



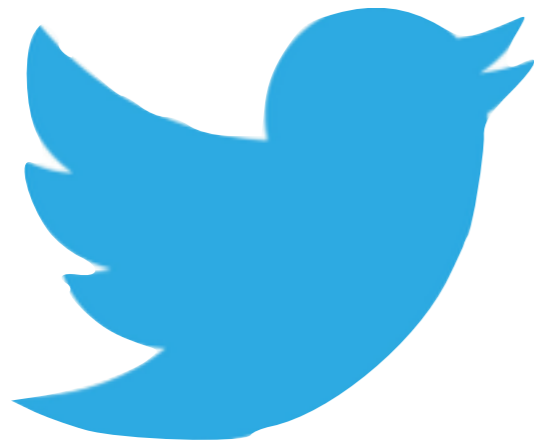
# Crowdsourcing Translation

**Chris Callison-Burch**

April 22, 2014

with Rui Yan, Mingkun Gao, Ellie Pavlick, Matt Post, Dmitry Kachaev,  
Ann Irvine, Omar Zaidan, Scott Novotney, and 10clouds

facebook.



TED

IDEAS WORTH SPREADING

courseera

IBM



duolingo

genngo

# Can Turkers Translate?

## Urdu

میں اس خطے میں ابتدائی 1994  
انسانوں کی باقیات جو تقریباً 8  
لاکھ سال پرانی مانی جاتی ہے ،  
دریافت کی گئیں جنہیں ہومو اینٹی  
سیسر یعنی 'بانی انسان' کا نام دیا  
گیا۔  
اس سے قبل 6 لاکھ پرانے انسان  
جنہیں سائنسی اصطلاح میں ہومو  
بیڈلبرجینسس کہا جاتا ہے ، اس  
خطے کے قدیم ترین رہائشی مانے  
جاتے تھے۔  
آثارِ قدیمہ کے ماہرین کا کہنا ہے  
کہ انہیں ایسے شواہد ملے ہیں جن  
سے پتہ چلتا ہے کہ اس خطے کے  
لوگ ڈھلائی کیے ہوئے اوزار بھی

## LDC Translation

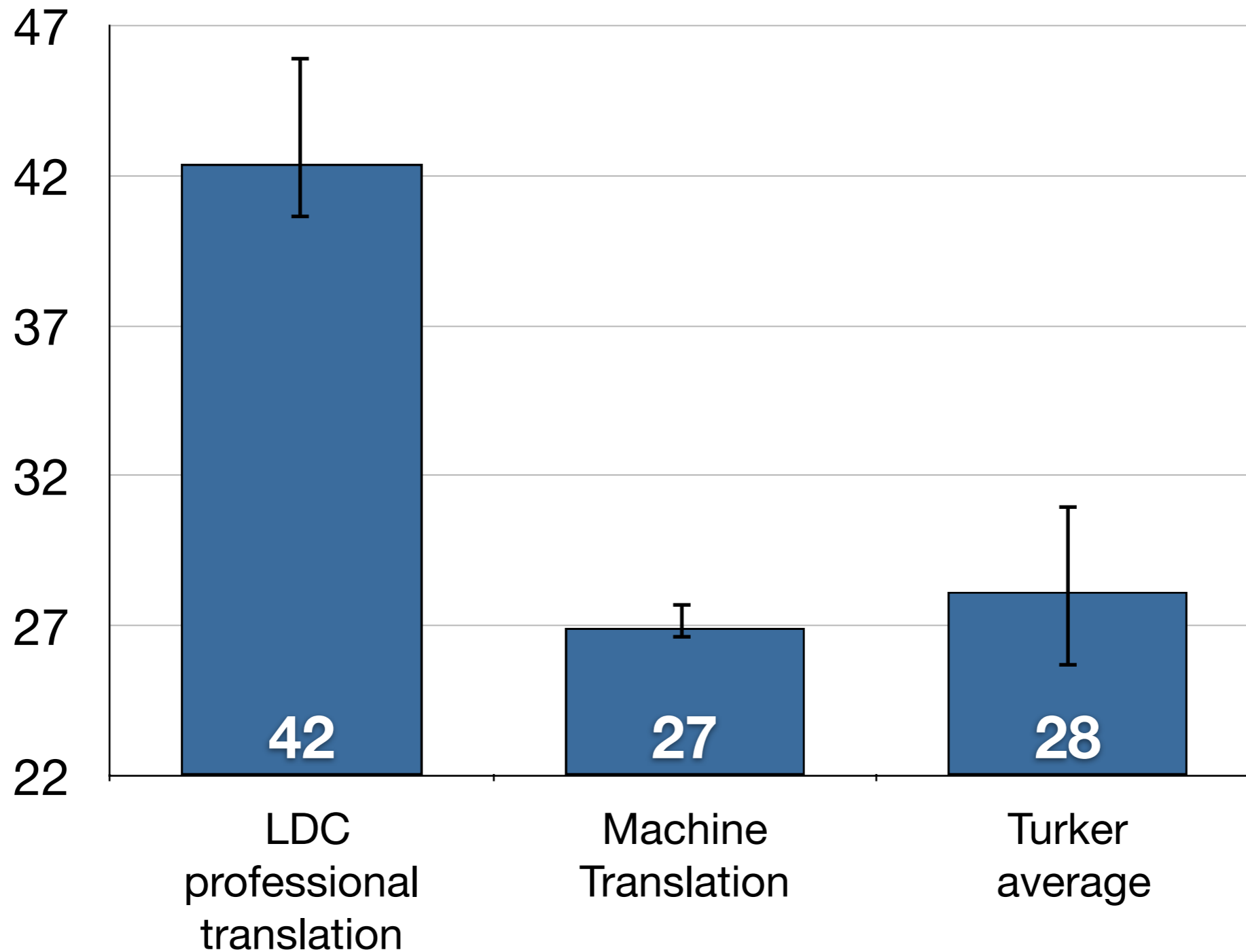
In 1994, the remains of  
early human beings who  
were believed to be eight  
hundred thousand years  
old were discovered who  
were given the name  
homo antecessor meaning  
the 'founder man'. Prior to  
this, the six hundred  
thousand years old man,  
called homo  
heidelebergensis in  
scientific terms, was  
believed to be the earliest  
resident of this area.

## Turk Translation

In 1994, the remains of  
pre-historic man, which  
are believed to be  
800,000 years old were  
discovered and they were  
named 'Home  
Antecessor' meaning 'The  
Founding Man'. Prior to  
that 6 lac years old  
humans, named  
as Homogenisens in  
scientific terms, were  
beleived to be the  
oldest dewellers of this  
area.

