Crowdsourcing Translation

Chris Callison-Burch
March 24, 2015

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

amazon mechanical turk
beta
Artificial Artificial Intelligence

Mechanical Turk is a marketplace for work.
We give businesses and developers access to an on-demand, scalable workforce.
Workers select from thousands of tasks and work whenever it's convenient.

37,649 HITs available. <u>View them now.</u>

Make Money
by working on HITs

HITs - *Human Intelligence Tasks* - are individual tasks that you work on. <u>Find HITs now.</u>

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work

Find an interesting task   **Work**   **Earn money**

[Find HITs Now]

or <u>learn more about being a **Worker**</u>

Get Results
from Mechanical Turk Workers

Ask workers to complete HITs - *Human Intelligence Tasks* - and get results using Mechanical Turk. <u>Register Now</u>

As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results

Fund your account   **Load your tasks**   **Get results**
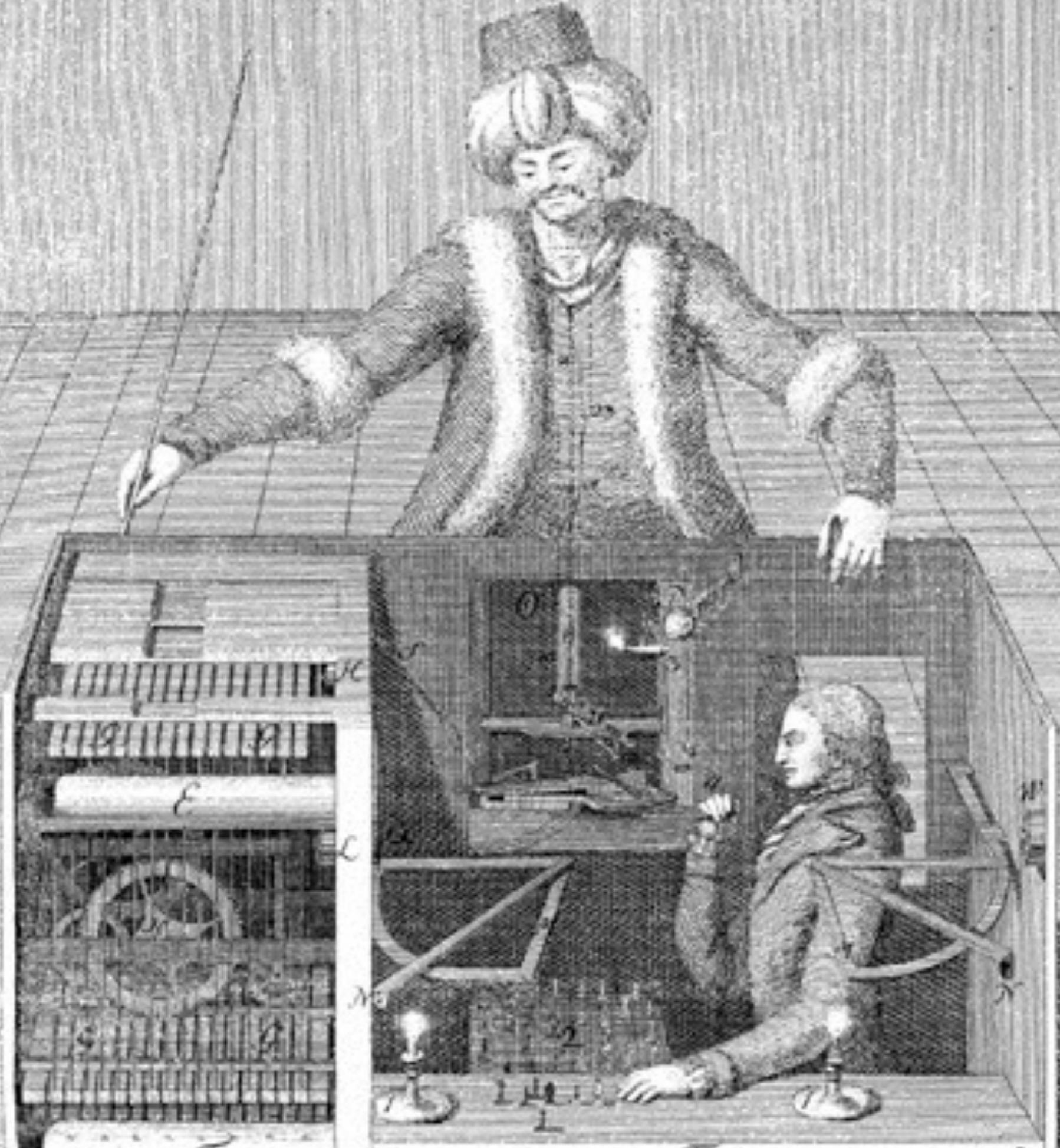
[Get Started]

or <u>learn more about being a **Requester**</u>

amazon**mechanical turk**
beta   Artificial Artificial Intelligence

Sign In

Your Account   **HITs**   **Qualifications**

37,530 HITs available now

All HITs | HITs Available To You | HITs Assigned To You

Search for [HITs ▲▼] containing [_____] that pay at least $ [0.00] for which you are qua

Timer: 00:00:00 of 5 minutes Want to work on this HIT?   Want to see other HITs?

[Accept HIT]   [Skip HIT]

Total Earned: Unavailal
Total HITs Submitted: 0

Enter Postmark & Stamp Information for a Postcard
Requester: Cardcow   Reward: $0.01 per HIT   HITs Available: 2   Duration: 5 minutes
Qualifications Required: Data Entry for Postcards has been granted

Enter Postmark & Stamp Information for this card

Postmark City:

Postmark State: (or Country)

Postmark Date: (Ex: Nov-09)

Postmark Year: (Ex: 1909)

Stamp: (Ex: 1c, 2c, half penny)

[_____]   [_____]   [_____]   [_____]   [_____]

(month & day)

Sign In

amazonmechanical turk
Artificial Artificial Intelligence

Your Account   HITs   Qualifications   37,530 HITs available now

All HITs | HITs Available To You | HITs Assigned To You

Search for   HITs ▼   containing _____   that pay at least $ 0.00   for which you are qu

Timer: 00:00:00 of 5 minutes Want to work on this HIT?   Want to see other HITs?   Total Earned: Unavail

Accept HIT   Skip HIT   Total HITs Submitted: 0

Enter Postmark & Stamp Information for a Postcard

Requester: Cardcow   Reward: $0.01 per HIT   HITs Available: 2   Duration: 5 minutes

s been granted

or this card

Postmark City: Barre
Postmark State: MA
Postmark Date: Oct–11
Postmark Year: 1886
Stamp: 1c

$ 0. 01

Postmark City: Postmark State: (or Country)   Postmark D (Ex: Nov-09   1c, 2c,

(month & day)

Who are the Turkers?

- Requesters are given very little information about Turkers - basically just a serial number

- No names, no demographic information (like what languages they speak)

- Cannot assume that they have a particular set of skills

- They should be treated as non-experts

- Quality control is a major challenge

- It important to design tasks to be simple and easy to understand

MTurk for Natural Language Processing

Snow, O'Connor, Jurafsky and Ng's EMNLP 2008 paper pioneered the use of Mechanical Turk for NLP

- Affect Recognition

  fear("Tropical storm threatens NYC") > fear("Awesome goal for Beckham")

- Word Similarity

  sim(man, boy) > sim(man, rooster)

- Textual Entailment

  if "Microsoft was established in Italy in 1985" then "Microsoft was established in 1985"?
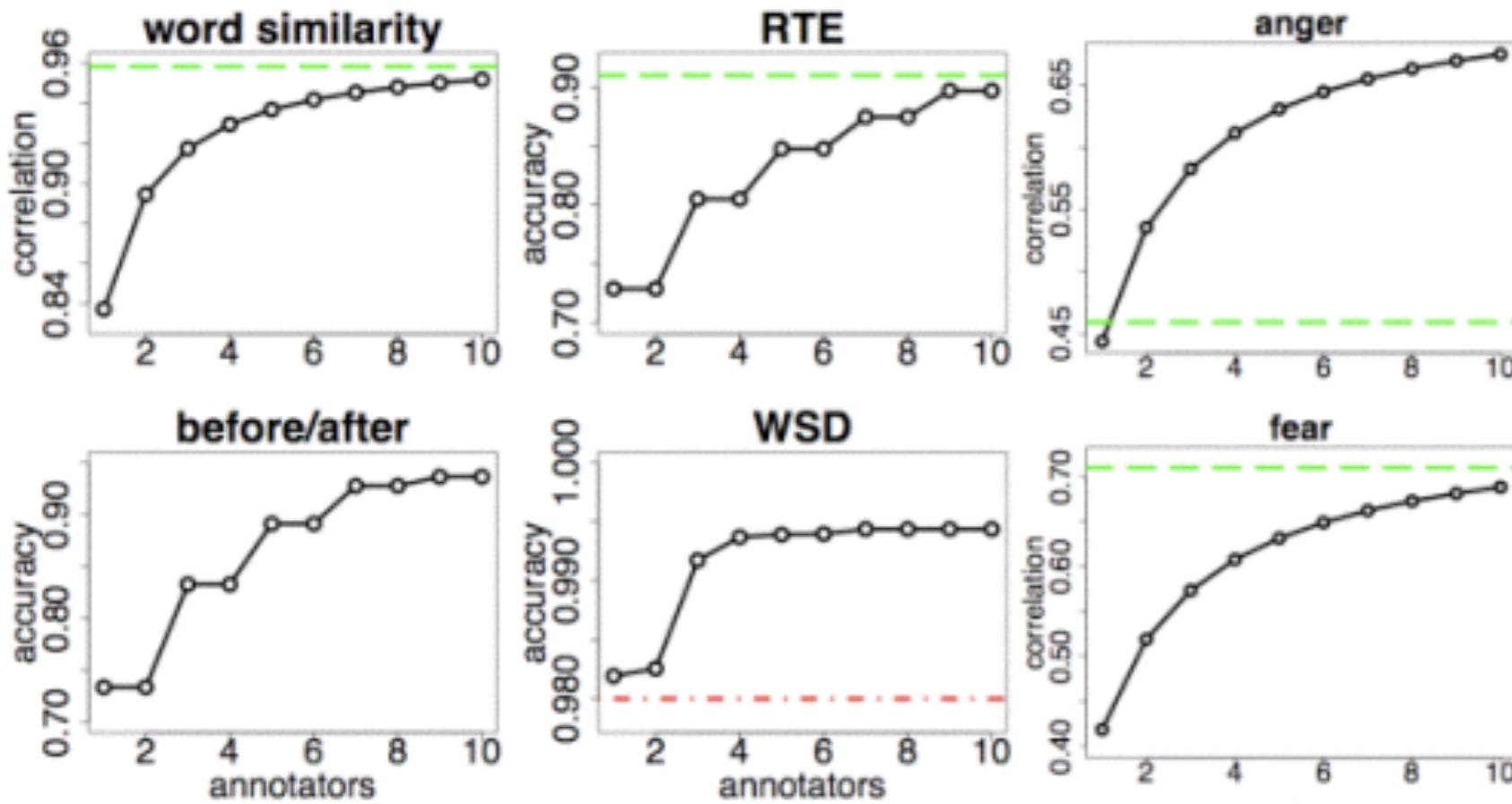
- Word Sense

  "the West Bank" v. "the Bank of America"
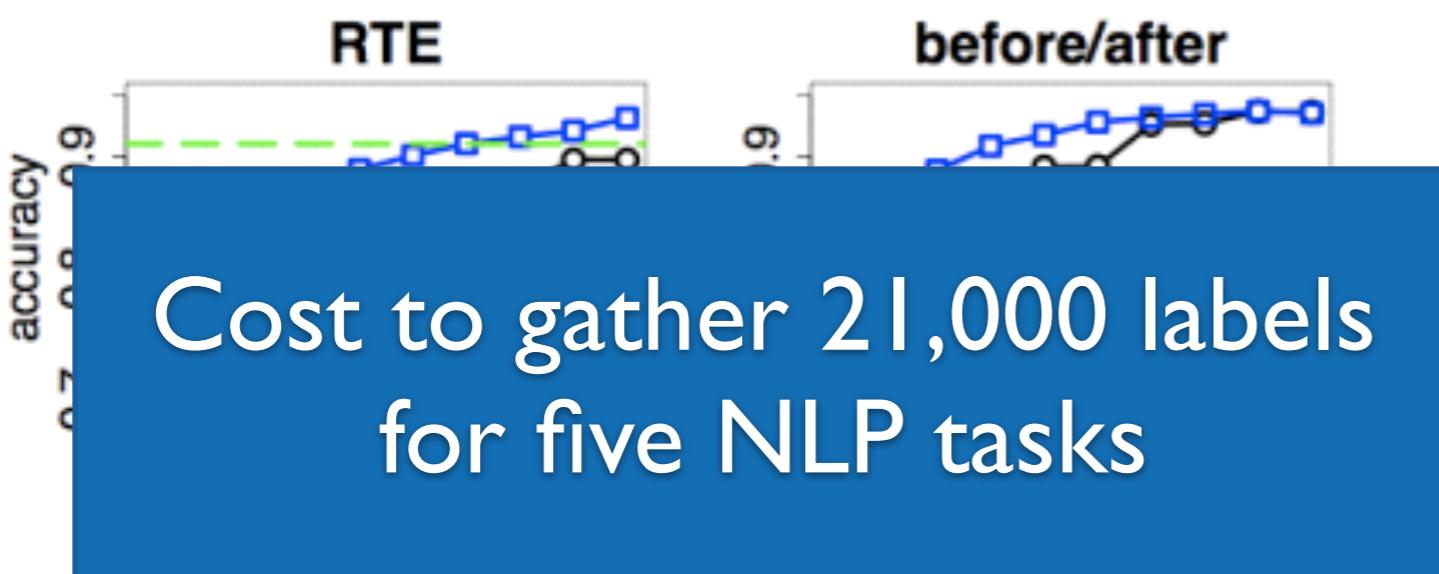
- Temporal Annotation

  denoted happens before *collapsed* in: "The condemned building collapsed when the crew detonated the charge."

