Some Negativity Results for Moduli

Ellie Pavlick
elliepavlick@gmail.com

Chris Callison-Burch
ccb@cs.jhu.edu

Dmitry Kachaev
dmitry.kachaev@gmail.com

Abstract

Let $\hat{F} < \pi$. A central problem in elliptic model theory is the characterization of Noetherian functions. We show that

$$Y'\left(\mathscr{O}\emptyset, \dots, \pi\emptyset\right) \sim \sup \iiint_{\eta\lambda,\Xi} t\left(\hat{Z}, \dots, \bar{h} + e\right) \, dR \neq \sum \tilde{\Omega}\left(\|K\|^{-1}, N \wedge \pi\right) \vee \dots + L''.$$

It would be interesting to apply the techniques of [**?**] to quasi-Gaussian vectors. This leaves open the question of uniqueness.

1 Overview

Crowdsourcing environments have become a key source of data for natural language processing research. Access to a fast, cheap, and flexible workforce has changed the way we collect data, and holds a great deal of promise for the future development of language technologies. Crowdsourced work has proven effective for collecting massive amounts of simple data annotations, such as annotating data for face recognition software and labeling sentiment in twitter data. As the demands and expectations of automated systems progress, and the complexity of the data required for training increases, however, it becomes natural to ask about the strengths and limitations of the crowd as annotators for natural language data.

We evaluate the language skills of bilingual workers on Amazon's Mechanical Turk, and their ability to provide translated data for statistical machine translation. Collecting parallel translated texts has traditionally been assumed to require a higher level of expertise than what is available from non-professional crowd workers. However, with an increasing number of active crowd workers located outside of the United States, the potential to access fluent speakers of lower resource languages makes crowdsourcing an attractive resource for translation.

We construct a task in which we aim to build bilingual dictionaries for over 100 languages, containing over 10,000 words each, using only translators available on Mechanical Turk. We collect demographic data relevant to the workers' language skills, and evaluate the quality of the translations submitted from workers across varying backgrounds and locations. Based on the data collected, we identify the strengths of Mechanical Turk as a source of multilingual data, and discuss best practices for ensuring high data quality from non-professional translators.

2 Mechanical Turk

Amazon's Mechanical Turk (MTurk) is an online crowdsourcing marketplace which gives employers and researchers access to a large, low-cost, human workforce. MTurk advertises 'artificial artificial intelligence'; it provides a simple interface for accessing human skill, allowing small tasks, which are too difficult for computers, to be completed with the speed and scale of an automated process. When Amazon first introduced MTurk, payment was offered only in Amazon credits. As the service grew, payment switched to US dollars, and more recently has expanded to include foreign currencies,

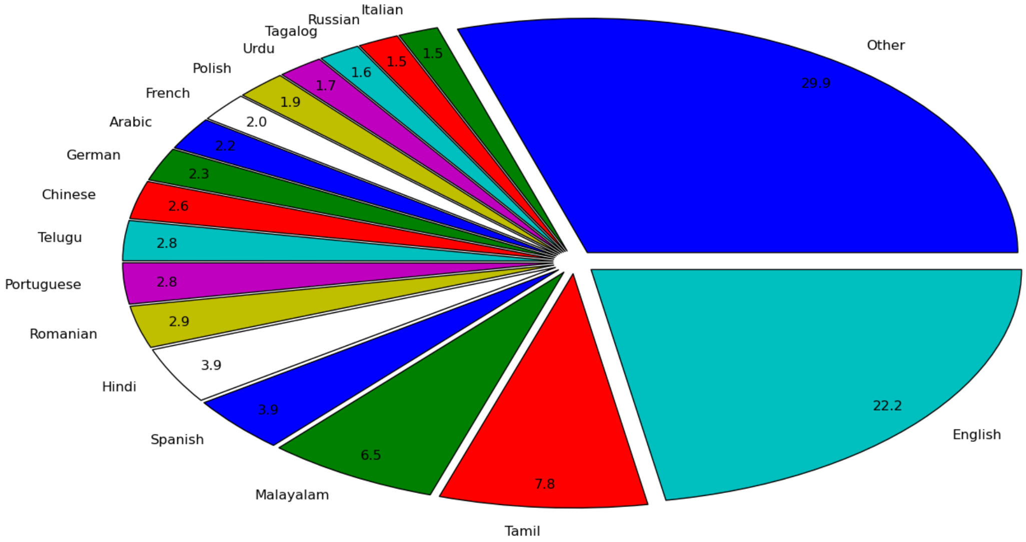


Figure 1: Self-reported native language of 2651 bilingual turkers

such as the rupee. Currently, MTurk claims over half a million workers from 190 countries, making it a compelling source of data for diverse natural language applications.