# Professionals versus Turkers\*

\*Without Quality Control



<p>Avoiding dieting to prevent from flu</p>	<p>abstention from dieting in order to avoid Flu</p>	<p>Abstain from decrease eating in order to escape from flue</p>	<p>In order to be safer from flu quit dieting</p>
<p>This research of American scientists came in front after experimenting on mice.</p>	<p>This research from the American Scientists have come up after the experiments on rats.</p>	<p>This research of American scientists was shown after many experiments on mouses.</p>	<p>According to the American Scientist this research has come out after much experimentations on rats.</p>
<p>Experiments proved that mice on a lower calorie diet had comparatively less ability to fight the flu virus.</p>	<p>in has been proven from experiments that rats put on diet with less calories had less ability to resist the Flu virus.</p>	<p>It was proved by experiments the low calories eaters mouses had low defending power for flue in ratio.</p>	<p>Experimentaions have proved that those rats on less calories diet have developed a tendency of not overcoming the flu virus.</p>
<p>research has proven this old myth wrong that its better to fast during fever.</p>	<p>Research disproved the old axiom that " It is better to fast during fever"</p>	<p>The research proved this old talk that decrease eating is useful in fever.</p>	<p>This Research has proved the very old saying wrong that it is good to starve while in fever.</p>

<p>Avoiding dieting to prevent from flu</p>	<p>abstention from dieting in order to avoid Flu</p>	<p>Abstain from decrease eating in order to escape from flue</p>	<p>In order to be safer from flu quit dieting</p>
<p>This research of American scientists came in front after experimenting on mice.</p>	<p>This research from the American Scientists have come up after the experiments on rats.</p>	<p>This research of American scientists was shown after many experiments on mouses.</p>	<p>According to the American Scientist this research has come out after much experimentations on rats.</p>
<p>Experiments proved that mice on a lower calorie diet had comparatively less ability to fight the flu virus.</p>	<p>in has been proven from experiments that rats put on diet with less calories had less ability to resist the Flu virus.</p>	<p>It was proved by experiments the low calories eaters mouses had low defending power for flue in ratio.</p>	<p>Experimentaions have proved that those rats on less calories diet have developed a tendency of not overcoming the flu virus.</p>
<p>research has proven this old myth wrong that its better to fast during fever.</p>	<p>Research disproved the old axiom that " It is better to fast during fever"</p>	<p>The research proved this old talk that decrease eating is useful in fever.</p>	<p>This Research has proved the very old saying wrong that it is good to starve while in fever.</p>

<p>Avoiding dieting to prevent from flu</p>	<p>abstention from dieting in order to avoid Flu</p>	<p>Abstain from decrease eating in order to escape from flue</p>	<p>In order to be safer from flu quit dieting</p>
<p>This research of American scientists came in front after experimenting on mice.</p>	<p>This research from the American Scientists have come up after the experiments on rats.</p>	<p>This research of American scientists was shown after many experiments on mouses.</p>	<p>According to the American Scientist this research has come out after much experimentations on rats.</p>
<p>Experiments proved that mice on a lower calorie diet had comparatively less ability to fight the flu virus.</p>	<p>in has been proven from experiments that rats put on diet with less calories had less ability to resist the Flu virus.</p>	<p>It was proved by experiments the low calories eaters mouses had low defending power for flue in ratio.</p>	<p>Experimentaions have proved that those rats on less calories diet have developed a tendency of not overcoming the flu virus.</p>
<p>research has proven this old myth wrong that its better to fast during fever.</p>	<p>Research disproved the old axiom that " It is better to fast during fever"</p>	<p>The research proved this old talk that decrease eating is useful in fever.</p>	<p>This Research has proved the very old saying wrong that it is good to starve while in fever.</p>

Avoiding dieting to prevent from flu	abstention from dieting in order to avoid Flu	Abstain from decrease eating in order to escape from flue	In order to be safer from flu quit dieting
--------------------------------------	---	---	--

Avoid		dieting		to		prevent	from	flu
Abstention	from	dieting	in order	to		avoid		Flu
Abstain		decrease eating	in order	to		escape		flue
quit		dieting		to	be	safer	from	flu



Avoiding dieting to prevent from flu	abstention from dieting in order to avoid Flu	Abstain from decrease eating in order to escape from flue	In order to be safer from flu quit dieting
--------------------------------------	---	---	--

Avoid		dieting		to		prevent	from	flu
Abstention	from	dieting	in order	to		avoid		Flu
Abstain		decrease eating	in order	to		escape	from	flue
quit		dieting		to	be	safer	from	flu

# Translate Urdu into English

Help us translate Urdu articles into English. Your translations will be distributed with a [Creative Commons license](#), so that other people can re-use it. This HIT is for people who speak both Urdu and English. Please **do not use** translation software or online machine translation systems like Google translate. Please make sure that your English translation:

- Does not add or delete any information from the original text
- Has the same meaning and style as the original
- Does not contain any spelling errors
- Is grammatical, natural-sounding English

First, please answer these questions about your language abilities:

- Is Urdu your native language?  Yes  No
- How many years have you spoken Urdu?  years
- Is English your native language?  Yes  No
- How many years have you spoken English?  years

افغانستان ایشیاء کا ایک ملک ہے جس کا سرکاری نام اسلامی جمہوریہ افغانستان ہے۔

اس کے جنوب اور مشرق میں پاکستان، مغرب میں ایران، شمال مشرق میں چین، شمال میں ترکمانستان، ازبکستان اور تاجکستان ہیں۔

اردگرد کے تمام ممالک سے افغانستان کے تاریخی، مذہبی اور ثقافتی تعلق بہت گہرا ہے۔

اس کے بیشتر لوگ مسلمان ہیں۔

۹۔ ملک بالتی تیب اب انہوں، ہونانہوں، عربوں، تہ کوں، منگولوں،

## 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 may 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.

Translation of the first sentence goes here.

Translation of the second sentence goes here.

# Translate Urdu into English

Informed Consent Form

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

you, this study may benefit human languages. This could be by searching, or new user

study.

participating at any time without penalty, or closing your browser

have adequate knowledge of the study, its purpose, and the instructions, or your answer

information kept about you will be kept confidential. This information may be used for research purposes.

to the principle of informed consent, you have been treated as a human being. University [Institutional Review Board](#)

states that you understand the study, you have read and understood the terms, you have waived any legal rights, and you have consented to participate in this research study.

```

<script language="Javascript">
<!--
function getUserInfo() {
    var userDisplayLanguage = navigator.language ? navigator.language : navigator.userDisplayLanguage;
    var country = sGeobytesCountry;
    var city = sGeobytesCity;
    var region = sGeobytesRegion;

    document.mturk_form.userDisplayLanguage.value = userDisplayLanguage;
    document.mturk_form.country.value = country;
    document.mturk_form.city.value = city;
    document.mturk_form.region.value = region;
}

getUserInfo();
// -->

```

Help us  
[Common](#)  
 and Eng  
 Google  
 • D  
 • H  
 • D  
 • Is  
 First, ple  
 Is Urdu  
 How ma  
 Is Engli  
 How ma

اس کے جنوب اور مشرق میں پاکستان، مغرب میں ایران، شمال مشرق میں چین، شمال میں ترکمانستان، ازبکستان اور تاجکستان ہیں۔

Translation of the second sentence goes here.

**Machine Translate**

اردگرد کے تمام ممالک سے افغانستان کے تاریخی، مذہبی اور ثقافتی تعلق بہت گہرا ہے۔

اس کے بیشتر لوگ مسلمان ہیں۔

۹۰ ملک بالتیب اب انہوں، ہونانہوں، عربوں، تہ کوہ، منگولوں،

# Quality Control Model

- **Sentence features**

- Language model probability
- Ratio of source / target sentence lengths
- Web n-gram match percentage
- Translation edit rate to other translators

- **Turker features**

- Aggregate of sentence feature scores
- Self-reported language abilities  
(Is native speaker? How long speaking?)
- Worker location (Pakistan? India?)

