

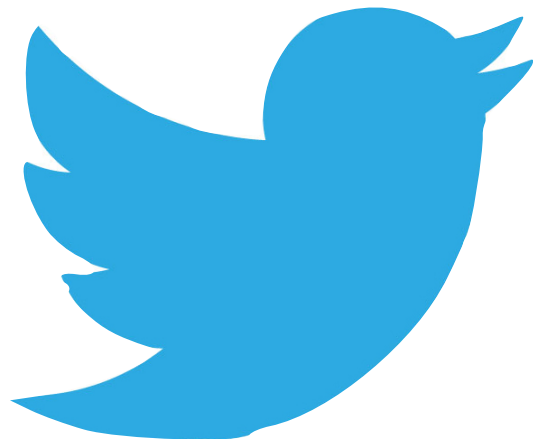
NETS 213: CROWDSOURCING
AND HUMAN COMPUTATION

Crowdsourcing and Translation

Professor Callison-Burch



facebook®



IBM

TED

IDEAS WORTH SPREADING

coursera



duolingo

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 .

...

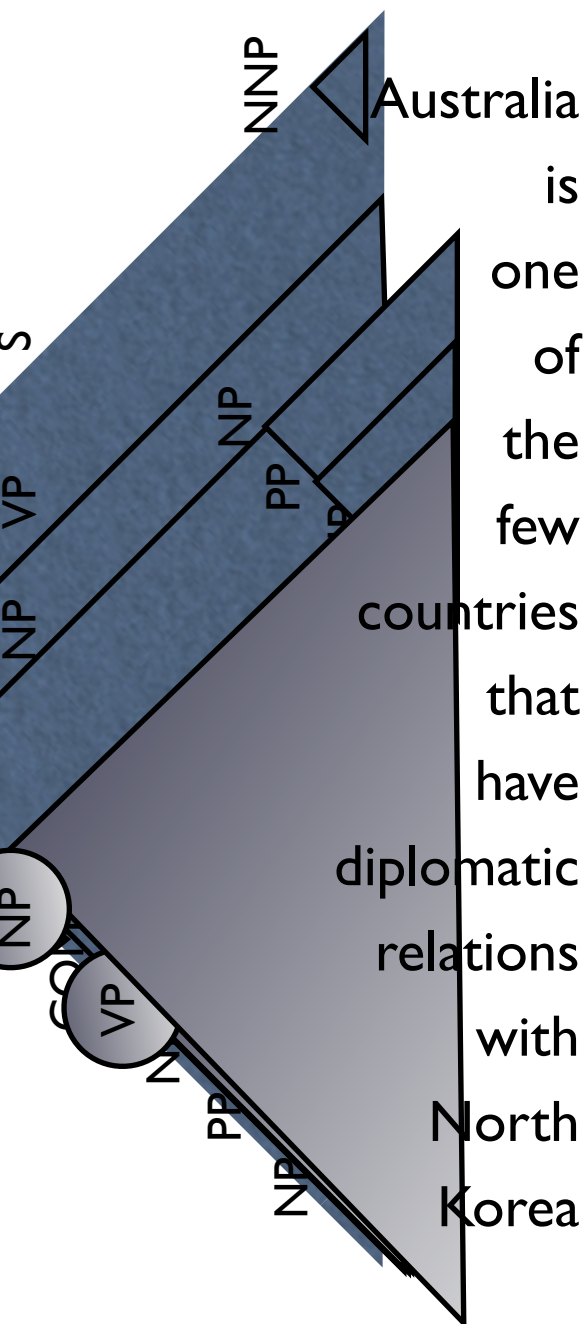
English

China's energy and raw materials
production up.

ANC calls for steps to prevent deaths in
police custody .

...