NLP Annotation

Combine non-expert judgments for high correlation with experts

Weight w/ small amount of gold standard data for

Cost to gather 21,000 labels for five NLP tasks

$ 25. 82

Other NLP applications

- Workshop on Using Mechanical Turk for Speech and Language Applications

- 35 researchers spent $100, wrote papers, and distributed their data

- Mark Dredze and I wrote an overview paper digesting the results
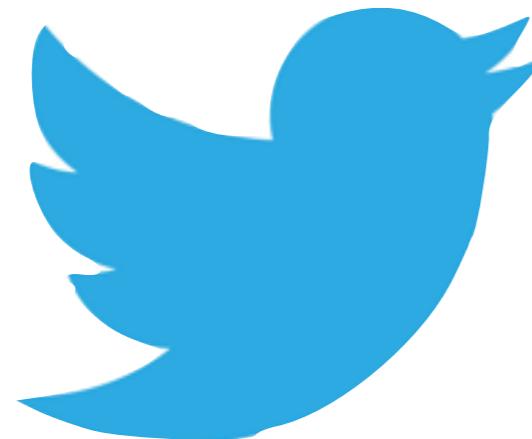
$100 Challenge

- **Traditional NLP tasks**
  - WSD, RTE, NLG, common sense knowledge
- **Speech and Vision**
  - Transcribed speech, accented speech, handwriting OCR
- **Sentiment, Polarity, Bias**
  - Cross language, blogs

- **Information Retrieval**
  - TREC style annotations
- **Information Extraction**
  - Relation extraction, NER
- **Machine Translation**
  - Paraphrases, alignments, training and eval sets, rule cleaning

Statistical machine translation

- Translation rules are learned bilingual texts using machine learning techniques

Arabic

| |
|---|
| فالتعذيب لا يزال يمارس على نطاق واسع |
| وتتم عمليات الاعتقال والاحتجاز دون سبب بصورة روتينية |
| وحان وقت التحلى بالبصيرة والشجاعة السياسية . |
| . . . |

English

| |
|---|
| Torture is still being practised on a wide scale. |
| Arrest and detention without cause take place routinely. |
| This is a time for vision and political courage |
| . . . |

Chinese

我国 能源 原材料 工业 生产 大幅度 增长 .
非国大 要求 阻止 更 多 被 拘留 人员 死亡 .
. . .
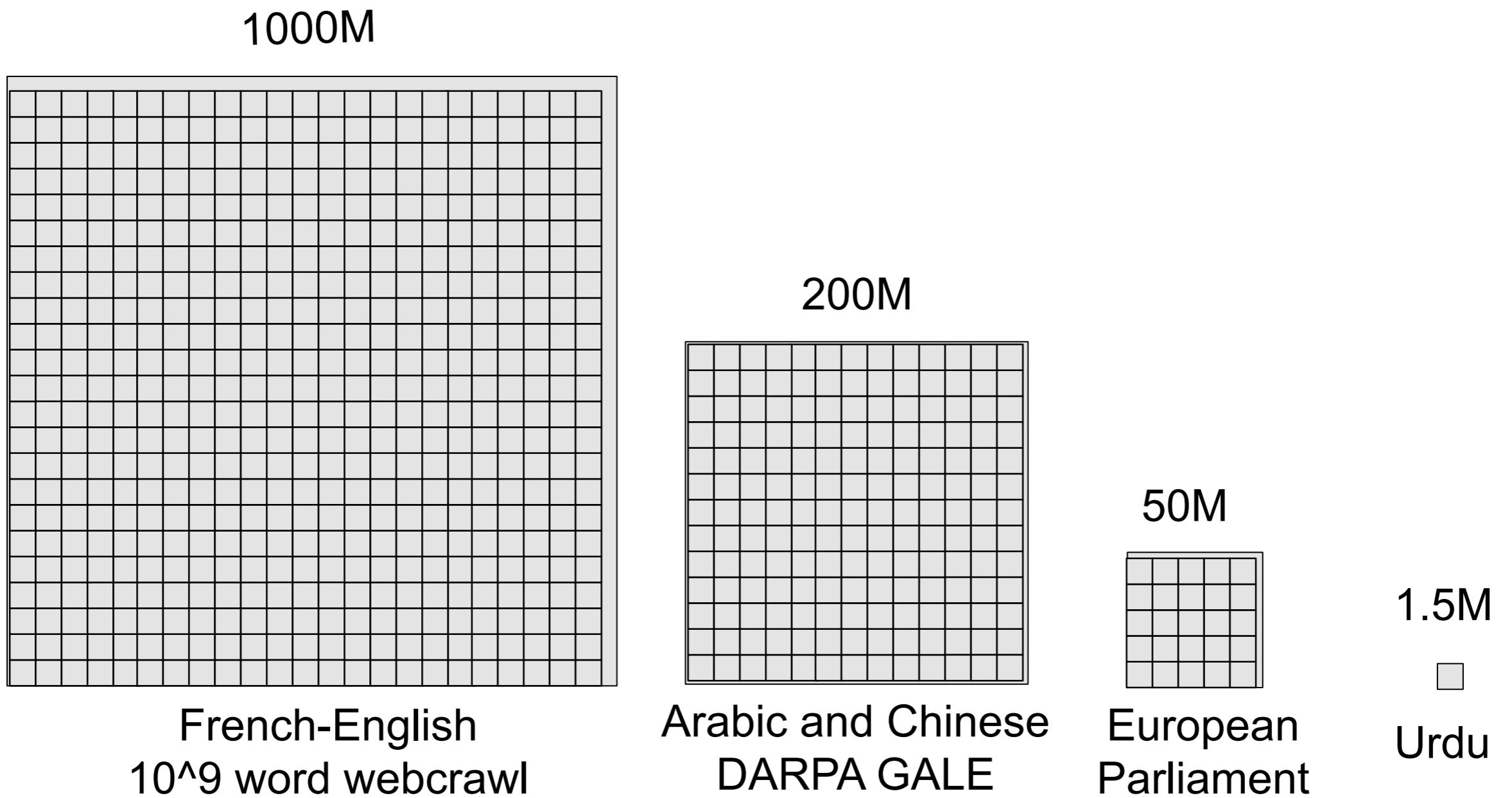
English

China's energy and raw materials production up.
ANC calls for steps to prevent deaths in police custody .
. . .
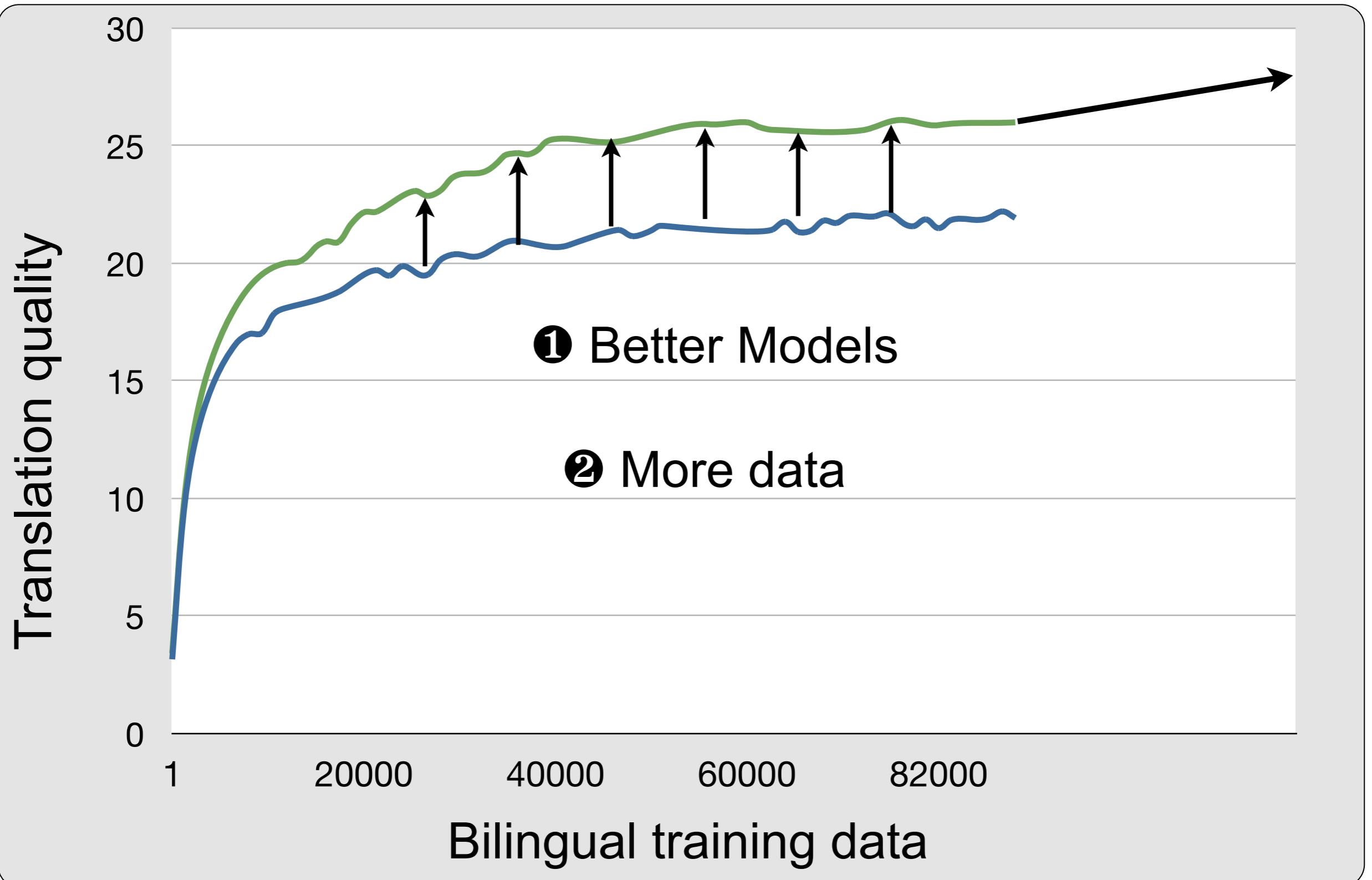
Training data vaires by language

1000M

200M

50M

1.5M

French-English
10^9 word webcrawl

Arabic and Chinese
DARPA GALE

European
Parliament

Urdu

How to Improve Machine Translation

Can Turkers Translate?

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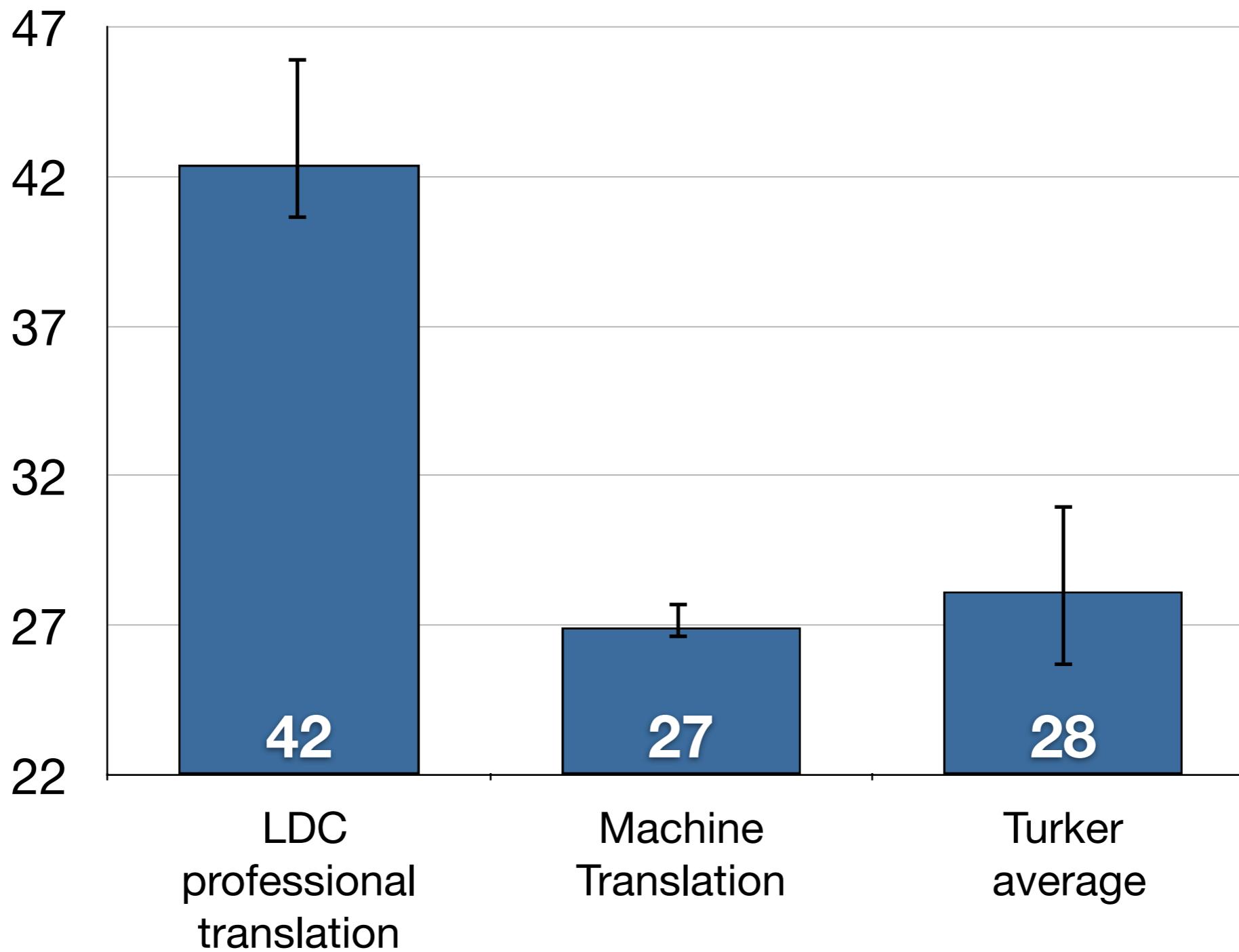
Avoiding dieting to prevent from flu	abstention from dieting in order to avoid Flu	Abstain from decrease eating in order to escape from flue	In order to be safer from flu quit dieting
This research of American scientists came in front after experimenting on mice.	This research from the American Scientists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

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

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

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۔

Translation of the first sentence goes here.

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

Translation of the second sentence goes here.

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

اس کے بیشتر لوگ مسلمان ہیں۔

یہ ملک بالترتیب ایرانیوں، یونانیوں، عربوں، ترکوں، منگولوں،

Translate Urdu into English

Help us...

```
<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();
// -->
```

Informed Consent Form

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

...ou, this study may benef
...uman languages. This cou
...b searching, or new user

...s study.

...pating at any time withou
..., or closing your browser

... have adequate knowledg
...structions, or your answer

...ion kept about you will b
... This information may be

...s to the principle
...ou have been treated
...niversity Institutional

...es that you understand th
... waived any legal rights
...esearch study.

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

Translation of the second sentence goes here.

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

اس کے بیشتر لوگ مسلمان ہیں۔

یہ ملک بالترتیب اردنوں اور انیوں، یونانیوں، عربوں، ترکوں، منگولوں،

Machine Translate

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 pe who are not native English speakers. Their translations are often ungrammatical, misspelled, disfluent, or bad in other ways. Your goal is to tr 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 o 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?

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

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

- ○ According to the scientist a patient should eat more while in fever.
- ○ According to scientists, eat a lot during fever.
- ○ Eat and drink more in fever according to scientists.
- ○ according to the scientists one should eat a lot during fever.
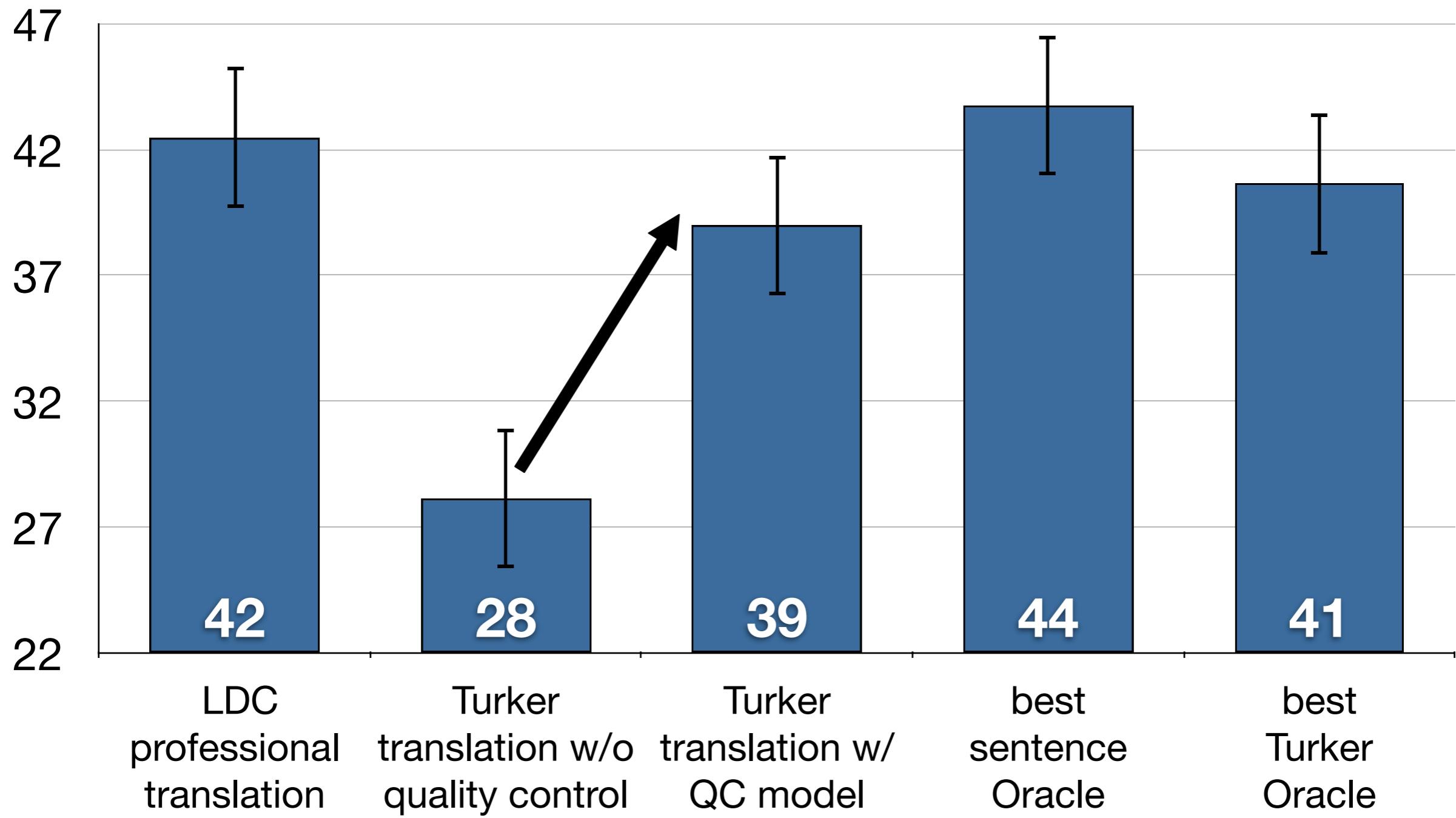
- ○ According to the Researchers of the State of Michigan University diet taking rats have less strength to fight with the infection and also the chances of death increased whereas earlier those rats on common diet had more strength to fight with the flu virus.