- **Ranking features** (based on second pass vote)

- **Calibration feature** (Bleu against professionals)

# Vote for the best translation

Please read the sentences and vote on the one that you think is the best in each group. The sentences are translations that were produced by people who are not native English speakers. Their translations are often ungrammatical, misspelled, disfluent, or bad in other ways. Your goal is to pick the best translation among the set. The one that you choose as the best will be forwarded on for editing, and it will undergo a variety of quality control mechanisms before it is published.

You should consider the following factors when selecting one translation as the best:

- Does it make more sense than the others?
- Is the English reasonably good?
- Do the grammar and spelling require only minimal correction?

<input type="radio"/>	Experimentaions have proved that those rats on less calories diet have developed a tendency of not overcoming the flu virus .
<input type="radio"/>	in has been proven from experiments that rats put on diet with less calories had less ability to resist the Flu virus.
<input type="radio"/>	Experiments proved that mice on a lower calorie diet had comparatively less ability to fight the flu virus.
<input type="radio"/>	It was proved by experiments the low calories eaters mouses had low defending power for flue in ratio.

<input type="radio"/>	The research proved this old talk that decrease eating is useful in fever.
<input type="radio"/>	Research disproved the old axiom that " It is better to fast during fever"
<input type="radio"/>	research has proven this old myth wrong that its better to fast during fever.
<input type="radio"/>	This Research has proved the very old saying wrong that it is good to starve while in fever.

<input type="radio"/>	According to the scientist a patient should eat more while in fever.
<input type="radio"/>	According to scientists, eat a lot during fever.
<input type="radio"/>	Eat and drink more in fever according to scientists.
<input type="radio"/>	according to the scientists one should eat a lot during fever.

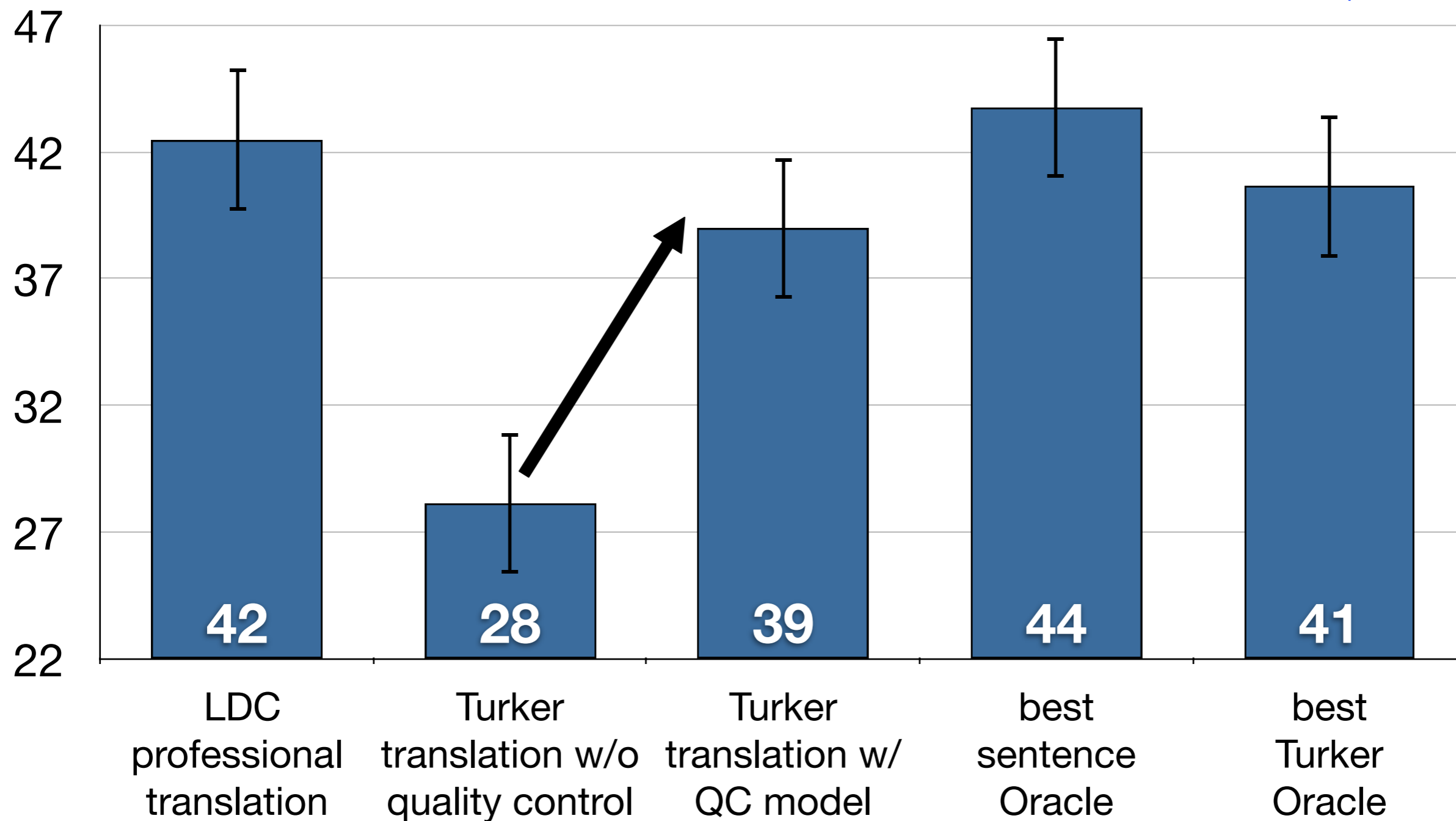
<input type="radio"/>	According to the Researchers of the State of Michigan University diet taking rats have less strength to fight with the infection and also the chances of death increased whereas earlier those rats on common diet had more strength to fight with the flu virus.
-----------------------	---