Extracting Translation Rules



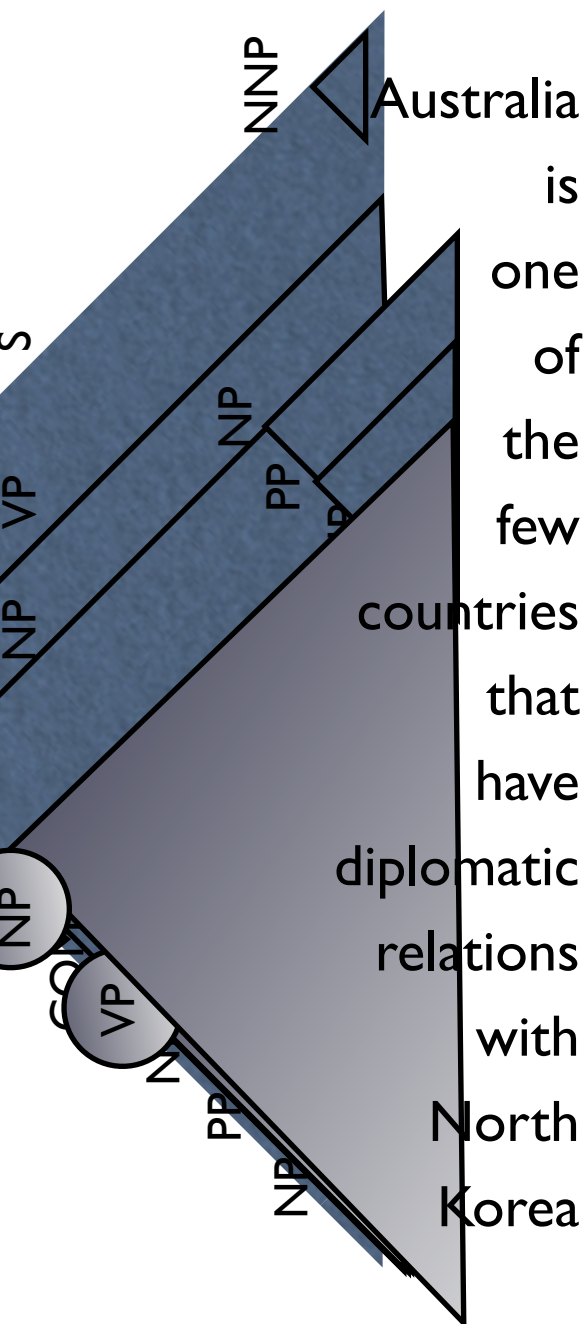
澳洲是与北韩有邦交的少数国家之一

●										
	●									
									●	
									●	
							●			
								●		
									●	
							●			
					●					
						●				
						●				
							●			
								●		
									●	
										●

VP → 与北韩有邦交,
have diplomatic relations
with North Korea

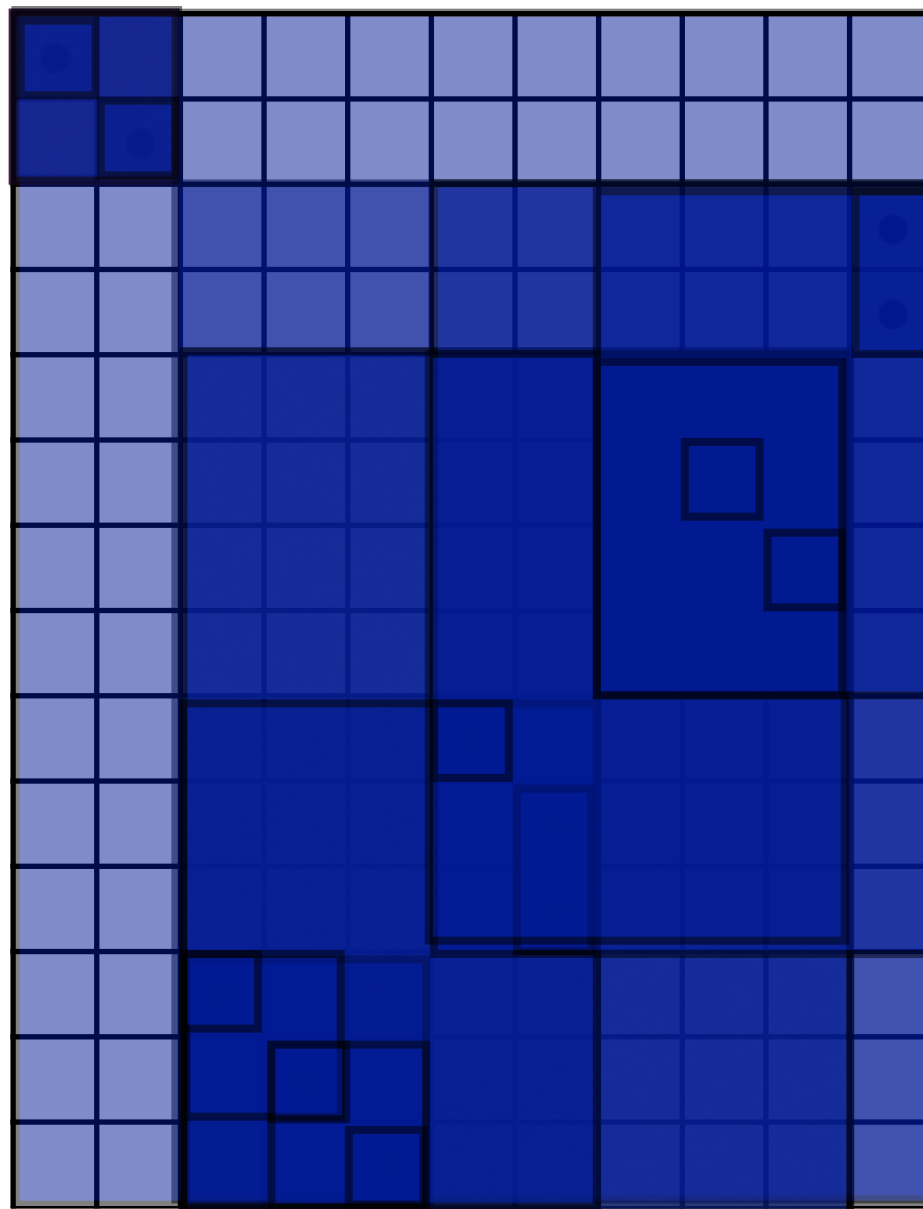
NP → 与北韩有邦交
的少数国家, the few
countries that have
diplomatic relations with
North Korea