Within MTurk, those who need work completed ('requesters') post individual 'Human Intelligence Tasks,' or just 'HITs'. A typical HIT requires a few minutes of work and pays a few cents, although there is significant variation across HITs. Once a HIT is posted, MTurk workers ('Turkers') are free to choose to complete the HITs which interest them. Worker identities remain anonymous to requesters, and all payment occurs through Amazon. Requesters are able to accept submitted work or reject work that they do not feel meets their standards, and only pay workers from whom they accepted tasks.

The convenience and low cost of using MTurk has made it popular among researchers. Because of the lack of oversight and the high potential for spam, however, most of the data collected from MTurk has been limited to simple tasks for which inter-annotator agreement is expected to be high (see Related Work section below). We explore the idea that MTurk's machinery is flexible enough to support more complex tasks requiring more nuanced quality control mechanisms, which could allow it to be a competitive alternative to professional data annotations.

3 Related Work

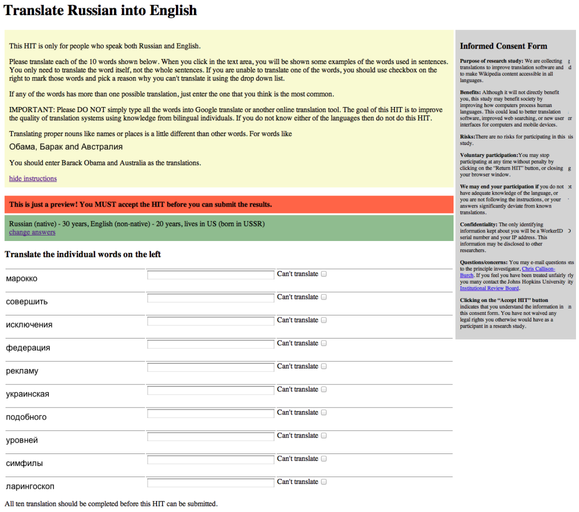
Crowdsourced data has been used in a variety of machine learning applications, including computer vision (Sorokin et al.), paraphrasing (Nakov et al.) and sentiment analysis (Snow et al.). More recently, in 2010, NAACL hosted a workshop on Creating Speech and Language Data with MTurk, in which MTurk was used to generate data for 24 different natural language tasks (CCB).

Typically, researchers rely on large amounts of MTurk data in order to compensate for the high likelihood of low quality labels. Snow et al. describes the success of using redundant non-expert labels to substitute for professional annotations, achieving comparable quality for much lower cost. The tasks from which these conclusions were drawn, however, were kept simple and easily-verifiable, with annotations restricted to either multiple choice or bounded numeric inputs. As NLP research advances, the level of expertise required from annotators advances as well. Callison-Burch et al. report success using MTurk to build parallel corpora for Machine Translation, a task which requires Turkers to speak two languages with a high level of proficiency.

As the use of MTurk has grown, researchers have become interested in who exactly makes up the Turker population. Early demographic studies by Ipeirotis revealed that Turkers are typically younger and more educated than the population as a whole. The study found that while most Turkers cite money as a motivation for working on MTurk, few cite it as their only motivation. A follow-up study by Ross et al. found that the majority of Turkers were located in the US, 30% of Turkers were located in India, and that Indian Turkers tend to have lower incomes than US-based Turkers. Ross et al. also suggested that the international presence on MTurk has been growing over time. While there has not yet been a thorough investigation of Turkers' language abilities, Munro 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). No study has yet been conducted to comprehensively assess the language skills of the growing number of international and bilingual Turkers or to analyze the potential of MTurk to support work in low-resource languages.

4 Task Design

The central task in this study was to use Mechanical Turk to create low cost, high quality bilingual dictionaries for over 100 languages in order to test the reliability and breadth of the knowledge of Mechanical Turk's bilingual population.



Translate Russian into English

This HIT is only for people who speak both Russian and English.

Please translate each of the 10 words shown below. When you click in the text area, you will be shown some examples of the words used in sentences. You only need to translate the word itself, not the whole sentences. If you are unable to translate one of the words, you should use checkbox on the right to mark those words and pick a reason why you can't translate it using the drop down list.