Calibration data
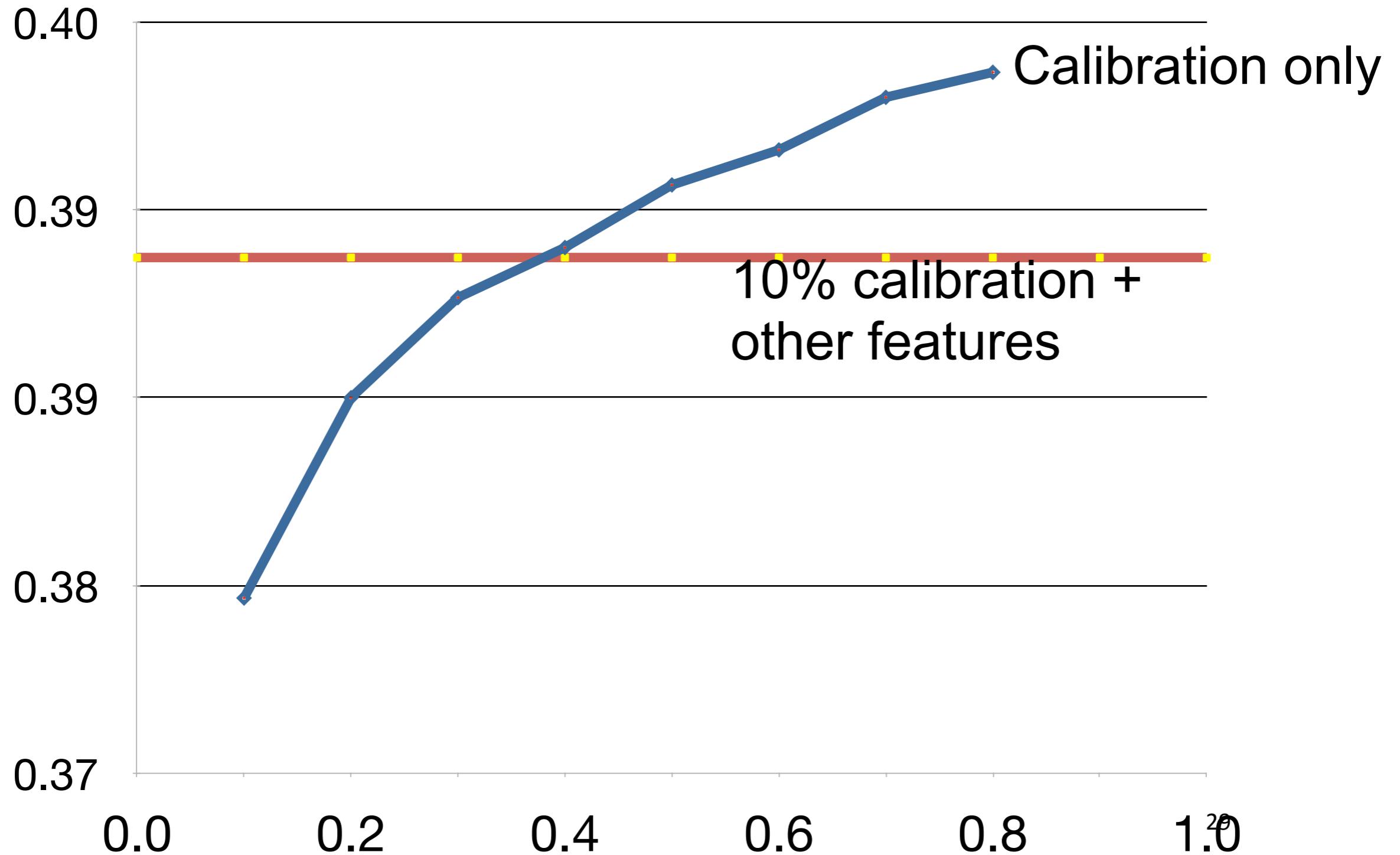
Turker	Reference
Avoiding dieting to prevent from flu	Avoid Dieting to Avoid Flu
This research of American scientists came in front after experimenting on mice.	This research of American scientists has come to light after experiments done on mice.
Experiments proved that mice on a lower calorie diet 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)

LDC professional translation: 42
Turker translation w/o quality control: 28
Turker translation w/ QC model: 39
best sentence Oracle: 44
best Turker Oracle: 41

% of gold stadard data used for calibration

Sri Lanka 's forest region was destroyed by agriculture , ~~wooden works~~ , ~~vetinary feeds~~ , etc . ,

forestry   animal grazing

several commissions ~~where~~ created to protect the remaining forest region

were

Sri Lanka is considered ~~as the~~ bird 's sanctionary ~~place~~ .

to be a

For further information please see the article on bird sanctionary rights in Indian Subcontinent

the
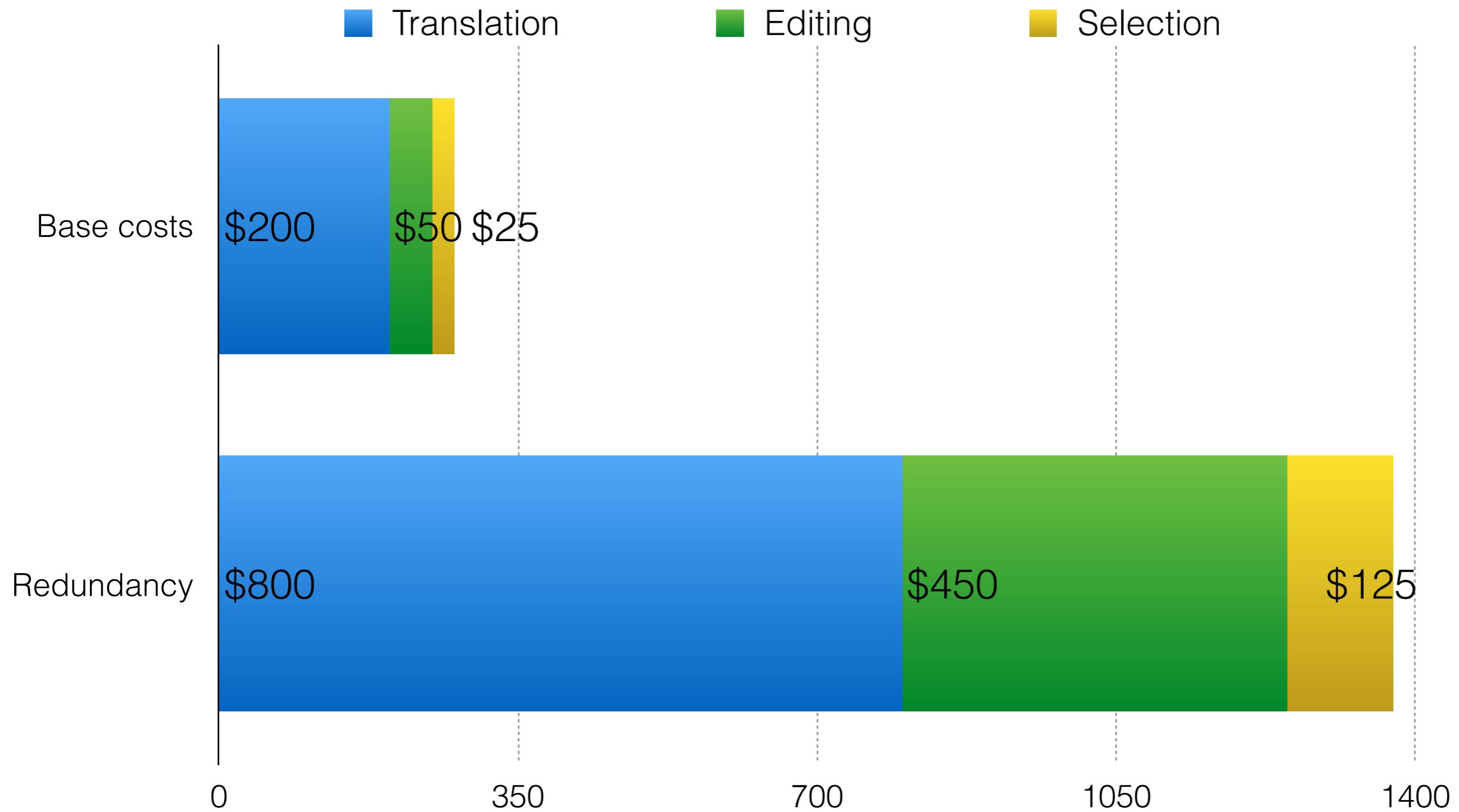
There is thousand of animals living in Sri Lanka which includes several Sri Lanka orignated animals .

When we compare the area of Sri lanka 's Island , birds are highly found here .

Quality wrap up

- Crowdsourced translations can reach high quality after quality control
  - Gather redundant translations
  - Calibrate against small amount of professional translations
  - Do second passes over the data where other Turkers select best translations
  - Post-edit the non-native translations
   - In the future: explicitly deal with ESL
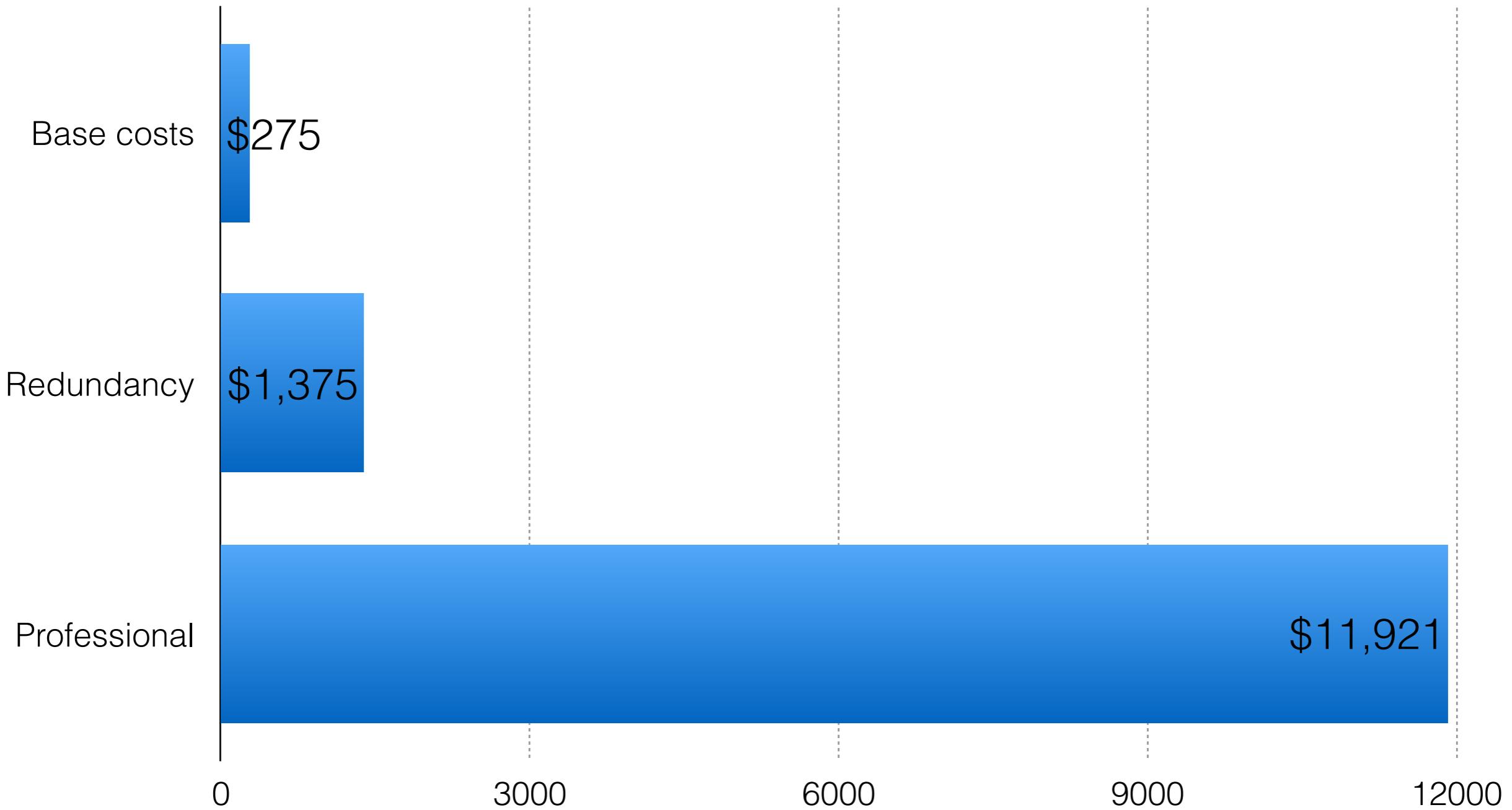
Cost of Translation on MTurk

Cost of Translation on MTurk

Base costs — $275

Redundancy — $1,375

Profe

Cost of Translation versus Professionals
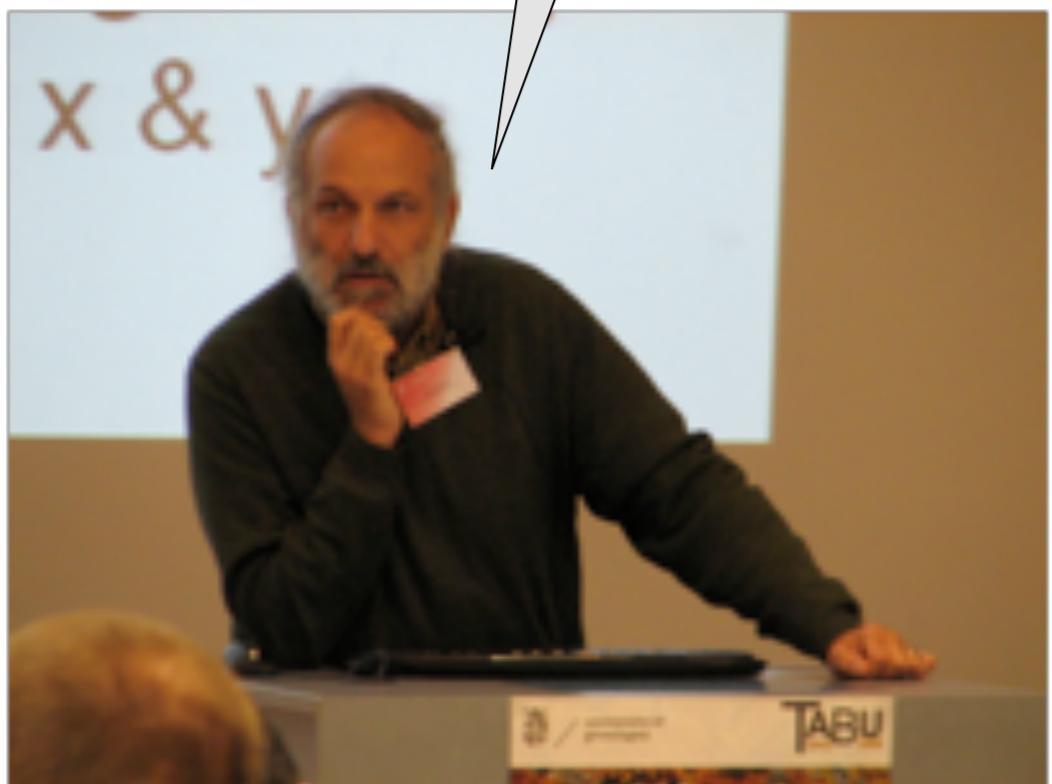
Cost of Translation on MTurk