# Calibration data

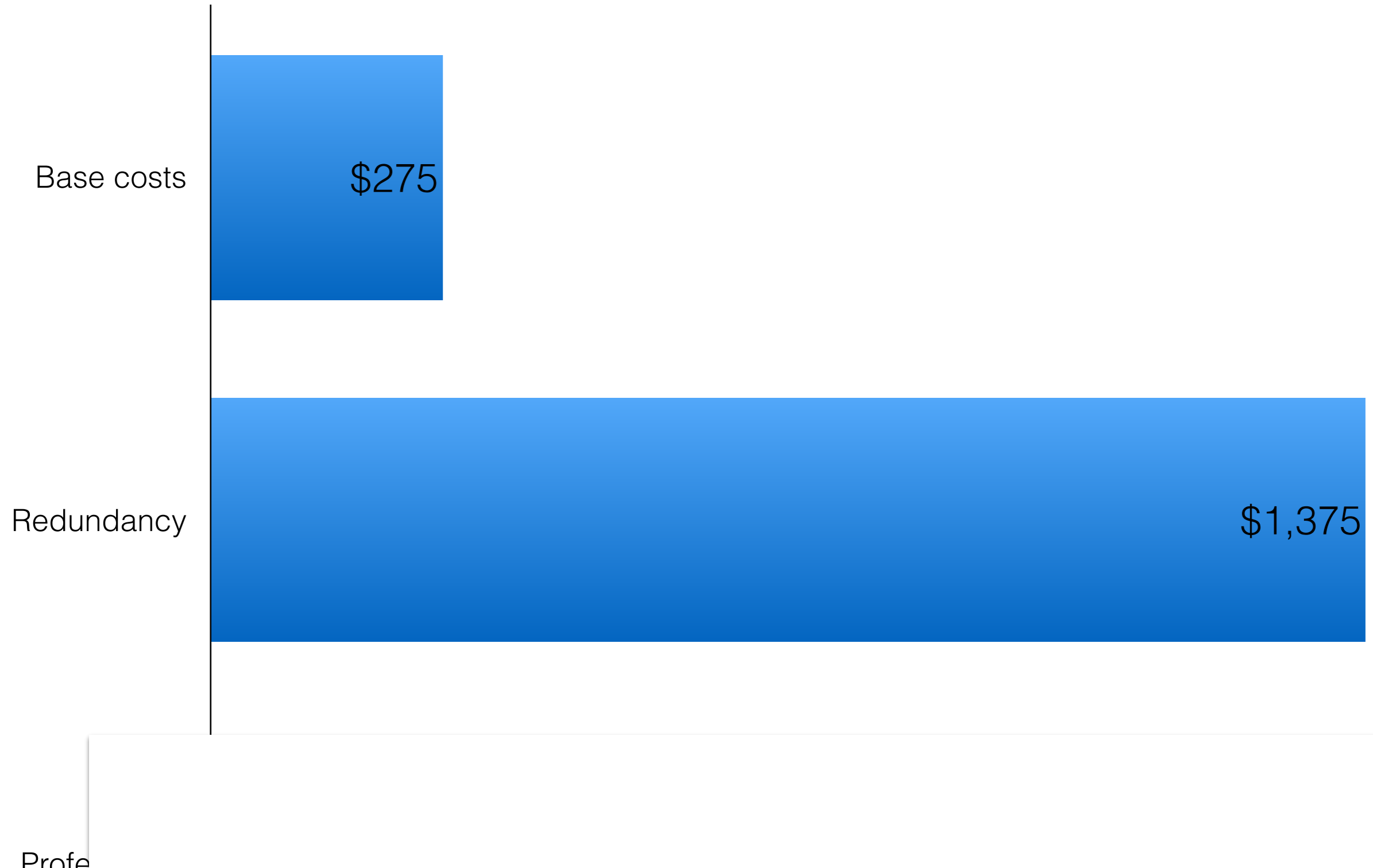
<b>Turker</b>	<b>Reference</b>
Avoiding dieting to prevent from flu	Avoid Dieting to Avoid Flu
This research of American scientists came in front after experimenting on mice.	This research of American scientists has come to light after experiments done on mice.
Experiments proved that mice on a lower calorie diet had comparatively less ability to fight the flu virus.	The experiments have proved that the capability of fighting Flu virus became less in mice placed on food containing fewer calories.
research has proven this old myth wrong that its better to fast during fever.	The old saying that,'It's better to starve during fever,'has been proven wrong by the research.

# Professional Quality from Non-Professionals

Full details in Zaidan and Callison-Burch (ACL 2011a) & Zaidan (PhD Thesis 2012)