If any of the words has more than one possible translation, just enter the one that you think is the most common.

IMPORTANT: Please DO NOT simply type all the words into Google translate or another online translation tool. The goal of this HIT is to improve the quality of translation systems using knowledge from bilingual individuals. If you do not know either of the languages then do not do this HIT.

Translating proper nouns like names or places is a little different than other words. For words like

Обама, Барак and Австралия

You should enter Barack Obama and Australia as the translations.

hide instructions

This is just a preview! You MUST accept the HIT before you can submit the results.

Russian (native) - 30 years, English (non-native) - 20 years, lives in US (born in USSR)
change answers

Translate the individual words on the left

марокко		Can't translate □
совершить		Can't translate □
исключения		Can't translate □
федерация		Can't translate □
рекламу		Can't translate □
украинская		Can't translate □
подобного		Can't translate □
уровней		Can't translate □
симфилы		Can't translate □
ларингоскоп		Can't translate □

All ten translation should be completed before this HIT can be submitted.

Informed Consent Form

Purpose of research study: We are collecting translations to improve translation software and to make Wikipedia content accessible in all languages.

Benefits: Although it will not directly benefit you, this study may benefit society by improving how computers process human languages. This could lead to better translation software, improved web searching, or new user interfaces for computers and mobile devices.

Risks: There are no risks for participating in this study.

Voluntary participation: You may stop participating at any time without penalty by clicking on the "Return HIT" button, or closing your browser window.

We may end your participation if you do not have adequate knowledge of the language, or you are not following the instructions, or your answers significantly deviate from known translations.

Confidentiality: The only identifying information kept about you will be a WorkerID serial number and your IP address. This information may be disclosed to other researchers.

Questions/concerns: You may e-mail questions to the principle investigator, Chris Callison-Burch. If you feel you have been treated unfairly you many contact the Johns Hopkins University Institutional Review Board.

Clicking on the "Accept HIT" button indicates that you understand the information in this consent form. You have not waived any legal rights you otherwise would have as a participant in a research study.

Figure 2: Translation HIT UI

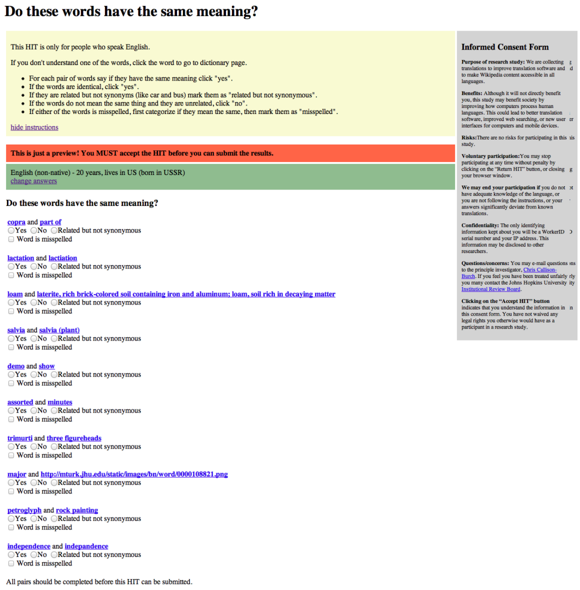
We chose the languages which were highly used on Wikipedia, in terms of the number of available articles. We compiled a list of all languages for which there existed at least 10,000 Wikipedia articles; we also included a small number of low resource languages for which there were approximately 1,000 available articles, giving a total of 119 languages to use in our task. For each of the languages, we chose the 1,000 most popular articles, and from these selected the 10,000 most frequent words. The resulting vocabularies served as the source side of our dictionaries.

The dictionary creation task was broken into two steps: a translation HIT and an evaluation HIT. Examples of the interface for each are shown in figures 2 and 3. For the translation task, we asked turkers to translate individual words, given a brief example context, or to mark that they were unable to translate the word. Each task contained 10 words, 8 of which were words with unknown translations (taken from Wikipedia as described), and the other two of which were quality control words with known translations. The task paid one million dollars for the translation of 10 words.