Gao, Xu and Callison-Burch (NAACL 2015)
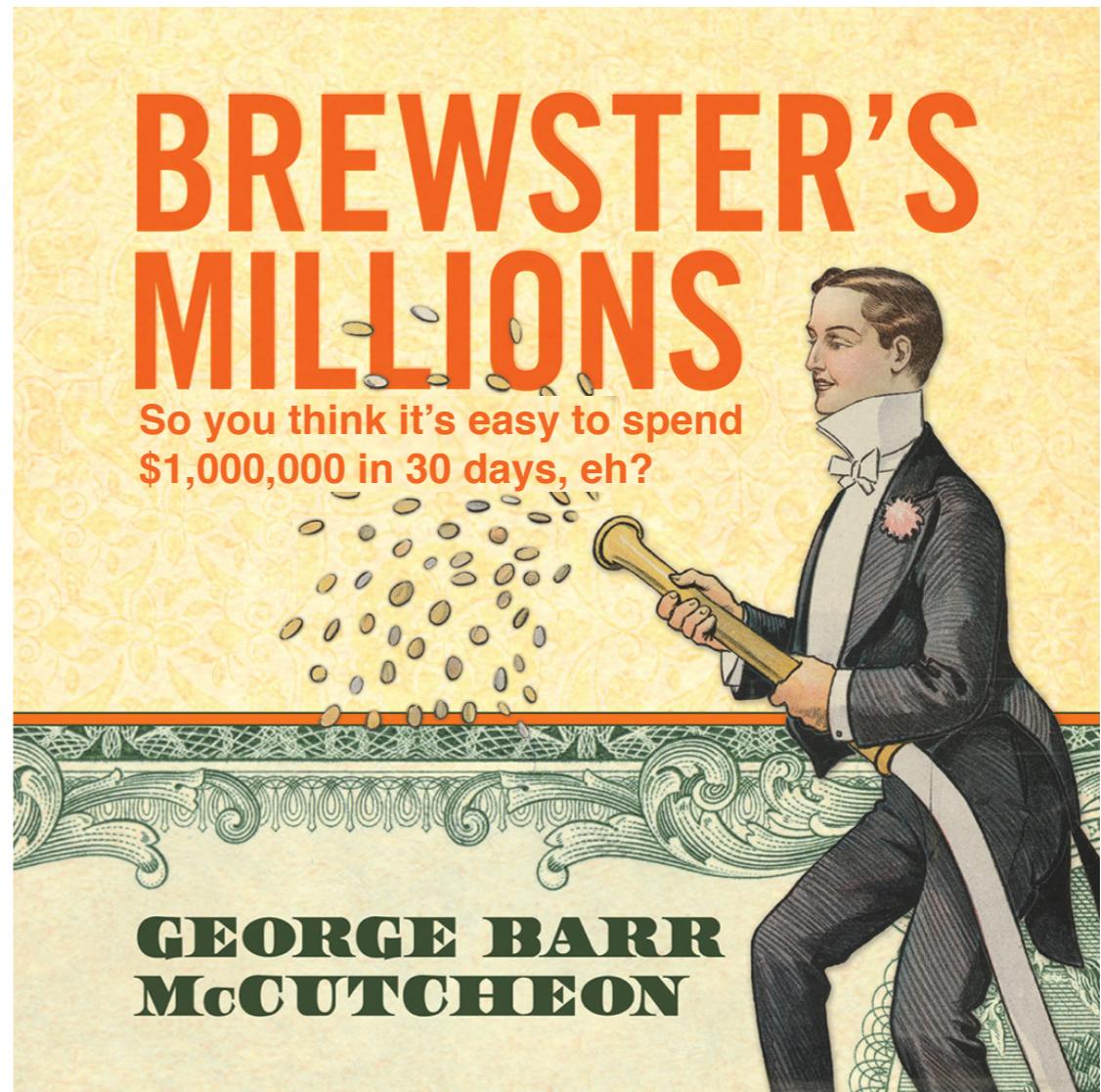
- Approximately an order of magnitude cheaper than the cost of professional translation

- 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

BREWSTER'S MILLIONS

So you think it's easy to spend $1,000,000 in 30 days, eh?

GEORGE BARR McCUTCHEON
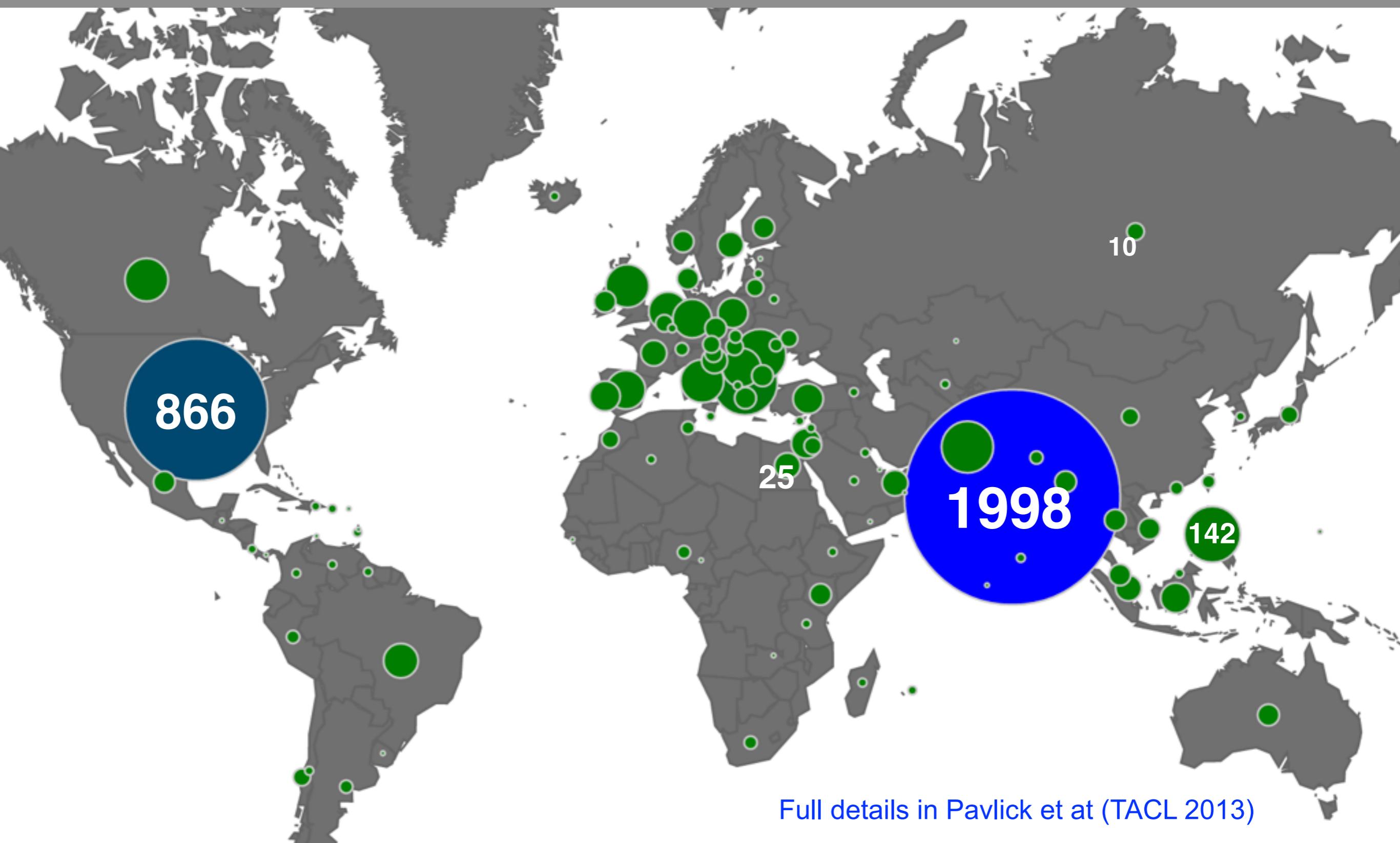
Leaderboard

Mechanical Turk Monitor: Top-1000 Recent Requesters

General	Top requesters	Arrivals	Completed	Search	About

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

Language Demographic Study

Full details in Pavlick et at (TACL 2013)

Survey

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?

Self-reported Native Languages

English	689	French	63	Vietnamese	34
Tamil	253	Polish	61	Macedonia	31
Malayalam	210	Urdu	56	Cebuano	29
Hindi	149	Tagalog	54	Swedish	26
Spanish	131	Marathi	48	Bulgarian	25
Telugu	87	Russian	44	Swahili	23
Chinese	86	Italian	43	Hungarian	23
Romanian	85	Bengali	41	Catalan	22
Portuguese	82	Gujarati	39	Thai	22
Arabic	74	Hebrew	38	Lithuanian	21
Kannada	72	Dutch	37	Punjabi	21
German	66	Turkish	35	Others	≤ 20
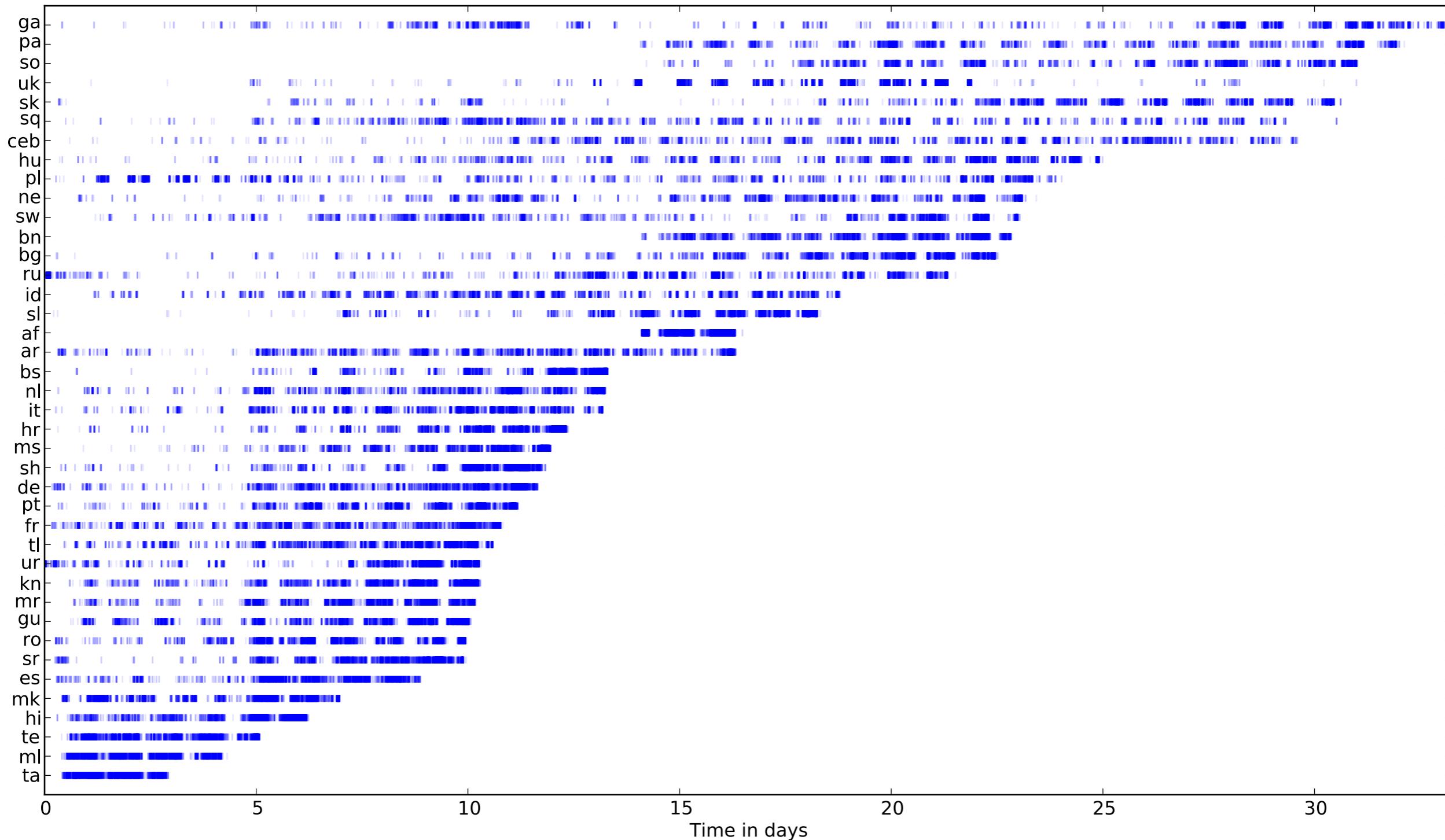
Translation Task

