than 50 active in-region Turkers. Languages are "high" quality if they average greater than 70% accuracy on our controls. Languages are "fast" if all of our translations (approx. 10,000 words each) were completed within two weeks.

tilingual systems. Based on our study, we can confidently recommend 13 languages as good candidates for future research: German, Spanish, French, Gujarati, Italian, Kannada, Malayalam, Dutch, Portuguese, Romanian, Serbian, Telugu, and Tagalog. These languages have large Turker populations who complete tasks quickly and accurately. Table 6 summarizes the strengths and weaknesses of all 100 languages covered in our study.

While the demographics reported are likely to shift overtime, as Amazon expands its methods of payments and reaches new countries, the data presented provides a valuable snapshot of the current state of MTurk, and the methods used can be applied generally in future research.

Unfiltered data can contain large amounts of noise, a variety of techniques can be incorporated into crowdsourcing pipelines to ensure high quality data. As a best practice, we suggest restricting workers to countries that plausibly speak the foreign language of interest, and embedding gold standard controls where possible, rather than relying solely on self-reported language skills.

Although our study targeted bilingual workers on Mechanical Turk, and neglected monolingual workers, we believe our results reliably represent the current speaker populations, since the vast majority of the work available on the crowdsourced platform is currently English-only. We therefore assume the number of non-English speakers is small. In the future, it may be desirable to recruit monolingual foreign workers. In such cases, we recommend other tests to validate their language abilities in place of our translation test. These could include performing narrative cloze, or listening to audio files containing speech in different language and identifying their language.

7 Data release

We plan to release all data and code used in this study upon publication of this paper. Our data release will include the raw data, along with bilingual dictionaries that are filtered to be high quality. It will include 256,604 translation assignments from 5,281 Turkers and 20,952 synonym assignments from 1,005 Turkers, along with meta information like geolocation and time submitted, plus external dictionaries used for validation. The dictionaries will contain 1.5M total translated words in 100 languages, along with code to filter the dictionaries based on different criteria.

References

Amazon. 2013. Service summary tour for requesters on Amazon Mechanical Turk. `https://requester.mturk.com/tour`.

Vamshi Ambati and Stephan Vogel. 2010. Can crowds build parallel corpora for machine translation systems? In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics.

Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. 2010. Active learning and crowd-sourcing for machine translation. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*.

Vamshi Ambati. 2012. *Active Learning and Crowdsourcing for Machine Translation in Low Resource Scenarios*. Ph.D. thesis, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Michael S. Bernstein, Greg Little, Robert C. Miller, Bjrn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2010. Soylent:

a word processor with a crowd inside. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*.

Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and Tom Yeh. 2010. VizWiz: nearly real-time answers to visual questions. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*.

Michael Bloodgood and Chris Callison-Burch. 2010. Large-scale cost-focused active learning for statistical machine translation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

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

Jia Deng, Alexander Berg, Kai Li, and Li Fei-Fei. 2010. What does classifying more than 10,000 image categories tell us? In *Proceedings of the 12th European Conference of Computer Vision (ECCV*, pages 71–84.

Maxine Eskenazi, Gina-Anne Levow, Helen Meng, Gabriel Parent, and David Suendermann. 2013. *Crowdsourcing for Speech Processing, Applications to Data Collection, Transcription and Assessment*. Wiley.

Siamak Faridani, Björn Hartmann, and Panagiotis G. Ipeirotis. 2011. What's the right price? pricing tasks for finishing on time. In *Third AAAI Human Computation Workshop (HCOMP'11)*.

Ulrich Germann. 2001. Building a statistical machine translation system from scratch: How much bang for the buck can we expect? In *ACL 2001 Workshop on Data-Driven Machine Translation*, Toulouse, France.

Panagiotis G. Ipeirotis. 2010a. Analyzing the mechanical turk marketplace. In *ACM XRDS*, December.

Panagiotis G. Ipeirotis. 2010b. Demographics of Mechanical Turk. Technical Report Working paper CeDER-10-01, New York University, Stern School of Business.

Ann Irvine and Alexandre Klementiev. 2010. Using Mechanical Turk to annotate lexicons for less commonly used languages. In *Workshop on Creating Speech and Language Data with MTurk*.

Ian Lane, Matthias Eck, Kay Rottmann, and Alex Waibel. 2010. Tools for collecting speech corpora via mechanical-turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Los Angeles.

Florian Laws, Christian Scheible, and Hinrich Schütze. 2011. Active learning with amazon mechanical turk. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland.

Matthew Lease, Jessica Hullman, Jeffrey P. Bigham, Juho Kim Michael S. Bernstein and, Walter Lasecki, Saeideh Bakhshi, Tanushree Mitra, and Robert C. Miller. 2013. Mechanical Turk is not anonymous. `http://dx.doi.org/10.2139/ssrn.2228728`.

Vili Lehdonvirta and Mirko Ernkvist. 2011. Knowledge map of the virtual economy: Converting the virtual economy into development potential. `http://www.infodev.org/en/Document.1056.pdf`, April. An InfoDev Publication.

M. Paul Lewis, Gary F. Simons, and Charles D. Fennig (eds.). 2013. Ethnologue: Languages of the world, seventeenth edition. `http://www.ethnologue.com`.

Greg Little, Lydia B. Chilton, Rob Miller, and Max Goldman. 2009. Turkit: Tools for iterative tasks on mechanical turk. In *Proceedings of the Workshop on Human Computation at the International Conference on Knowledge Discovery and Data Mining (KDD-HCOMP '09)*, Paris.

Matthew Marge, Satanjeev Banerjee, and Alexander Rudnicky. 2010. Using the Amazon Mechanical Turk to transcribe and annotate meeting speech for extractive summarization. In *Workshop on Creating Speech and Language Data with MTurk*.

George A. Miller. 1995. WordNet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Robert Munro and Hal Tily. 2011. The start of the art: Introduction to the workshop on crowdsourcing technologies for language and cognition studies. In *Crowdsourcing Technologies for Language and Cognition Studies*, Boulder.

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

Gabriel Parent and Maxine Eskenazi. 2011. Speaking to the crowd: looking at past achievements in using crowdsourcing for speech and predicting future challenges. In *Proceedings Interspeech 2011, Special Session on Crowdsourcing*.

Alexander J. Quinn and Benjamin B. Bederson. 2011. Human computation: A survey and taxonomy of a growing field. In *Computer Human Interaction (CHI)*.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using Amazon's Mechanical Turk. In *Workshop on Creating Speech and Language Data with MTurk*.

Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers?: Shifting demographics in Amazon Mechanical Turk. In *alt.CHI session of CHI 2010 extended abstracts on human factors in computing systems*, Atlanta, Georgia.

Yaron Singer and Manas Mittal. 2011. Pricing mechanisms for online labor markets. In *Third AAAI Human Computation Workshop (HCOMP'11)*.

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

Alexander Sorokin and David Forsyth. 2008. Utility data annotation with amazon mechanical turk. In *First IEEE Workshop on Internet Vision at CVPR*.

Luis von Ahn. 2005. *Human Computation*. Ph.D. thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Omar F. Zaidan and Chris Callison-Burch. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1220–1229. Association for Computational Linguistics.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine translation of Arabic dialects. In *The 2012 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.