Extracting Translation Rules



Australia
is
one
of
the
few
countries
that
have
diplomatic
relations
with
North
Korea

澳洲是与北韩有邦交的少数国家之一



VP → 与北韩有邦交,
have diplomatic relations
with North Korea

NP → 与北韩有邦交
的少数国家, the few
countries that have
diplomatic relations with
North Korea

Example translation for Urdu

	Urdu	English
S →	NP① VP②	NP① VP②
VP →	PP① VP②	VP② PP①
VP →	V① AUX②	AUX② V①
PP →	NP① P②	P② NP①
NP →	<i>hamd ansary</i>	<i>Hamid Ansari</i>
NP →	<i>na}b sdr</i>	<i>Vice President</i>
V →	<i>namzd</i>	<i>nominated</i>
P →	<i>kylve</i>	<i>for</i>
AUX →	<i>taa</i>	<i>was</i>

NP**1**
△
hamd ansary

NP**2**
△
na}b sdr

P**3**
|
kylve

V**4**
|
namzd

AUX**5**
|
taa

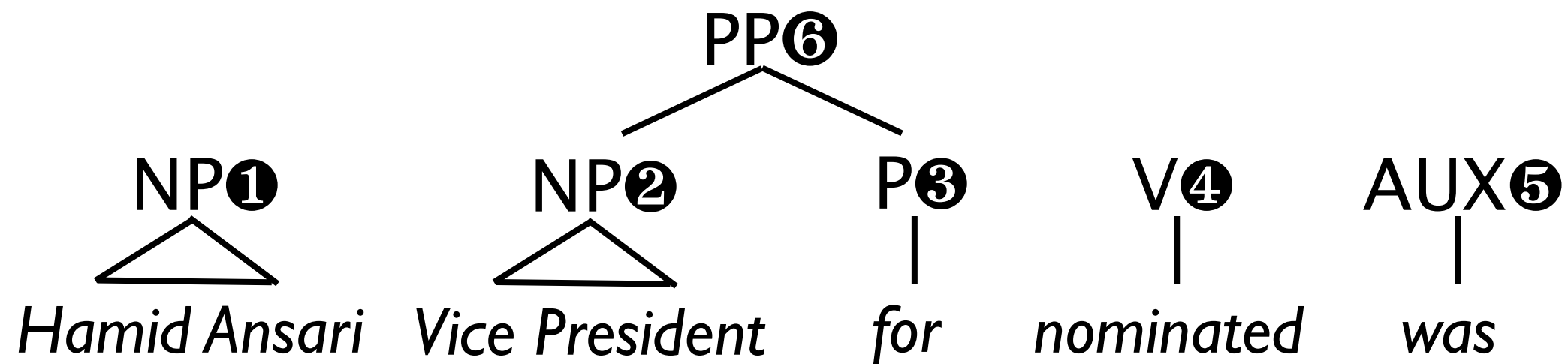
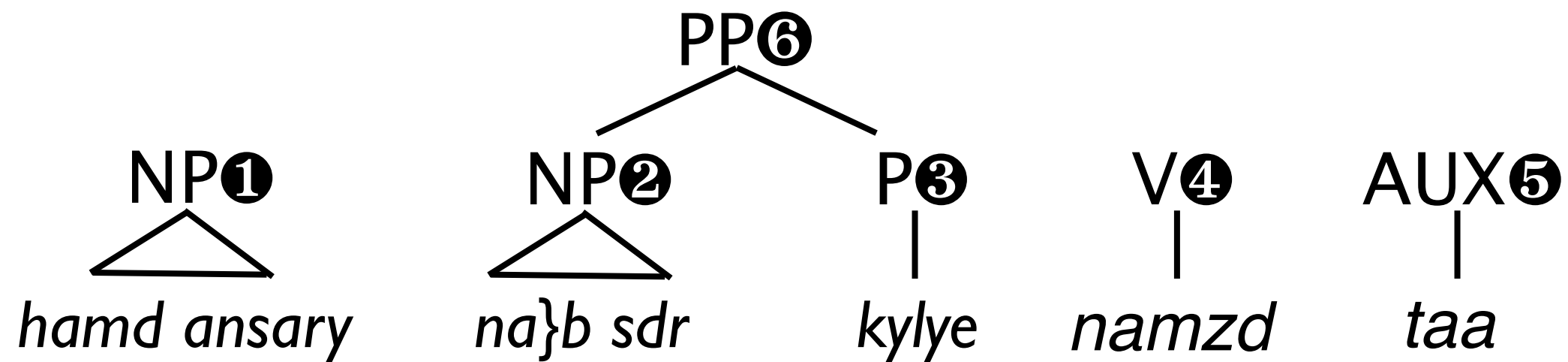
NP**1**
△
Hamid Ansari

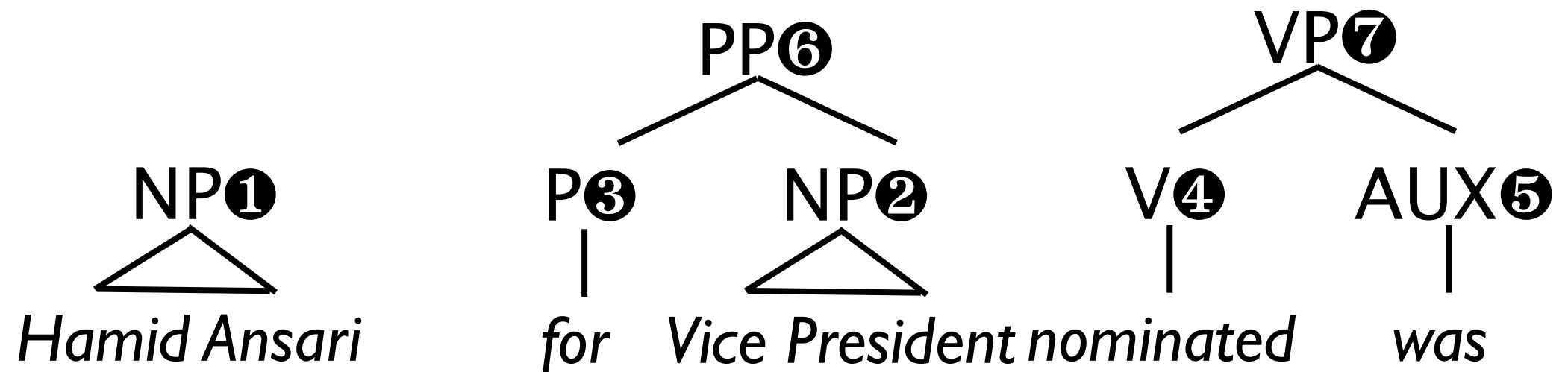
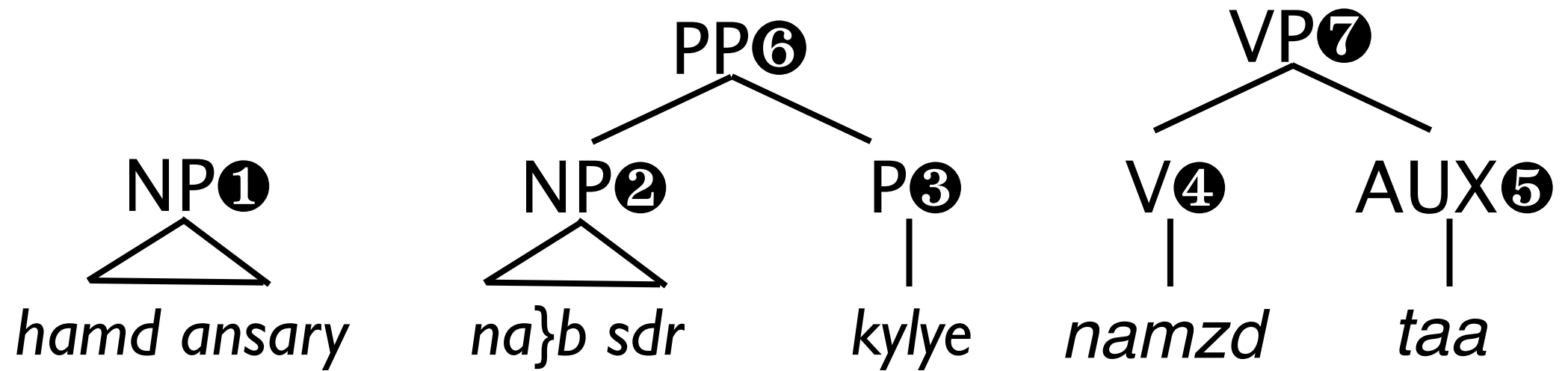
NP**2**
△
Vice President