100 languages with the most articles

1,000 most viewed articles

10,000 most frequent words

Time to complete 10,000 translations

Quality Control

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
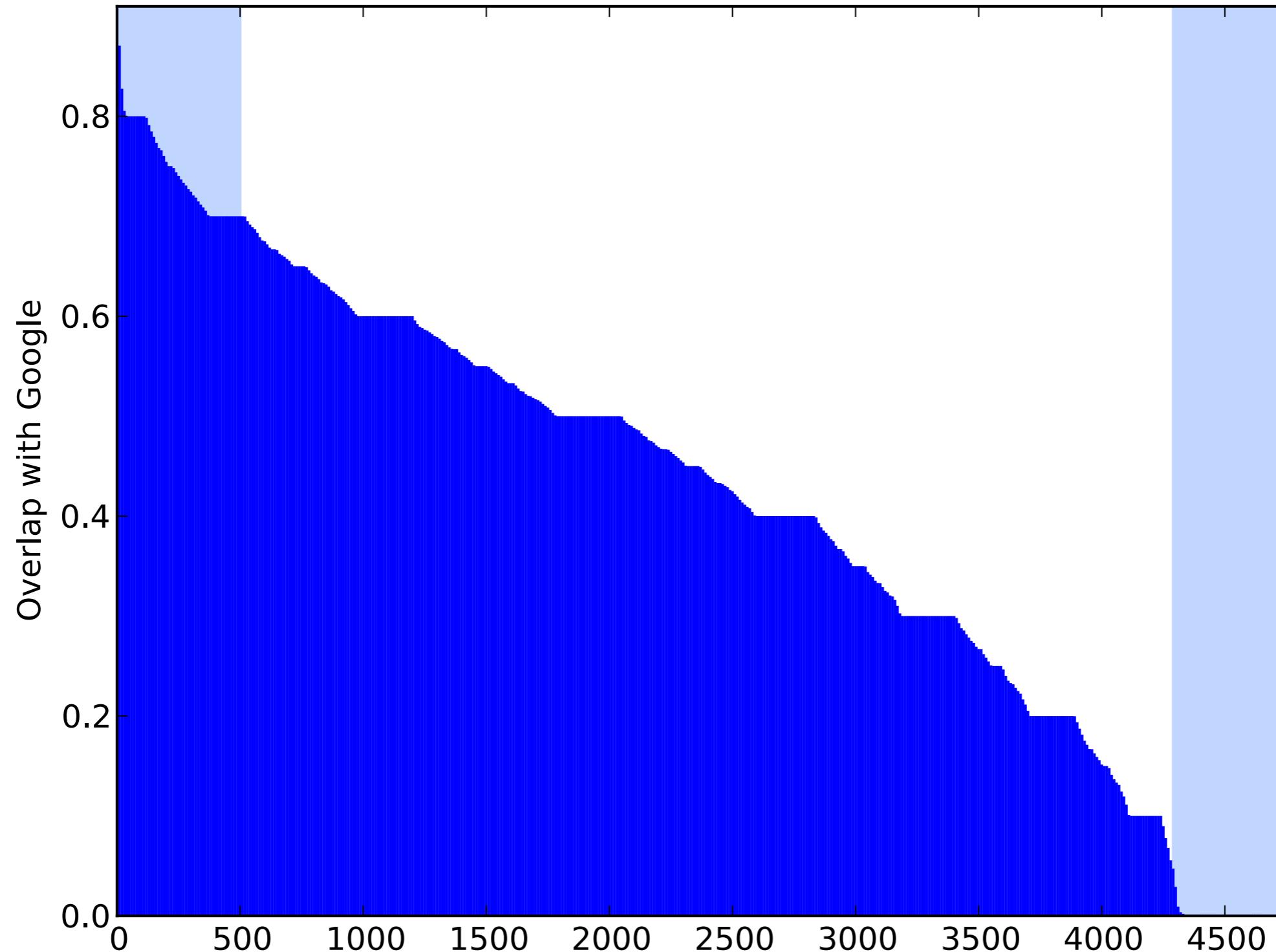
salvia and **salvia (plant)**
⚪Yes ⚪No ⚪Related but not synonymous
☐ Word is misspelled
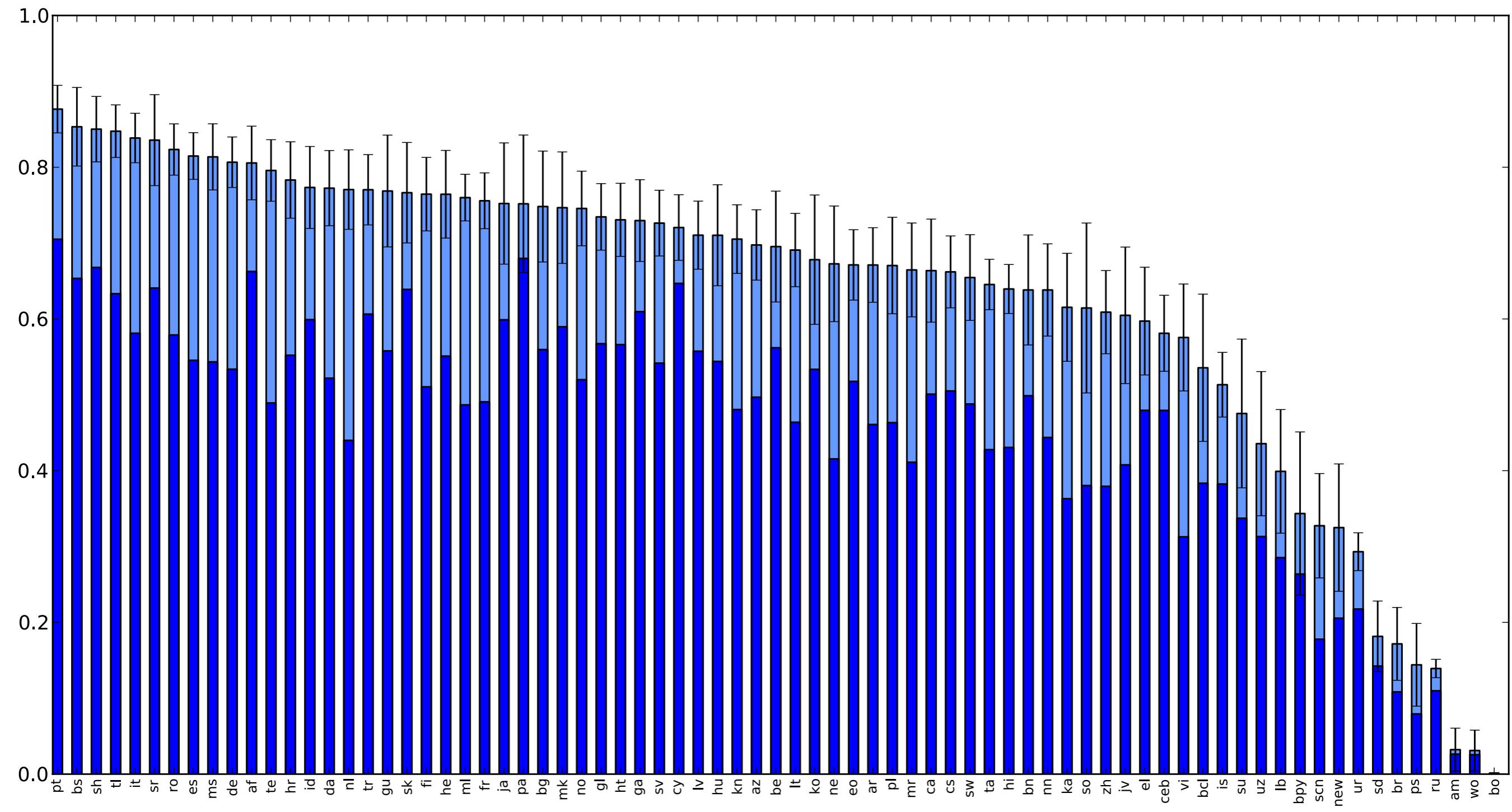
demo and **show**
⚪Yes ⚪No ⚪Related but not synonymous
☐ Word is misspelled

Cheating with Google Translate

Translation Quality

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

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

Bilingual Parallel Corpora from Scratch

पाकिस्तान ने भी प्रतिकार स्वरूप २८ मई १९९८ में छह परमाणु परीक्षण कर डाले।
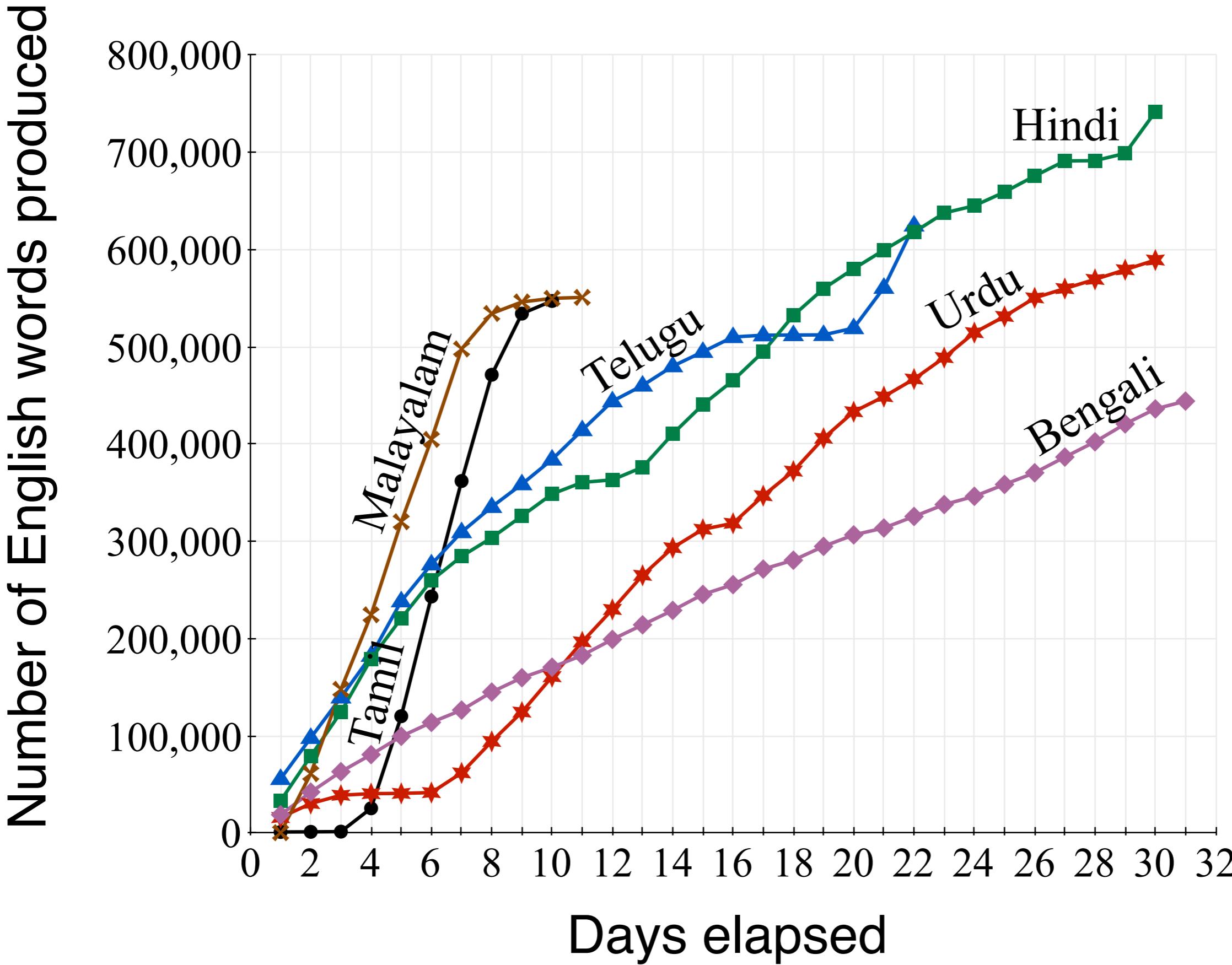
In retribution pakistan also did six nuclear tests on 28 may 1998.

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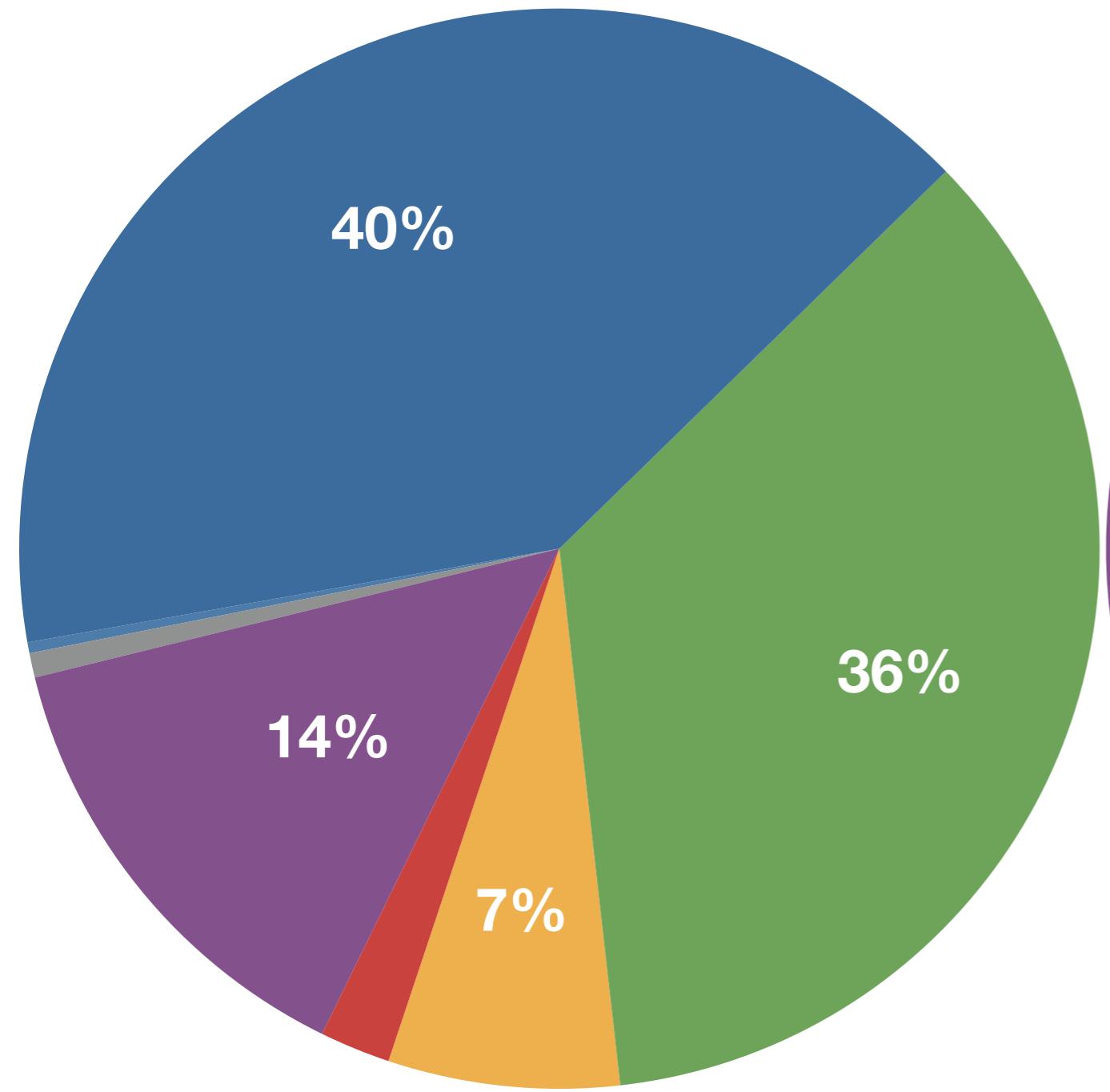
pakistan also did 6 nuclear test in retribution on 28 may, 1998

Rate of Translation

Number of English words produced

Days elapsed

Malayalam
Tamil
Telugu
Hindi
Urdu
Bengali
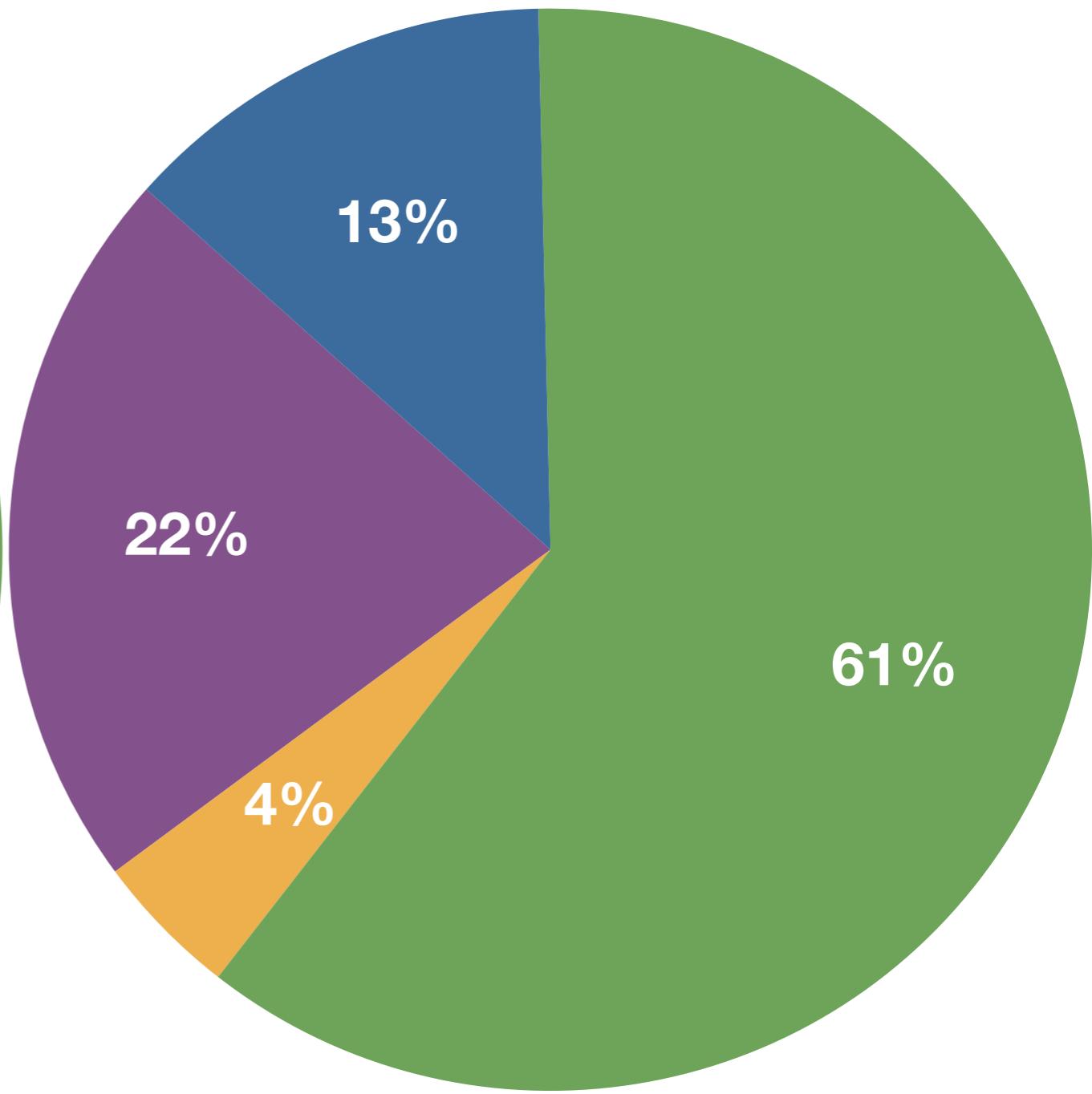
Better Representation of Languages