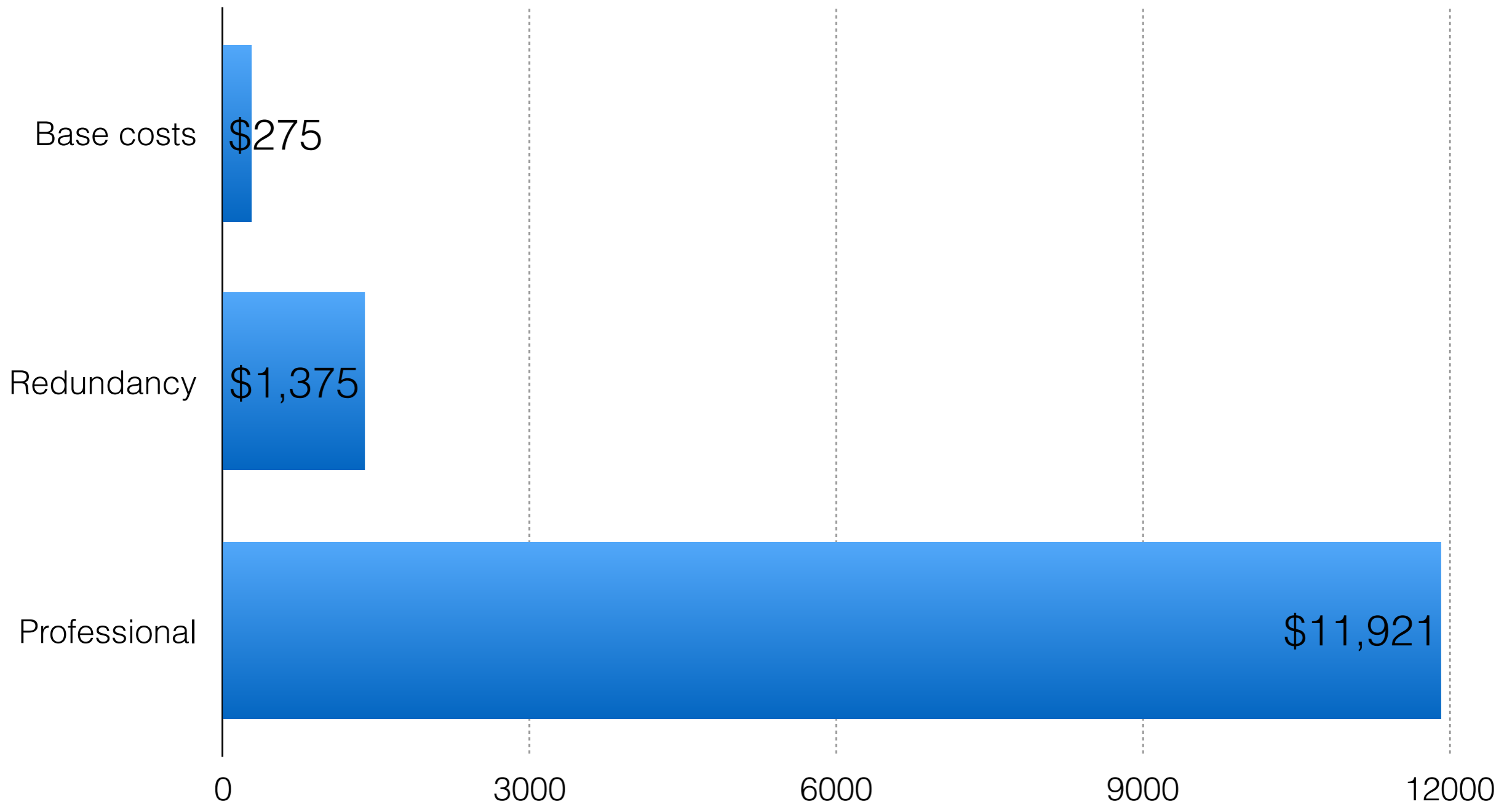


# Cost of Translation on MTurk





# Cost of Translation versus Professionals



# Cost of Translation on MTurk

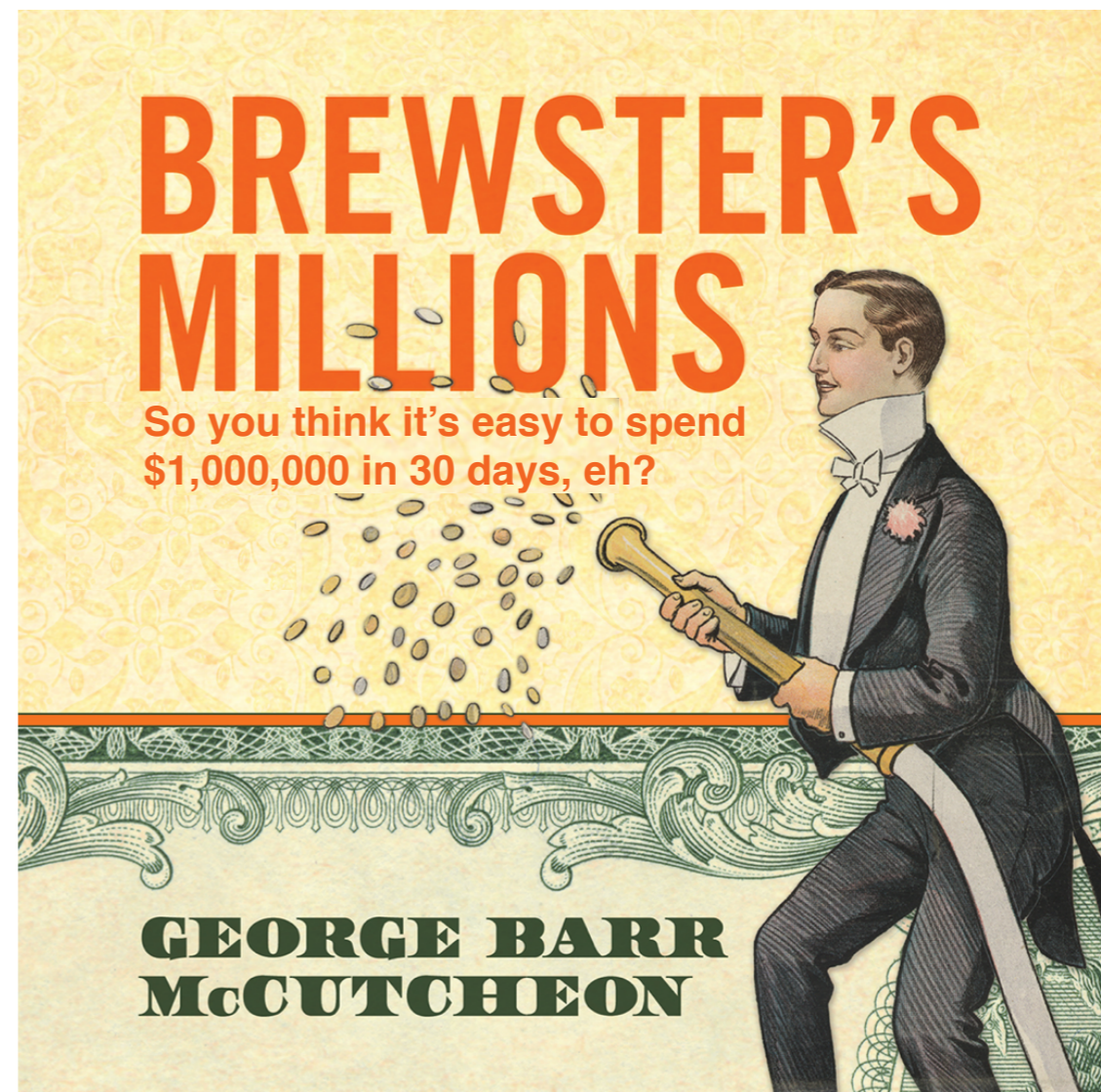
- Approximately an order of magnitude cheaper than the cost of professional translation
- Further reductions possible
  - reduce dependency on gold standard data
  - reduce amount of redundant translations collected
  - predict whether we need to get another translation
- It now seems feasible to collect enough data to train a statistical machine translation system

# \$100,000 challenge

Surely it can't scale. I bet you can't spend my money fast enough.