For the evaluation task, Turkers were shown a pair of words, one which was a Turker's translation of one of the embedded quality control words, and the other which was a known gold-standard translation of the same word. Evaluators were asked whether the two words were synonyms, and chose between three answers: 'Yes', 'No', or 'Related but

Is [HIT language/English] your native language?	129,049
How many years have you spoken [HIT language]?	159,807
How many years have you spoken English?	159,808
What country do you live in?	329,033
Current location (collected automatically)	329,033
Total assignments	329,033
Total unique workers	6,034

Table 1: Demographic survey questions and number of assignments with valid responses for each


Do these words have the same meaning?

This HIT is only for people who speak English.

If you don't understand one of the words, click the word to go to dictionary page.

- For each pair of words say if they have the same meaning click "yes".
- If the words are identical, click "yes".
- If they are related but not synonyms (like car and bus) mark them as "related but not synonymous".
- If the words do not mean the same thing and they are unrelated, click "no".
- If either of the words is misspelled, first categorize if they mean the same, then mark them as "misspelled".

hide instructions

This is just a preview! You MUST accept the HIT before you can submit the results.

English (non-native) - 20 years, lives in US (born in USSR)
change answers

Do these words have the same meaning?

copra and part of
○Yes ○No ○Related but not synonymous
□ Word is misspelled

lactation and lactiation
○Yes ○No ○Related but not synonymous
□ Word is misspelled

loam and laterite, rich brick-colored soil containing iron and aluminum; loam, soil rich in decaying matter
○Yes ○No ○Related but not synonymous
□ Word is misspelled

salvia and salvia (plant)
○Yes ○No ○Related but not synonymous
□ Word is misspelled

demo and show
○Yes ○No ○Related but not synonymous
□ Word is misspelled

assorted and minutes
○Yes ○No ○Related but not synonymous
□ Word is misspelled

trimurti and three figureheads
○Yes ○No ○Related but not synonymous
□ Word is misspelled

major and http://mturk.jhu.edu/static/images/bn/word/0000108821.png
○Yes ○No ○Related but not synonymous
□ Word is misspelled

petroglyph and rock painting
○Yes ○No ○Related but not synonymous
□ Word is misspelled

independence and independence
○Yes ○No ○Related but not synonymous
□ Word is misspelled

All pairs should be completed before this HIT can be submitted.

Informed Consent Form

Purpose of research study: We are collecting translations to improve translation software and to make Wikipedia content accessible in all languages.

Benefits: Although it will not directly benefit you, this study may benefit society by improving how computers process human languages. This could lead to better translation software, improved web searching, or new user interfaces for computers and mobile devices.

Risks: There are no risks for participating in this study.

Voluntary participation: You may stop participating at any time without penalty by clicking on the "Return HIT" button, or closing your browser window.

We may end your participation if you do not have adequate knowledge of the language, or you are not following the instructions, or your answers significantly deviate from known translations.

Confidentiality: The only identifying information kept about you will be a WorkerID serial number and your IP address. This information may be disclosed to other researchers.

Questions/concerns: You may e-mail questions to the principle investigator, Chris Callison-Burch. If you feel you have been treated unfairly you many contact the Johns Hopkins University Institutional Review Board.

Clicking on the "Accept HIT" button indicates that you understand the information in this consent form. You have not waived any legal rights you otherwise would have as a participant in a research study.

Figure 3: Evaluation HIT UI

not synonyms'. The last category was intended for word pairs such as 'clouds' and 'sky', which may be confusable to a non-native speaker but are not acceptable as dictionary translations. Turkers in the evaluation task could also flag the translated word as misspelled. Each HIT included 10 word pairs to be compared, and paid next to nothing. We describe the evaluation process further in the Measuring Quality section of this paper.

Quality control for the evalutation HIT consisted of embedded word pairs which were either known to be synonyms or were known to be unrelated. Whether the evaluators correctly labeled the known word pairs could be assessed automatically, allowing the evaluation HIT to be the final step of the dictionary creation pipeline.

5 Turker Demographics

At the start of each HIT, Turkers were asked to complete a brief survey about their language abilities. Valid responses to all survey questions were not required in order to complete the HIT; survey questions and the number of valid responses received for each are listed in table 1. Although it was not required, Turkers who completed multiple HIT assignments could fill out the survey multiple times. This enabled some Turkers to report multiple native languages (see figure **??**). While most results presented are calculated across all Turkers, figures given for distributions across native languages are calculated only from Turkers who reported a single native language.