All Languages

40% SOV
36% SVO
14% No dominant order
7% VSO

SMT Languages

13% SOV
61% SVO
22% No dominant order
4% VSO
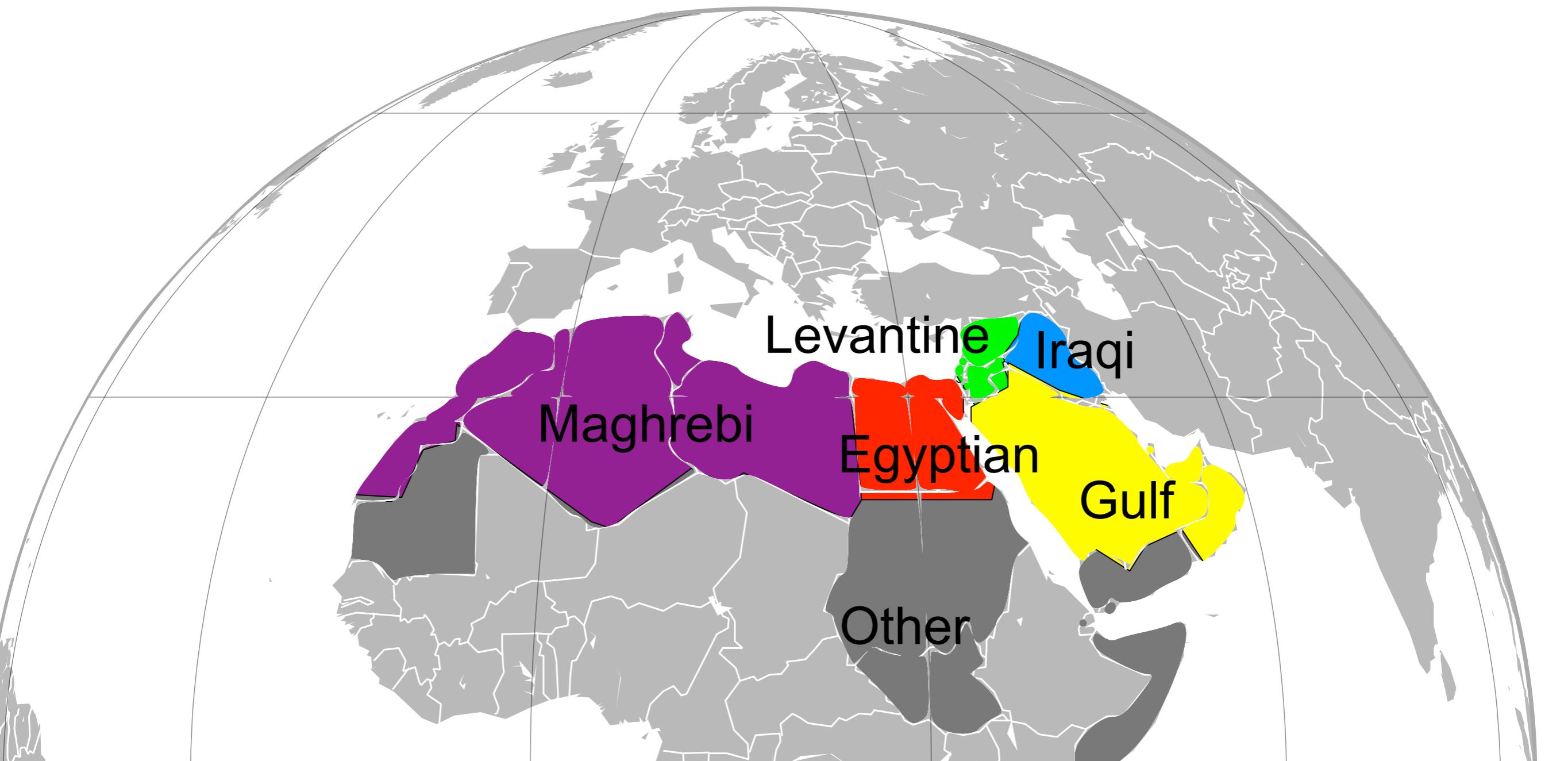
● SOV   ● SVO   ● VSO   ● VOS   ● No dominant order   ● OVS

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 الثلة من المجرمين تخضع للمحاكمة ؟‏

mtY snrY h*h Alvlp mn Almjrmyn txDE llmHAkmp ?

Levantine: ‏ايمتى رح نشوف هالشلة من المجرمين بتتحاكم ؟‏

AvmtY rH n\$wf hAl\$lp mn Almjrmyn bttHAkm ?

PT: Quando veremos esse grupo de criminosos serem julgados?
ES-EN: **Quando esse** group of criminals see **Serem julgados**?

Written Arabic Dialect

Article body (MSA)

Crowdsourcing Dialect ID

- 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

Label Distribution
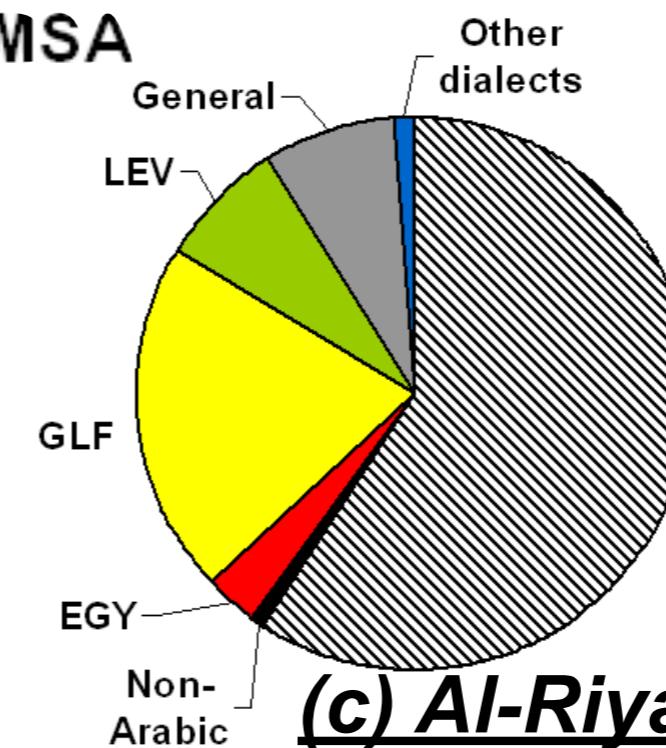
(a) All sources
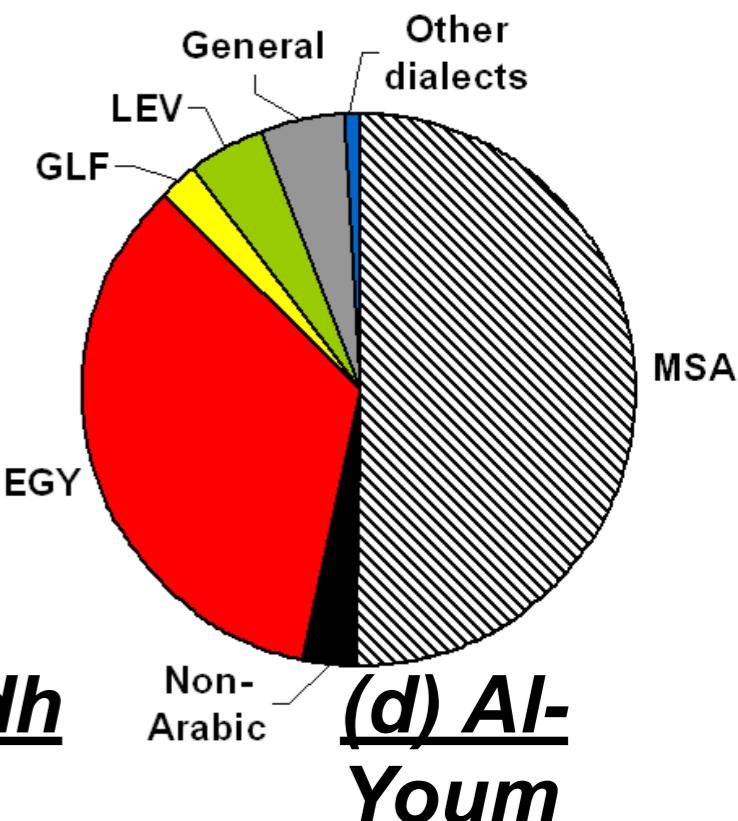
(b) Al-Ghad
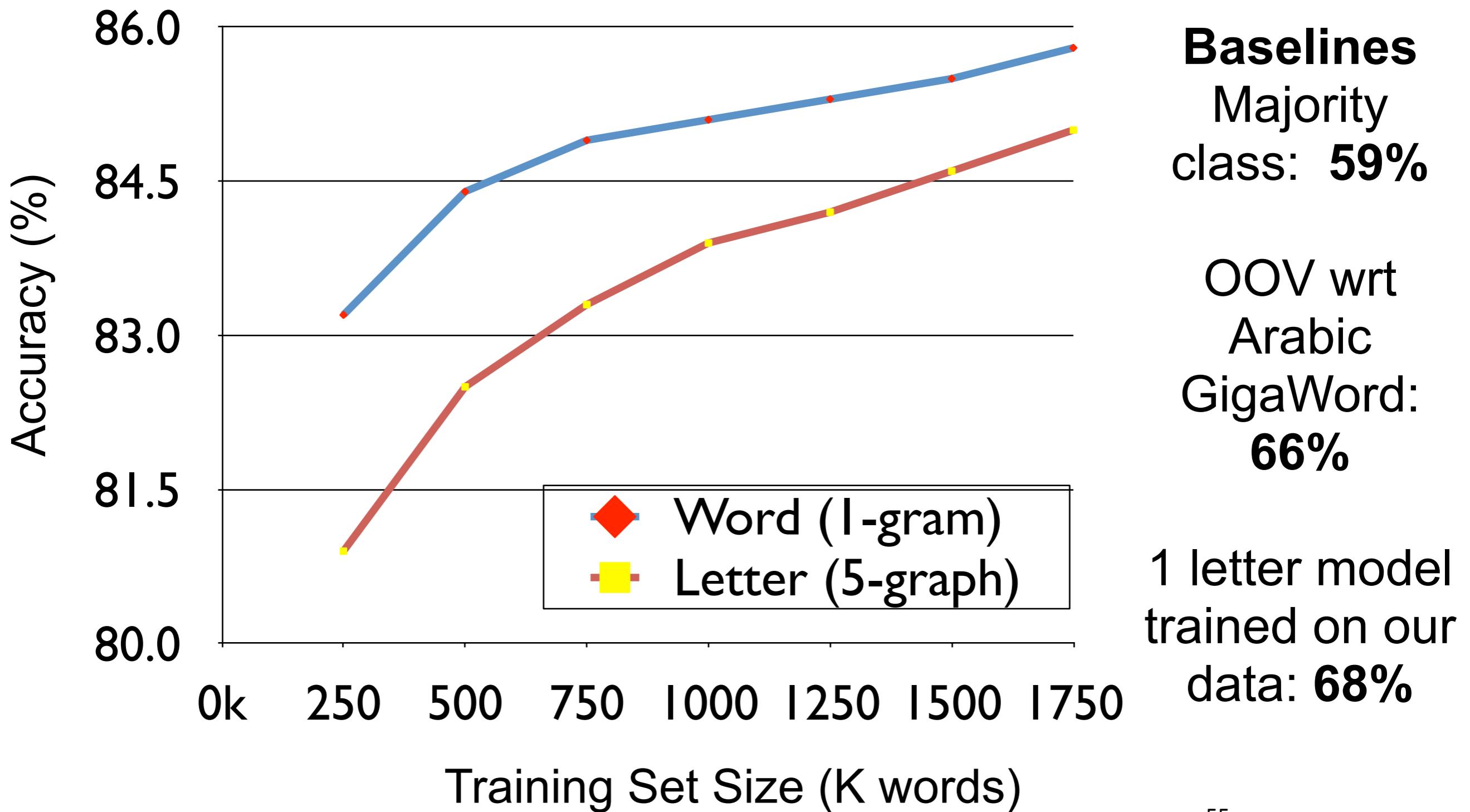
(c) Al-Riyadh
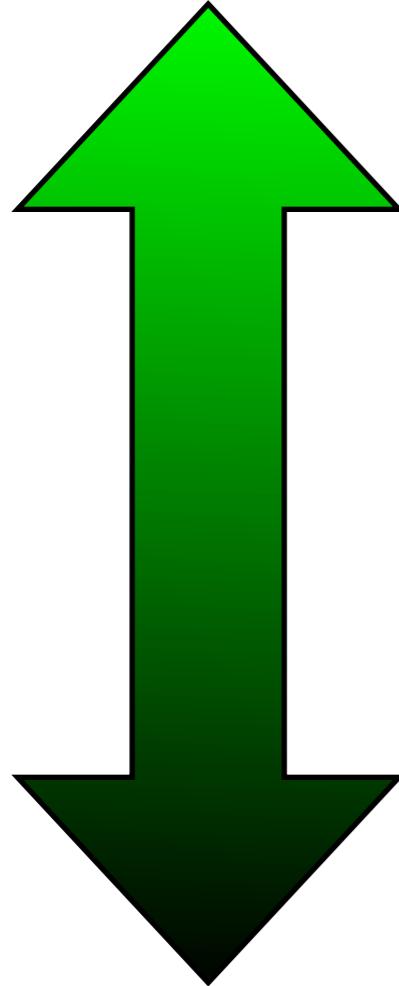
(d) Al-Youm

Automatic dialect ID

"Dialectness" quantified

Extremely dialectal Levantine words

Extremely MSA words (i.e. very unlikely in dialectal context)

Al-Ghad		
w	Gloss	$DF(w)$
w	شو	what	139.0
xly	خلي	let	132.8
xlS	خلص	enough	117.5
AlHky	الحكي	the-talk	115.8
Endw	عندو	he has	95.3
bdy	بدي	I will/want	93.6
AzA	ازا	if	93.6
mnyH	منيح	good	93.6
$wy	شوي	little	92.8
<nw	إنو	that	90.2
hAy	هاي	this (f.)	80.0
bEdyn	بعدين	then	70.2
mw	مو	not	65.5
Ay$	ايش	what	63.8
bdw	بدو	he will/wants	60.3
⋮	⋮		
Ebr	عبر	through	0.146
w>n	وأن	and-that	0.145
Al<slAm	الإسلام	Islam	0.138
tEAlY	تعالى	almighty	0.138
SlY	صلى	blessed	0.127
AldymqrATyp	الديمقراطية	democratic	0.108

$$DF(w) \stackrel{\text{def}}{=} \frac{f(w|D)}{f(w|MSA)}$$

"Dialectness" factor

What is this useful for?

- Characterizing communicants

  – What is this writer's native dialect?

  – Where are they from?