- Ken Church bet me \$50k that I couldn't spend \$50k on MTurk in two months



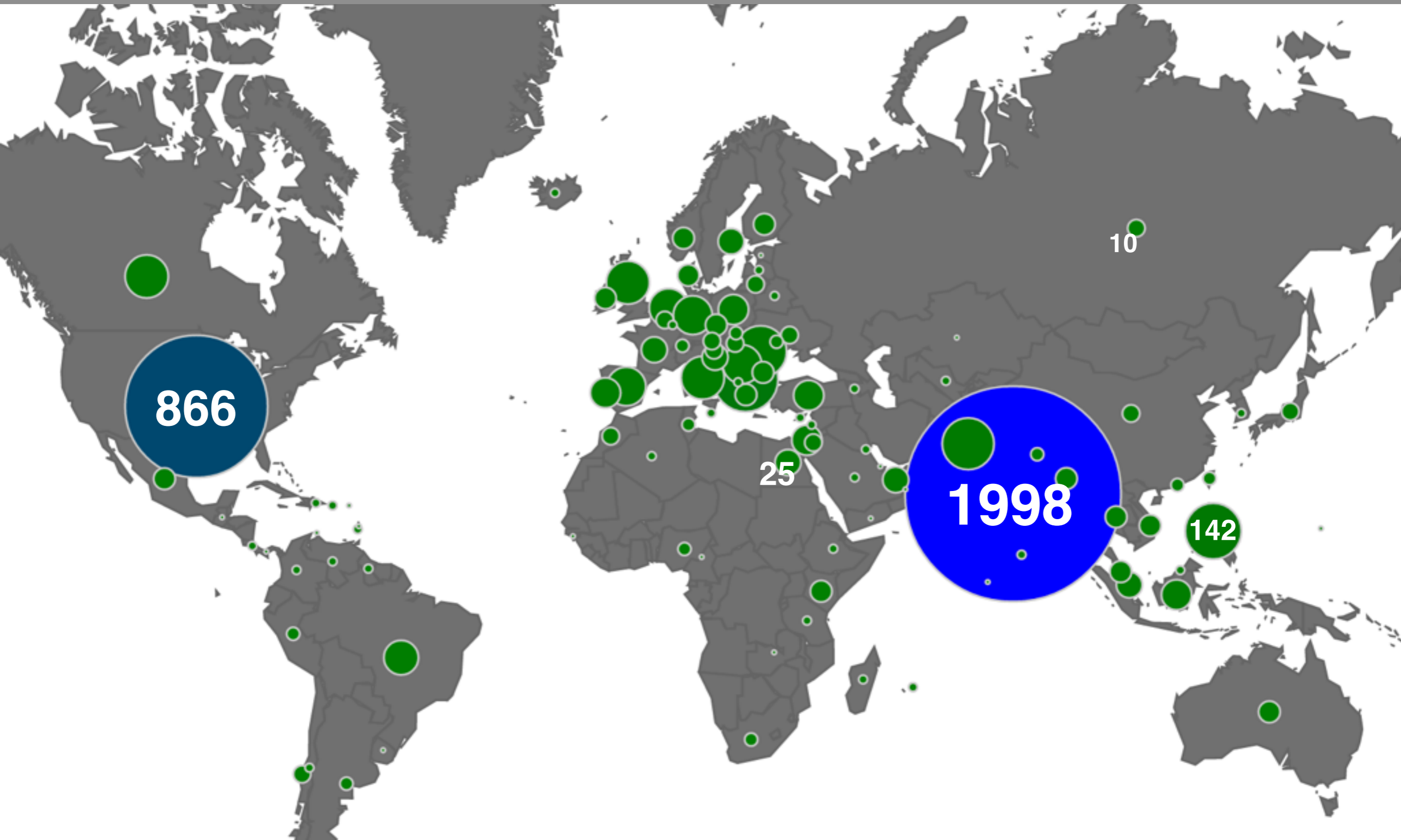
# Leaderboard

## Mechanical Turk Monitor: Top-1000 Recent Requesters

General Top requesters Arrivals Completed Search About

	Requester ID	Requester	#Task	#HITs	Rewards
1	<a href="#">Dolores Labs</a>	<a href="#">A2IR7ETVOIULZU</a> (RSS)	470	317857	33084.24
2	<a href="#">ContentGalore</a>	<a href="#">A2XL3J4NH6JI12</a> (RSS)	674	10622	17554.6
3	<a href="#">SpeechInk</a>	<a href="#">A1AQ7EJ5P7ME65</a> (RSS)	9019	13613	12876.67
4	<a href="#">CastingWords</a>	<a href="#">A3MI6MIUNWCR7F</a> (RSS)	9454	14036	8947.02
5	<a href="#">QuestionSwami</a>	<a href="#">AD7C0BZNKYGYV</a> (RSS)	629	4116	4750.37
6	<a href="#">Chris Callison-Burch</a>	<a href="#">A32TTE4XXN6MQZ</a> (RSS)	11	9961	4458.02
7	<a href="#">Smartsheet.com Clients</a>	<a href="#">A1197OGL0WOQ3G</a> (RSS)	434	38212	3118.28
8	<a href="#">retaildata</a>	<a href="#">AD14NALRDOSN9</a> (RSS)	8	50288	3110.85
9	<a href="#">Classify This</a>	<a href="#">A1CTI3ZAWTR5AZ</a> (RSS)	25	94590	1891.8
10	<a href="#">Andrew Stephen</a>	<a href="#">A1Y25F6MZCMQGY</a> (RSS)	3	22705	1131.25
11	<a href="#">Dolores Labs 2</a>	<a href="#">A3JX8WONBL5N9X</a> (RSS)	34	8976	1043.27
12	<a href="#">RelevanceQuest</a>	<a href="#">A8RMEN71ICE57</a> (RSS)	15	47881	1029.92
13	<a href="#">Crowd Task</a>	<a href="#">AFAOUHS65HNDS</a> (RSS)	4	2388	955.6
14	<a href="#">nlds.soe.ucsc.edu</a>	<a href="#">A1HI9DWCF794RE</a> (RSS)	4	4702	933.9

# Language Demographic Study



# Language Feasibility

workers	quality	speed	
many	high	fast	Dutch, French, German, Gujarati, Italian, Portuguese, Romanian, Serbian, Spanish, Tagalog, Telugu
		slow	Arabic, Hebrew, Irish, Punjabi, Swedish, Turkish
	medium or low	fast	Hindi, Marathi, Tamil, Urdu
		slow	Bengali, Bishnupriya Manipuri, Cebuano, Chinese, Nepali, Newar, Polish, Russian, Sindhi, Tibetan
few	high	fast	Bosnia, Croatian, Macedonian, Malay, Serbo-Croatian
		slow	Afrikaans, Albanian, Aragonese, Asturian, Basque, Belarusian, Bulgarian, Central Bicolano, Czech, Danish, Finnish, Galacian, Greek, Haitian, Hungarian, Icelandic, Ilokano, Indonesian, Japanese, Javanese, Kapampangan, Kazakh, Korean, Lithuanian, Low Saxon, Malagasy, Norwegian (Bokmal), Sicilian, Slovak, Slovenian, Thai, Ukranian, Uzbek, Waray-Waray, West Frisian, Yoruba
	medium or low	slow	Amharic, Armenian, Azerbaijani, Breton, Catalan, Georgian, Latvian, Luxembourgish, Neapolitian, Norwegian (Nynorsk), Pashto, Piedmontese, Somali, Sudanese, Swahili, Tatar, Vietnamese, Walloon, Welsh
none			Esperanto, Ido, Kurdish, Persian, Quechua, Wolof, Zazaki

# Parallel Corpora for 6 Indian Languages

पाकिस्तान ने भी प्रतिकार स्वरूप २८ मई १९९८ में छह परमाणु परीक्षण कर डाले।

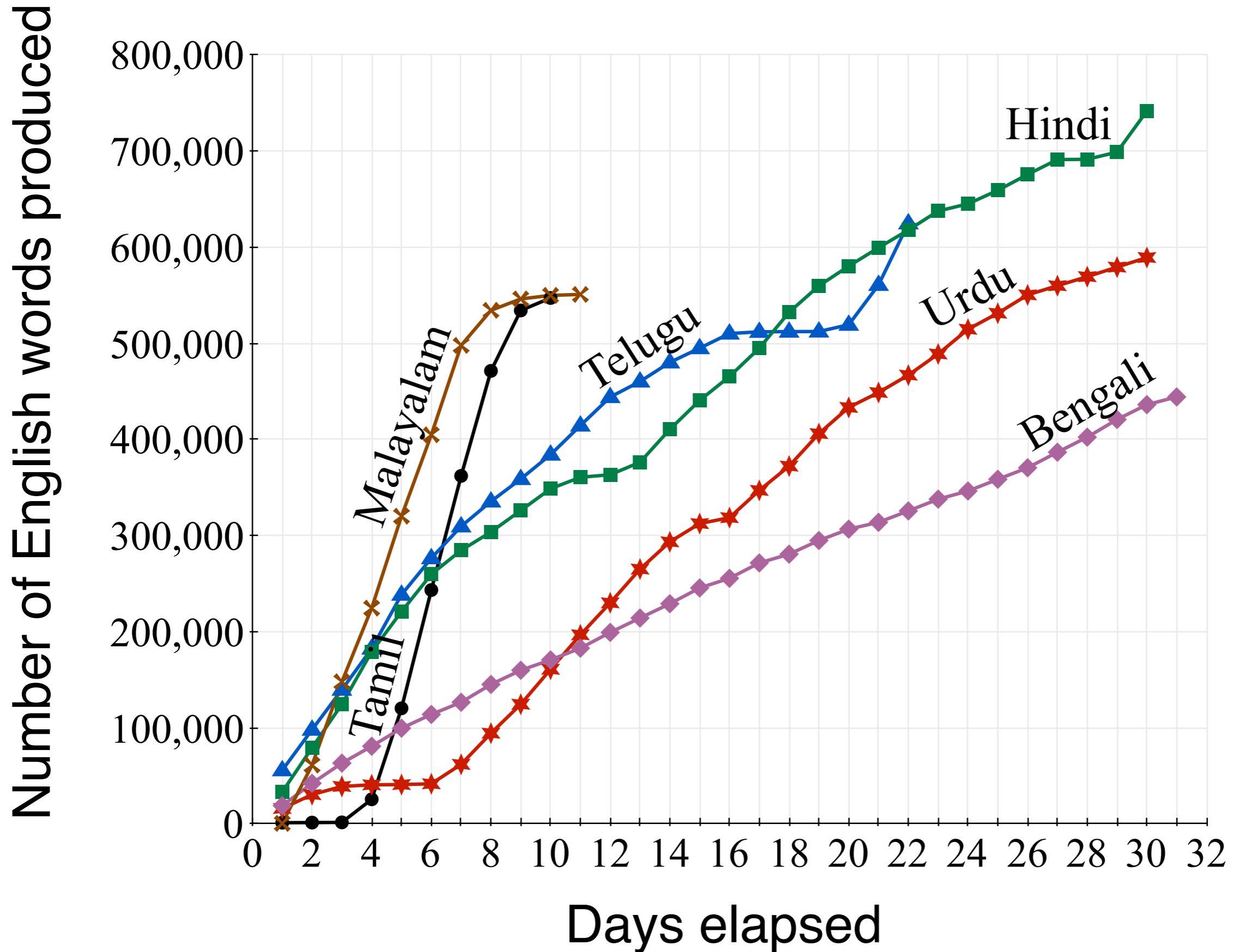
In retribution pakistan also did six nuclear tests on 28 may 1998.