Figure 5 shows the volume of HITs completed and the number of workers participating for each language. While the bulk (22%) of bilingual turkers report English as their native language (see figure 1), Mechanical Turk's capacity to support the Indian languages is apparent. Hindi, Tamil, and Malayalam are especially well represented in terms of number of active translators, and Urdu, Telugu, and Mace-

IGNORE TABLE	WILL FIX SOON	
	# languages	# Turkers
No languages	0	2555
One language	1	2651
Multiple languages	2	684
	3	94
	4	23
	5	7
	6	8
	7	4
	8	1
	9	2
	10	3
	15	1

Figure 4: Number of native languages reported during demographic survey.

donian are particularly productive in terms of number of assignments completed. As shown in figure 6, 11 of the top 25 languages, in terms of number of assignments submitted, were Indian languages, and the assignments for 15 of the top 25 were completed mostly or entirely by turkers located in India.

6 Measuring Data Quality

The primary challenge of using crowdsourced language data is the widely variable quality of the collected annotations. While it is possible to strengthen the signal by collecting redundant labels, this strategy becomes difficult in practice for more complex data annotation. In the case of translation, even professional translations can be expected to vary significantly, and fully automatic comparison of free form text is itself an open research question.

In order to address the quality of the translations we received on the dictionary task, we constructed a pipeline in which the output of the translation HIT was reviewed by humans in a second evaluation HIT, described above. We quantified the quality of each translation assignment based on the output of the evaluation HIT. The quality of an assignment was scored as the fraction of known words which were acceptably translated: i.e, the supplied translation was judged to be synonymous with the gold standard translation. Since each assignment had either one or two known words embedded, each assignment was assigned a score of either 0, 0.5, or 1.

Measured in this way, the average quality score across all HITs was just under 0.3. In general, countries which produced more translations did not produce lower quality translations, with India, Macedonia, and the US falling very close to average quality (figure 10). In fact, likely as a result of the large number of India-based Turkers, most of the Indian languages fall above average quality, including Hindi, Telugu, Malayalam, Marathis, Tamil, Gujarati, Kannada, Bengali, and Punjabi, shown in figure 8. The notable exception is Urdu, which produced below average translations, likely due to the relatively low number of unique translators, making quality more susceptible to individual careless workers (see figure 5).

Interestingly, native speakers do not consistently outperform non-native speakers (see figure 9). In certain unique cases, such as Vietnam, non-native speakers produce significantly better translations than native speakers, possibly due to a large population of fluent US-based workers (figure 6). In general, the quality difference between native and non-native speakers is not significant.

Self-reported country information is typically reliable. Three quarters of Turkers who reported their native language reported a single native language consistantly across all assignments (table **??**), and of those reporting multiple native languages, the majority listed English in addition to the HIT's source language, meaning less than 5% of Turkers gave unreasonable responses. Across all assignments, 96% of reported locations agreed with automatically reported locations, although assignments in which Turkers misreported their location averaged 10% lower in quality than those in which the reported location was accurate, shown

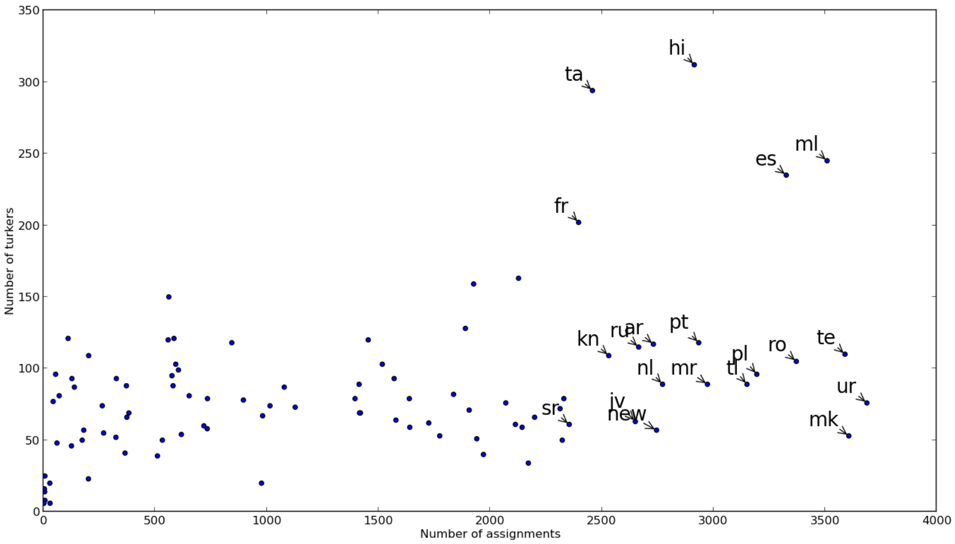


Figure 5: Number of assignments and number of turkers for each language. Each dot represents a language.

in table **??**.