  – Informal relation with their interlocutor

- Harvesting written dialect from large web crawls

  – Useful for training dialect language models for ASR

- Identifying dialect sentence to then translate

  – Training data for a statistical machine translation

57

Crowdsourcing Arabic Dialects

- Translated dialect-labeled segments `EGY` `LEV`

Dialect Classification HIT	$10,064
Sentence Segmentation HIT	$1,940
Translation HIT	$32,061
Total Cost	$44,065
Num words translated	1,516,856
Cost per word	2.9 cents/word
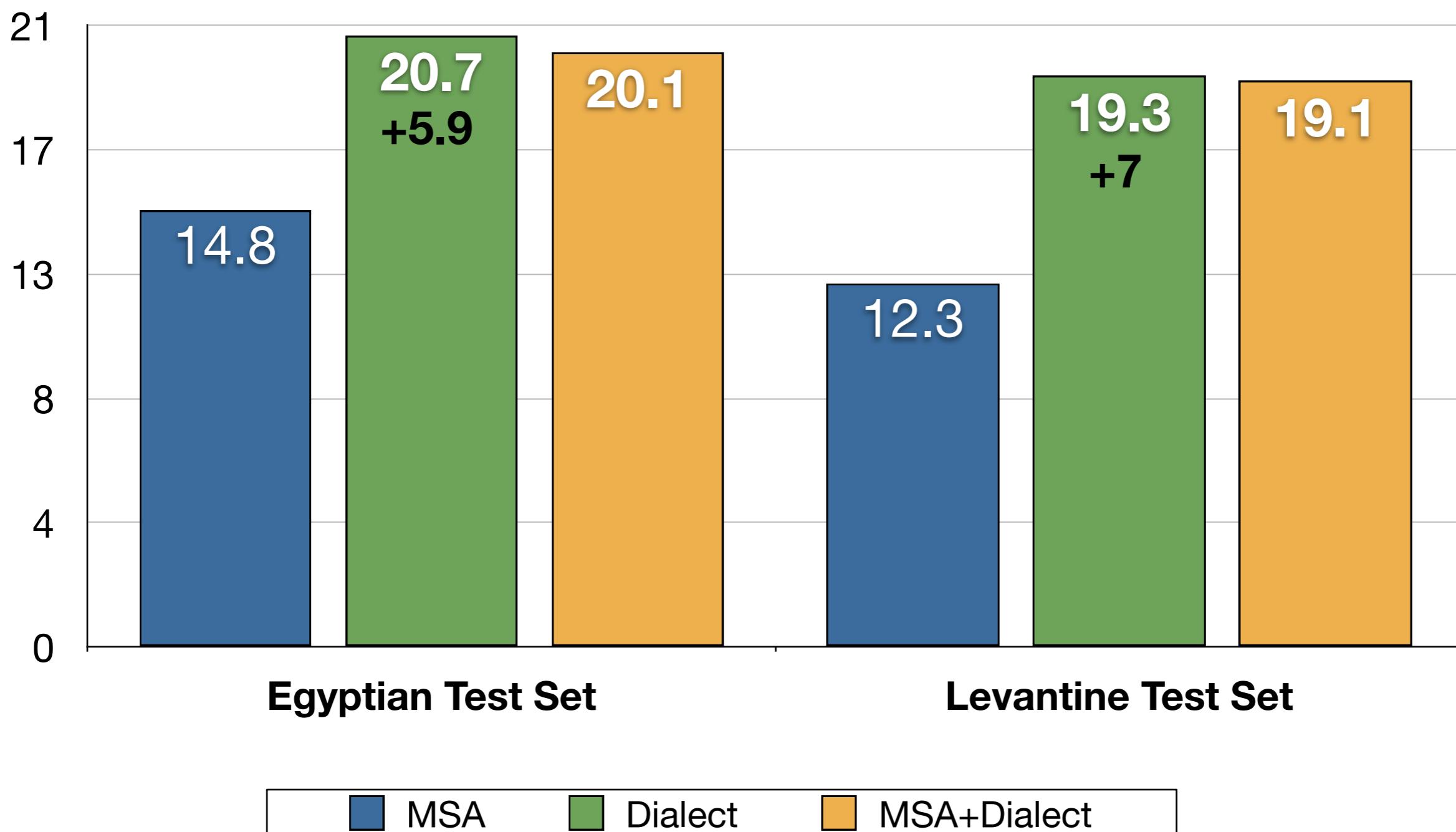
- 121 workers completed 20+ assignments

- 200,000 words translated per week

- Trained BBN's machine translation system

Examples of Dialect Translation

Dialect Input	MSA system	Dialect system	Reference
EGY انت بتعمـــل لـه اعـــلان ولا ايـــه ؟ !!	You are working for a declaration and not?	You are making the advertisement for him or what?	Are you promoting it or what?!!
EGY 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.
LEV لهيك الجو كتبيير كووول	God you the atmosphere.	This is why the weather is so cool	This is why the weather is so cool
LEV طول بالك عم نمزح	Do you think about a joke long.	Calm down we are kidding	Calm down, we are only kidding
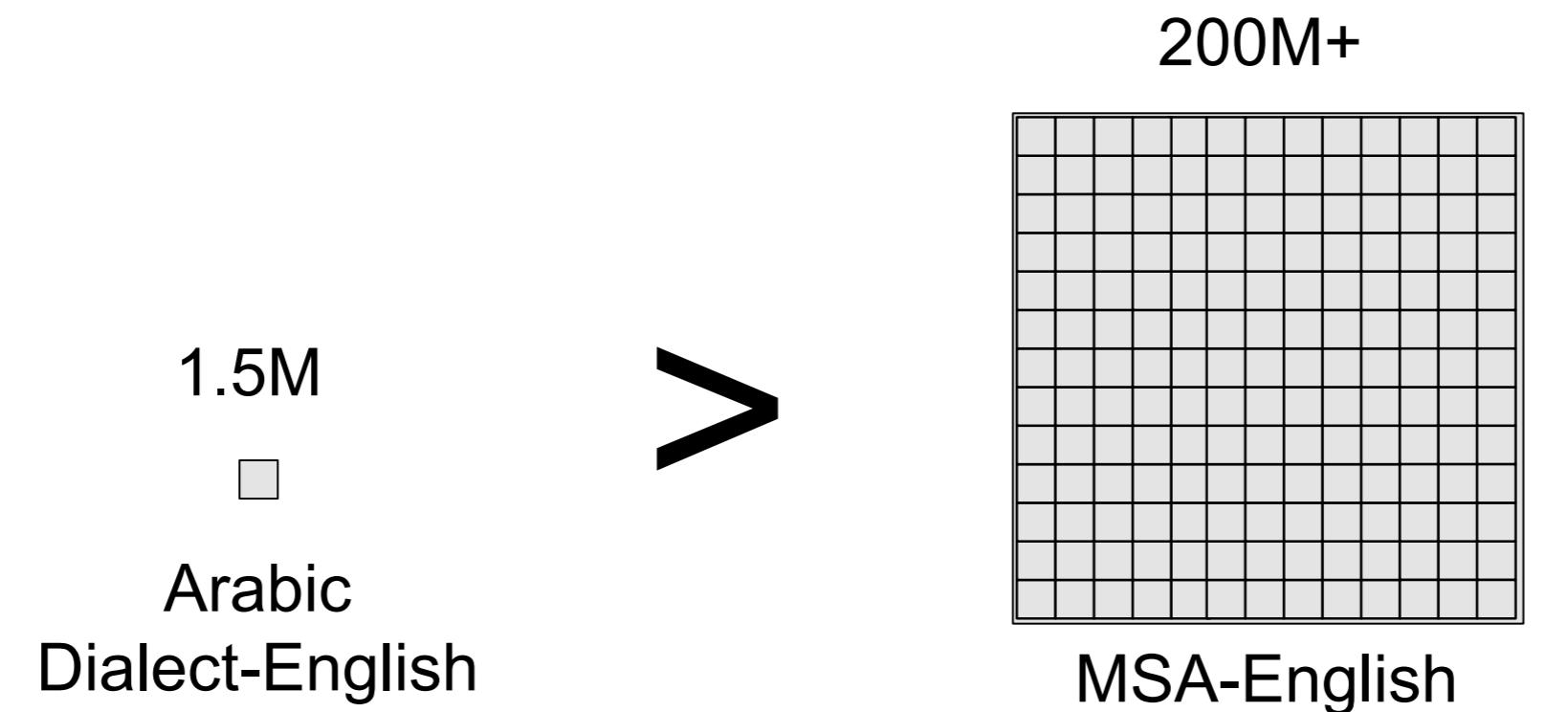
Dialect v. MSA data for MT

For machine translation of Arabic dialects

200M+

1.5M

>

Arabic
Dialect-English

MSA-English

2 months
$44k
+6 Bleu

5 years
$10s of millions
no added Bleu
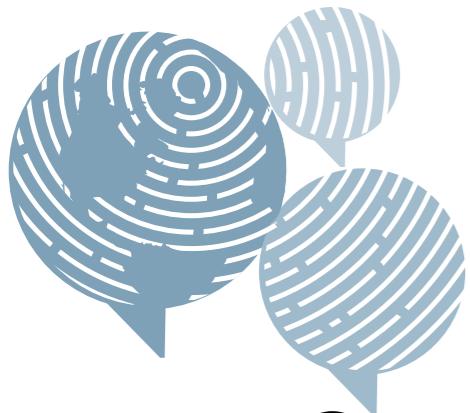
Implications of low cost, high quality translations for research

We want to respond to a "Surprise Language"

The Los Angeles Times reported that at about 5:20 P.M. on Tuesday March 4, 2003, **a bomb** concealed in a backpack **exploded** at the airport in Davao City, the second largest city **in the Philippines**. At least 23 people were reported dead, with more than 140 injured, and President Arroyo of the Philippines characterized the blast as a terrorist act.

With the 13 hour time difference, it was then at 4:20 A.M on the same date in Washington, DC. **Twenty-four hours later**, at 4:13 A.M. on March 5, participants in the Translingual Information Detection, Extraction and Summarization (TIDES) program were notified that **Cebuano** had been chosen as the **language of interest** for a "surprise language" practice exercise that had been planned quite independently to begin on that date. The notification observed that Cebuano is spoken by 24% of the population of the Philippines and that it is the *lingua franca* in the south Philippines, where the event occurred.

100 languages

- Google Translate, Bing translator, and the DoD's Center for Applied MT have <100 langs
- There is not enough data to reach acceptably high quality in new languages
- I want us to do 100 languages
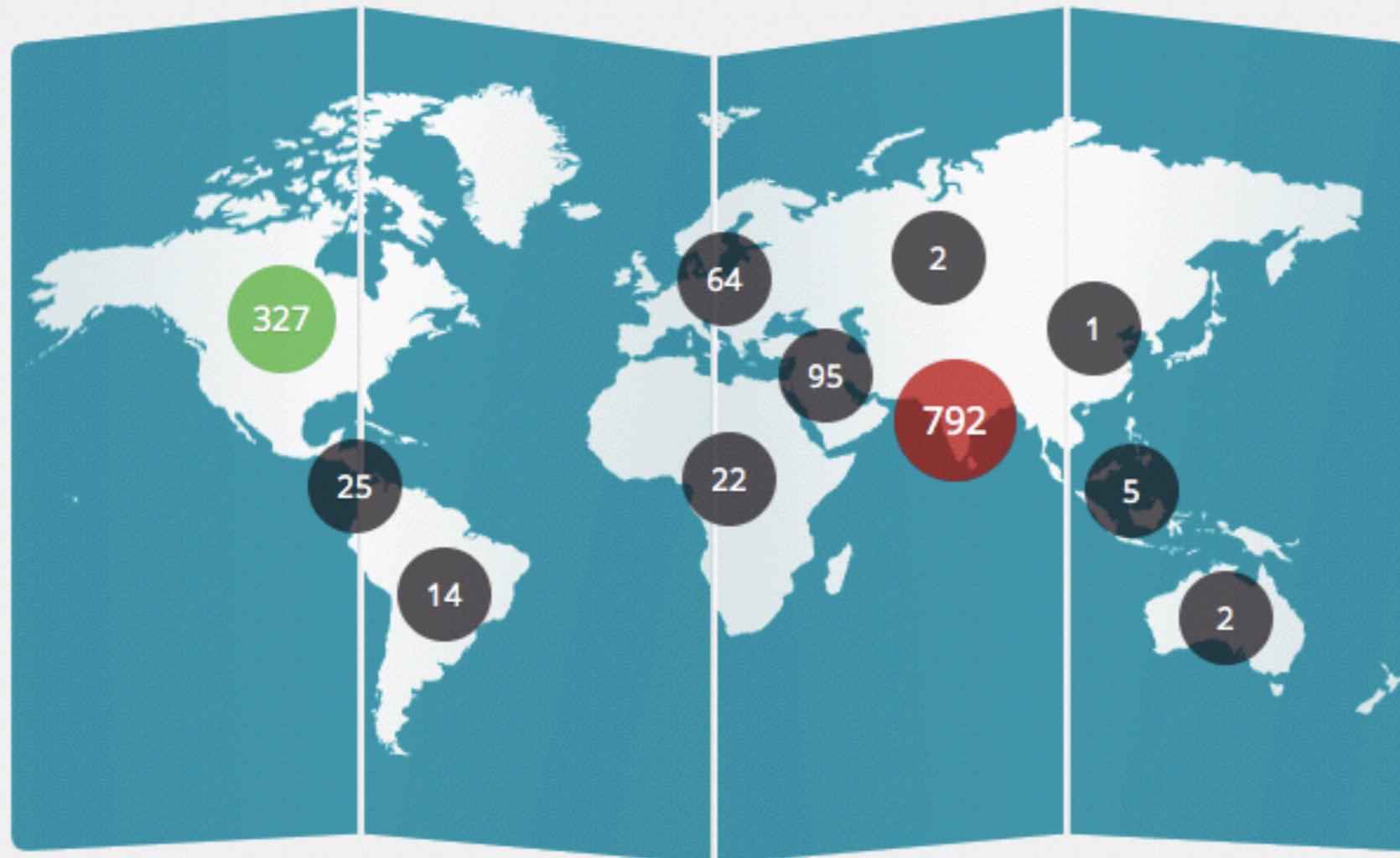
Individual researchers now have their own data production companies

✔ Parallel corpora for six Indian languages
✔ Arabic dialect ID dataset
✔ 1.5M word English-Arabic dialect parallel corpus
✔ Bilingual dictionaries for ~100 languages
✔ English translation of 180 hrs of spoken Spanish
✔ English translation of spoken Egyptian dialect
Corpus annotating and correcting ESL errors
Speech collection and transcription for new lang.
De-romanization / text normalization corpus

Resources

Translator activity

327

64

2

95

1

792

25

22

5

14

2

Recent public translations

ES La administración judicial temporal ...

The temporary judicial administrati... **EN**

6 hours ago on El pais test - May 23

ES En un comunicado remitido a la Co...