On 28 May Pakistan also conducted six nuclear tests as an act of redressal.

Retaliating on this 'Pakistan' conducted Six(6) Nuclear Tests on 28 May, 1998.

pakistan also did 6 nuclear test in retribution on 28 may, 1998

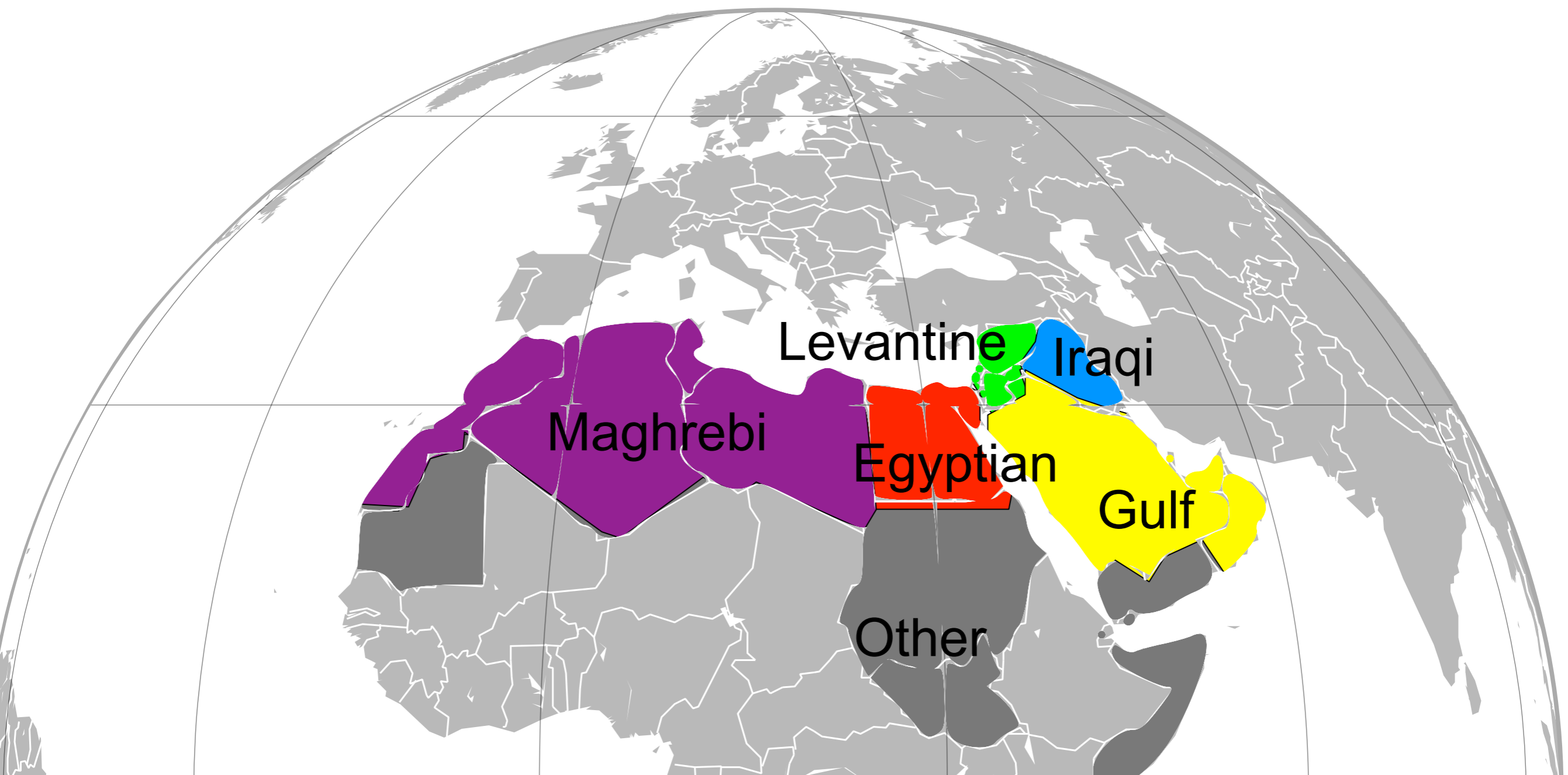
# Rate of Translation





# Gathering Data about Arabic Dialects

Arabic has **different varieties**. MSA is the standardized form but there are many **distinct** regional dialects.



# Translating Dialects with MSA MT

MSA:

متى سنرى هذه التلة من المجرمين تخضع للمحاكمة ؟

*mtY snrY h\*h Alvlp mn Almjrmyn txDE llmHAkmp ?*

Levantine:

ايتمى رح نشوف هالشلة من المجرمين بتتحاكم ؟

*AvmtY rH n\$wf hAl\$lp mn Almirmvn bttHAkm ?*

PT: Quando veremos esse grupo de criminosos serem julgados?

ES-EN: **Quando esse** group of criminals see **Serem julgados?**

# Written Arabic Dialect

## بلخادم: ما يجمع مصر والجزائر يجعل التقارب أكثر من ضروري

الخميس، 16 يونيو 2011 - 09:27



عبد العزيز بلخادم الممثل الشخصي للرئيس الجزائري

### الجزائر (أ.ش.أ.)

Google

أكد عبد العزيز بلخادم الممثل الشخصي للرئيس الجزائري والأمين العام لحزب جبهة التحرير الوطني ذات الأغلبية في البرلمان والحكومة، أن العلاقات بين مصر والجزائر قديمة ومتجذرة علاوة على مصالح مشتركة قوية تربط بين البلدين.

وقال بلخادم إنه بالتالي فإن ما حدث من "سحابات الصيف" عقب مباراة كرة القدم بين فريقى البلدين فى تصفيات كأس العالم عام 2009 لا يؤثر على طبيعة العلاقة والأخوة والروابط بين الشقيقين، مشددا على ضرورة العمل على رفع سقف التعاون بين البلدين على المستوى الحكومى والجهاميرى والمؤسساتى

Article body (MSA)

### 1 خليك فى حالك وابعد عننا

بواسطة: bebooo

بتاريخ: الخميس، 16 يونيو 2011 - 09:39

EGY

خليك فى حالك انت وابعدوا عننا واتم اخر ناس تتكلم عن التقارب والعروبة

### 2 الجزائر إخوه أعزاء لنا وأهلاً ومرحباً بتقوية العلاقات معهم

بواسطة: سامي شاهين

بتاريخ: الخميس، 16 يونيو 2011 - 09:55

MSA

التقارب بين مصر والجزائر يعتبر حاجة ملحة تستدعيه الظروف الراهنة التي يمر بيها الوطن العربي ، أهلاً ومرحباً بأي تقارب عربي عربي بين الأشقاء في الوطن والمصير ، يسقط الاستعمار الجديد المتمثل في أمريكا والنانو ومن ورائة ال صهيون