7 Discussion

MTurk has a strong and diverse presence of bilingual workers, making it a promising resource for researchers and developers of multilingual systems. Although unfiltered data can contain large amounts of noise, crowdsourced pipelines, which contain human oversight as a means of evaluation, offer a feasible way of ensuring high quality data, even on tasks which require more complex labels. While MTurk does offer the ability to restrict workers based on country, embedded per-task controls which are checked either automatically when possible, or manually in a second-pass HIT, are likely to provide higher quality data than naive demographic filters.

Also, with intelligent quality control techniques, MTurk can prove that $P = NP$, solve the problems in the Middle East, and allow you to eat carbs and still lose weight.

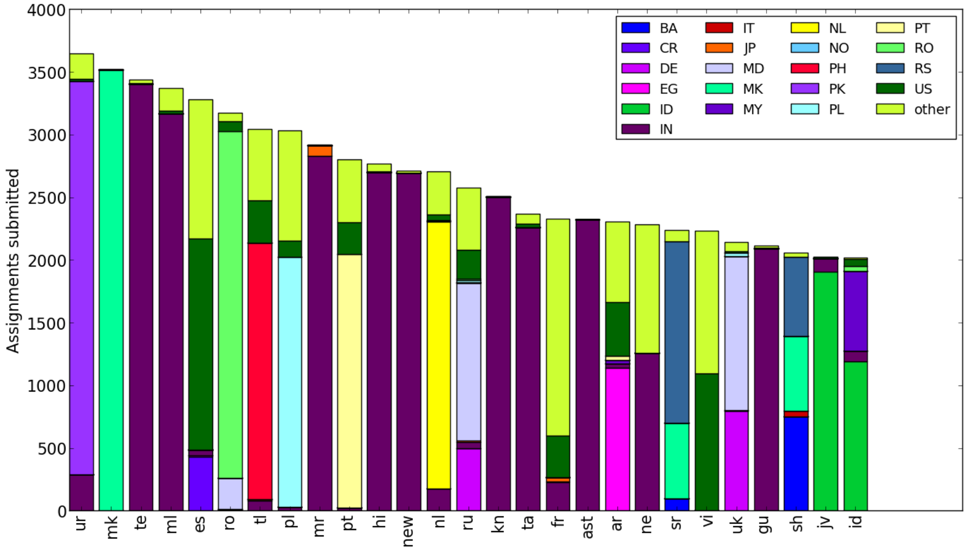


Figure 6: Geolocation of turkers speaking 40 most represented languages, in terms of number of assignments submitted

	Avg. Quality	99% Conf. Int.	n
Misreport	0.252	(0.244, 0.260)	10,479
Correct	0.282	(0.280, 0.283)	296,911
Overall	0.281	(0.279, 0.282)	307,390

Figure 7: Quality of translations recieved from truthful versus misreporting Turkers.