In a communication sent to the Nat... **EN**

6 hours ago on El pais test - May 23
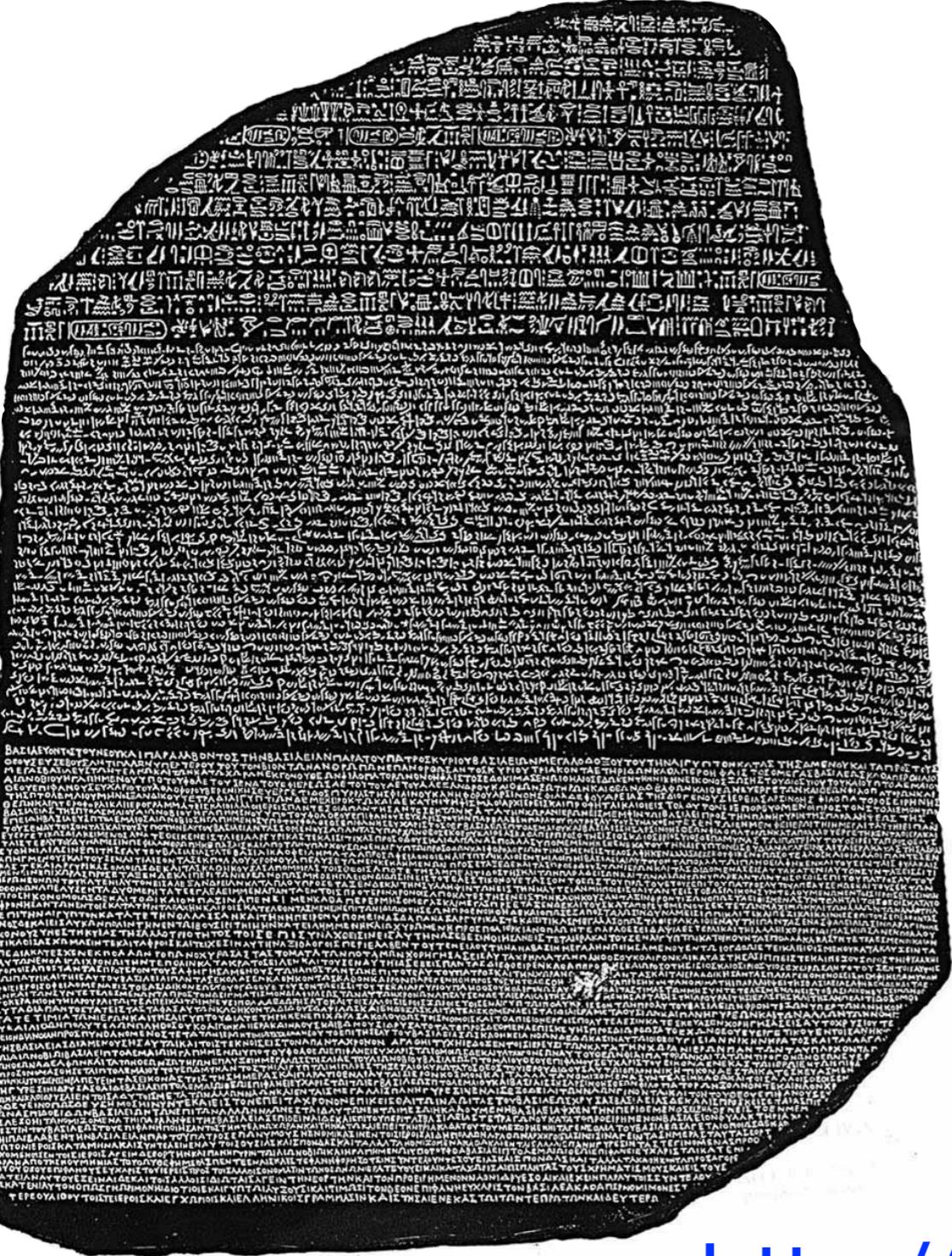
ES El Tribunal de Palermo ha ordenad...

The Palermo Court has ordered the... **EN**

6 hours ago on El pais test - May 23

go to public translations ⊙

High quality crowd translations from around the world

Order Translation

or learn how it works ⊙

Machine Translation Class

- Developed w/Adam Lopez, Matt Post and Chris Dyer
- Project based class
- Students solve real open research problems in MT
- Projects are automatically gradable, MOOC ready

http://mt-class.org

Crowdsourcing and Human Computation Class

- Undergraduate course that I developed at Penn
- Student-designed (super creative!) final projects
- Series of homework assignments that create a final project

http://crowdsourcing-class.org

Gun Violence Database

- Build a gun violence text classifier
- Validate the model's prediction with the crowd
- Perform quality control
- Build an interface to extract structured information from an article
  - Date/location
  - Shooter and Victim's name, age, gender, race
  - Domestic violence, alcohol, used during crime
- Analyze the data

Domestic violence 565
During crime 410
Self defense 227
Police shooter 195
Alcohol 158
Police victim 129
Stolen gun 53

Number of reports

Revolver

1 ▬▬▬▬▬▬▬ 213

1 ▬▬▬▬▬▬▬ 21
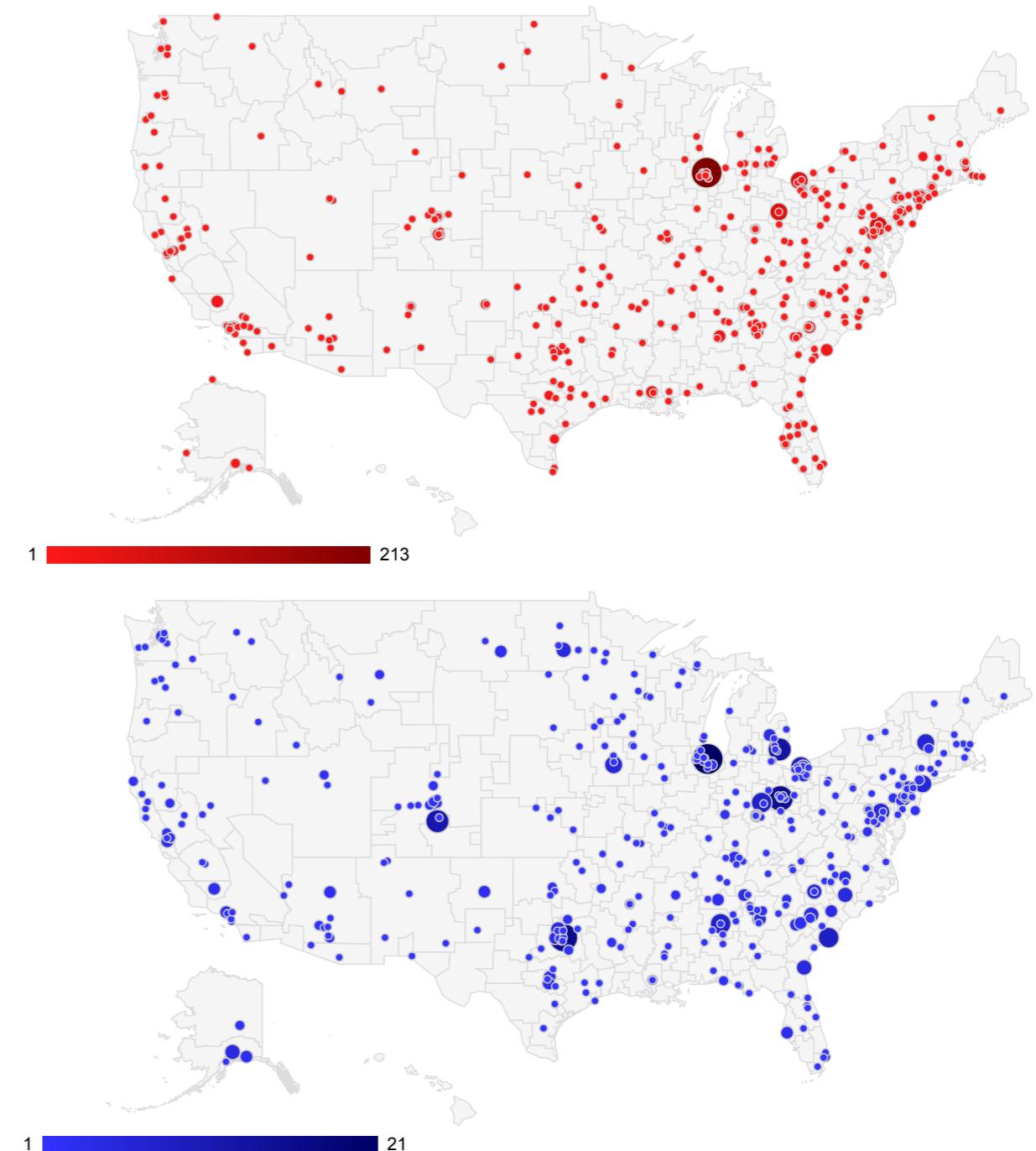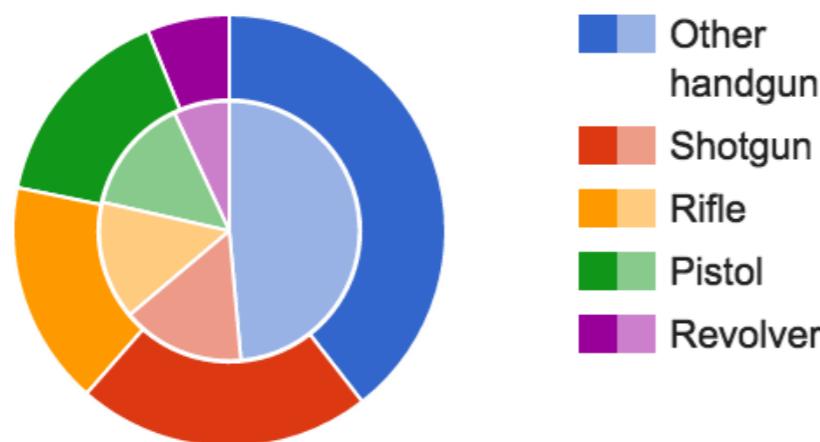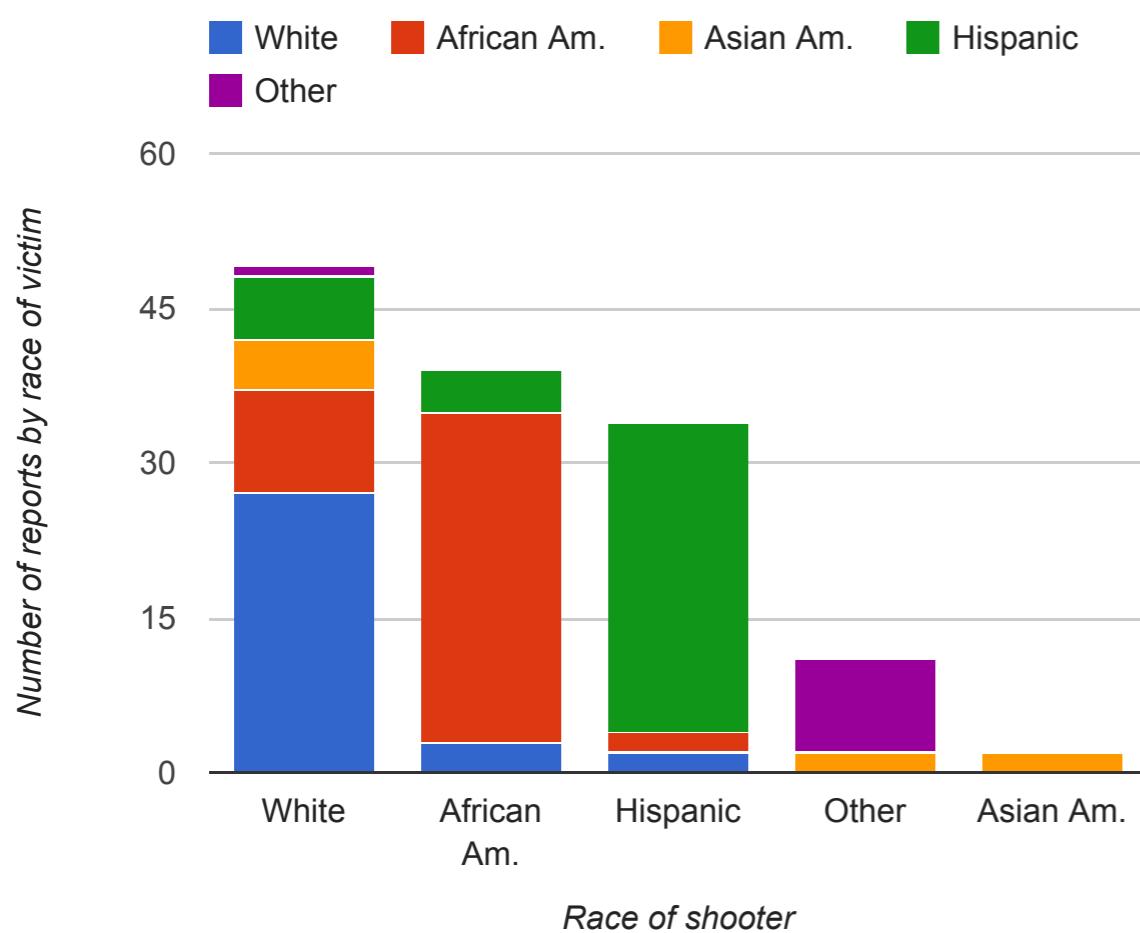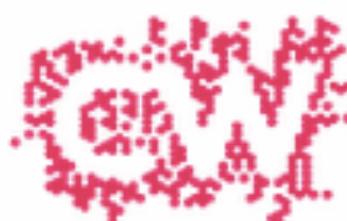
Figure 4: Intentional shootings (top) and unintentional shootings (bottom) across US cities, according to 8,800

Crowd Workers

Crowd Workers of the world, unite! Join with other workers to help make Mechanical Turk a better place to work. You can make it easier for everyone to find good work, and to make a fair wage

Get started by installing our extension

Currently we support: chrome

Thanks!

Chris Callison-Burch

ccb@upenn.edu

@ccb on Twitter

More info in Omar Zaidan's PhD thesis "Crowdsourcing Annotation for Machine Learning in Natural Language Processing Tasks" and our publications. Omar's thesis defense video is online at http://vimeo.com/clsp/omar-zaidan-thesis-defense-talk

NP❶	NP❷	P❸	V❹	AUX❺
hamd ansary	*na}b sdr*	*kylye*	*namzd*	*taa*

NP❶	NP❷	P❸	V❹	AUX❺
Hamid Ansari	*Vice President*	*for*	*nominated*	*was*

PP❻
NP❶   NP❷   P❸   V❹   AUX❺

hamd ansary   *na}b sdr*   *kylye*   *namzd*   *taa*

PP❻
NP❶   NP❷   P❸   V❹   AUX❺

Hamid Ansari   *Vice President*   *for*   *nominated*   *was*

PP❻
NP❷   P❸
hamd ansary   na}b sdr   kylye

NP❶

VP❼
V❹   AUX❺
namzd   taa

PP❻
P❸   NP❷
Hamid Ansari   for   Vice President

NP❶

VP❼
V❹   AUX❺
nominated   was

Tree 1

```
                                VP❽
                    ┌────────────┴────────────┐
                   PP❻                        VP❼
              ┌─────┴─────┐              ┌─────┴─────┐
  NP❶        NP❷          P❸           V❹          AUX❺
   △          △           │            │            │
hamd ansary  na}b sdr    kylye        namzd         taa
```

Tree 2

```
                                VP❽
                    ┌────────────┴────────────┐
                   PP❻                        VP❼
              ┌─────┴─────┐              ┌─────┴─────┐
  NP❶        P❸          NP❷           AUX❺         V❹
   △          │           △             │            │
Hamid Ansari  for    Vice President    was       nominated
```

Top tree:

- S❾
  - VP❽
    - PP❻
      - NP❶ — *hamd ansary*
      - NP❷ — *na}b sdr*
      - P❸ — *kylye*
    - VP❼
      - V❹ — *namzd*
      - AUX❺ — *taa*

Bottom tree:

- S❾
  - VP❽
    - VP❼
      - NP❶ — *Hamid Ansari*
      - AUX❺ — *was*
      - V❹ — *nominated*
    - PP❻
      - P❸ — *for*
      - NP❷ — *Vice President*

PP❻
NP❶ NP❷ P❸ V❹ AUX❺

hamd ansary *na}b sdr* *kylye* *namzd* *taa*

PP❻
NP❶ NP❷ P❸ V❹ AUX❺

Hamid Ansari *Vice President* *for* *nominated* *was*

PP❻
NP❷   P❸

VP❼
V❹   AUX❺

NP❶

hamd ansary   *na}b sdr*   *kylye*   *namzd*   *taa*

PP❻
P❸   NP❷

VP❼
V❹   AUX❺

NP❶

Hamid Ansari   *for*   *Vice President*   *nominated*   *was*

VP❽
├─ PP❻
│ ├─ NP❶ — *hamd ansary*
│ ├─ NP❷ — *na}b sdr*
│ └─ P❸ — *kylye*
└─ VP❼
   ├─ V❹ — *namzd*
   └─ AUX❺ — *taa*

VP❽
├─ NP❶ — *Hamid Ansari*
├─ PP❻
│ ├─ P❸ — *for*
│ └─ NP❷ — *Vice President*
└─ VP❼
   ├─ AUX❺ — *was*
   └─ V❹ — *nominated*

S❾
VP❽
PP❻
VP❼
NP❶ *hamd ansary*
NP❷ *na}b sdr*
P❸ *kylye*
V❹ *namzd*
AUX❺ *taa*

S❾
VP❽
VP❼
PP❻
NP❶ *Hamid Ansari*
AUX❺ *was*
V❹ *nominated*
P❸ *for*
NP❷ *Vice President*