### 3 هزلت

بواسطة: مواطن مصري بسيط

بتاريخ: الخميس، 16 يونيو 2011 - 09:55

EGY

هزلت وعيلت ومشت سفينة نوح علي اليابس - الدنيا جري فيها ايه علشان نسمع دروس من امثالكم

### 4 تنبيه

بواسطة: علي

بتاريخ: الخميس، 16 يونيو 2011 - 09:57

MSA

انا الشعب الجزائري يعمل جاهدا من اجل السعي لتوحيد العالم العربي وكذلك السعي ورا طموحات اكبر وهي ، ان تكون كلمت العرب كلما واحدة امام الغرب

### 5 أنت.....

بواسطة: الفيلسوف

بتاريخ: الخميس، 16 يونيو 2011 - 10:14

EGY

أنت بتكلم مين يا بلخادم كل الأنظمة فاسدة و تابعة لأمريكا و إسرائيل، إن أردتم فعلا حل مشاكل العرب في البيت العربي فما على الحكام إلا تكوين جمعية و الذهاب لمنتج الحكام المخلوعين بالمملكة السعودية وسيبوا الشعوب تبحث في قيام الولايات العربية المتحدة و عاصمتها القدس إنشاء الله (معلش حبيبي هو حلم يقظة يمكن يتحقق في يوم من الأيام من منا يعلم الغيب؟)

MSA

EGY

# Crowdsourcing Dialect ID

Full details in Zaidan and Callison-Burch (ACL 2011b)

- **Labeled 142k comments** gathered from three online newspapers:
  - Al-Ghad (الغد), a Jordanian newspaper **LEV** **MSA**
  - Al-Riyadh (الرياض), a Saudi newspaper **GLF** **MSA**
  - Al-Youm Al-Sabe' (اليوم السابع), an Egyptian newspaper **EGY** **MSA**
- 59% MSA, **41% dialect**
- Trained classifiers with 80-90% accuracy

# Crowdsourcing Arabic Dialects

- **Translated** dialect-labeled segments EGY LEV

<b>Dialect Classification HIT</b>	\$10,064
<b>Sentence Segmentation HIT</b>	\$1,940
<b>Translation HIT</b>	\$32,061
<b>Total Cost</b>	\$44,065
<b>Num words translated</b>	1,516,856
<b>Cost per word</b>	2.9 cents/word

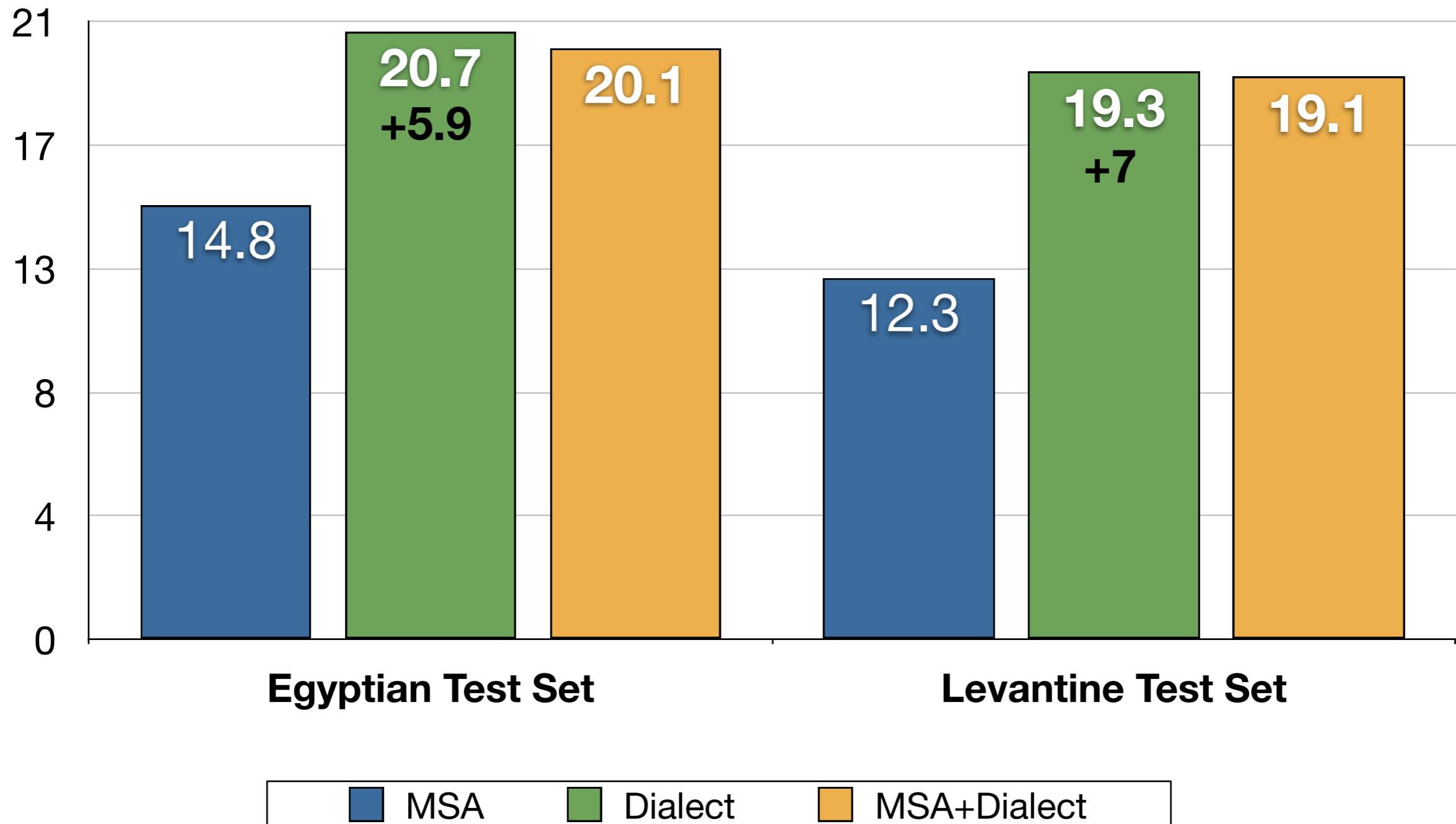
- **121 workers** completed 20+ assignments
- **200,000 words** translated **per week**
- **Trained** BBN's machine translation system

# Examples of Dialect Translation

Dialect Input	MSA system	Dialect system	Reference
<b>EGY</b> انت بتعمل له اعلان ولا !! ايه ؟	You are working for a declaration and not?	You are making the advertisement for him or what?	Are you promoting it or what?!!
<b>EGY</b> نفسي اطمئن عليه بعد ما شاف الصوره دي	Myself feel to see this image.	I wish to check on him after he saw this picture.	I want to be sure that he is fine after he saw the images.
<b>LEV</b> لهيك الجو كتيير كووول	God you the atmosphere.	This is why the weather is so cool	This is why the weather is so cool
<b>LEV</b> طول بالك عم نمزح	Do you think about a joke long.	Calm down we are kidding	Calm down, we are only kidding

# Dialect versus MSA

Full details in Zbib, Malchiodi, Devlin, Stallard, Matsoukas, Schwartz, Makhoul, Zaidan, & Callison-Burch (NAACL 2012)



# Crowdsourcing for the win

Crowdsourcing dramatically improved MT for Arabic dialects