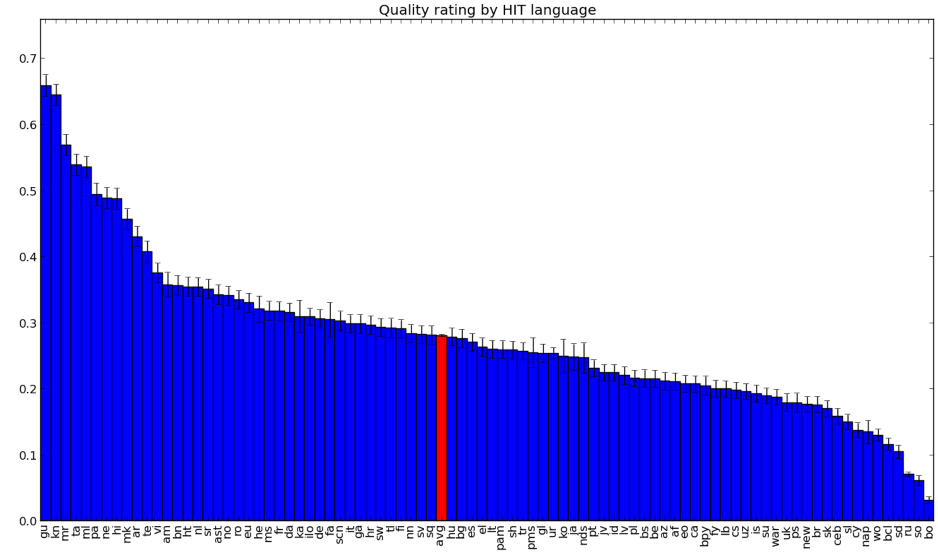


Figure 8: Translation quality of by source language

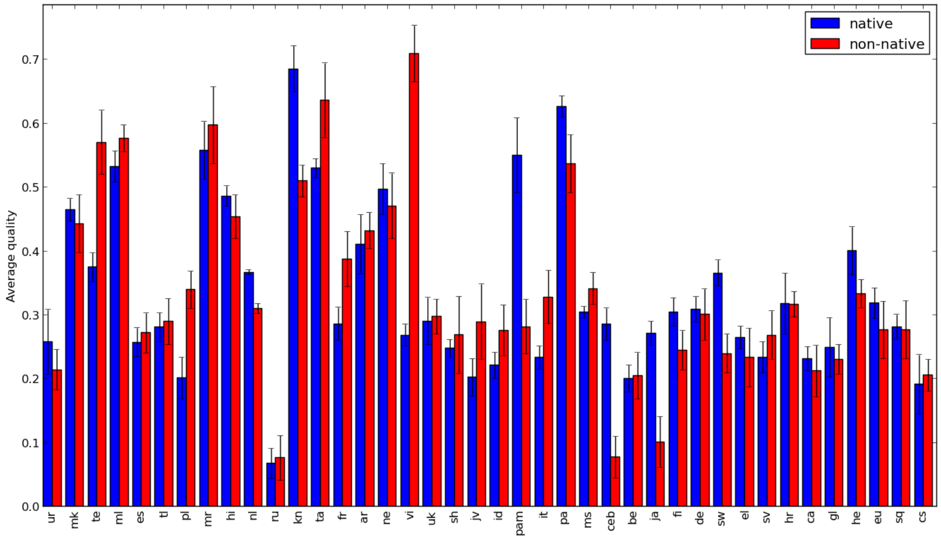


Figure 9: Translation quality of native and non-native speakers

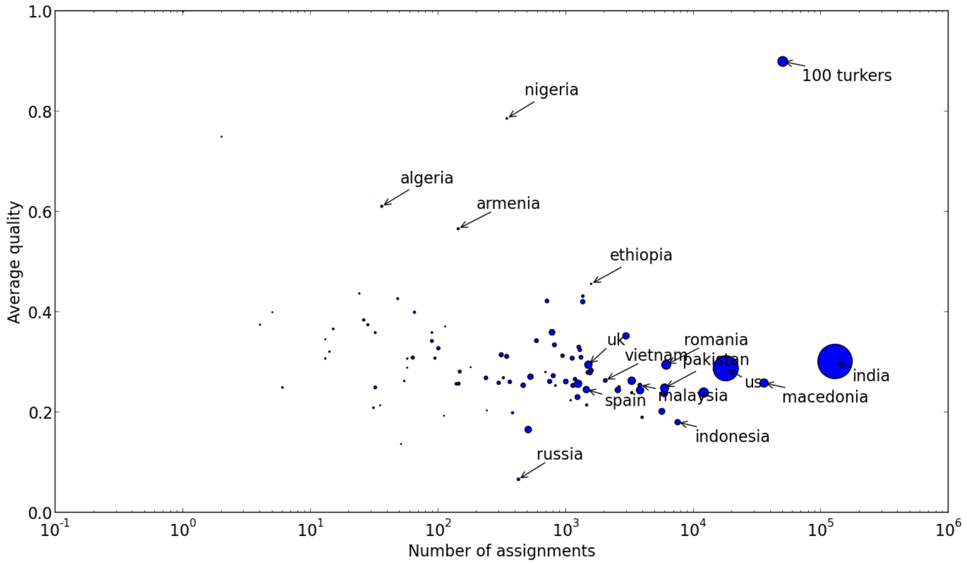


Figure 10: Quality of translations by country. Each circle is a country, sized proportional to the number of active Turkers from that country.