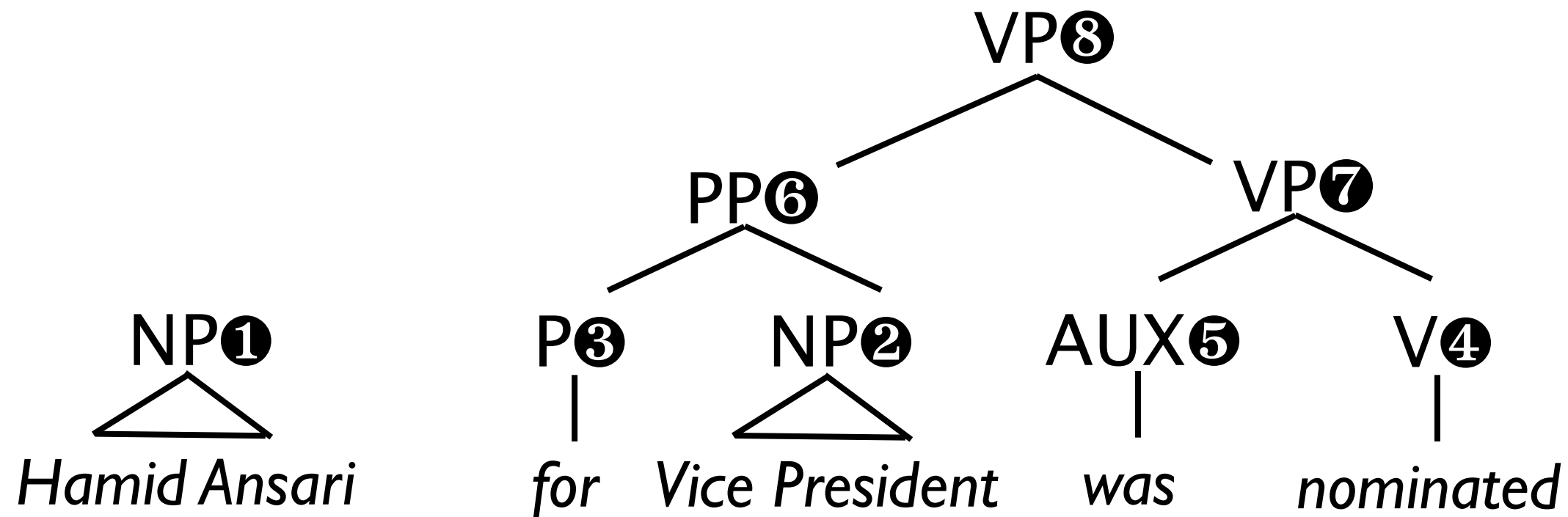
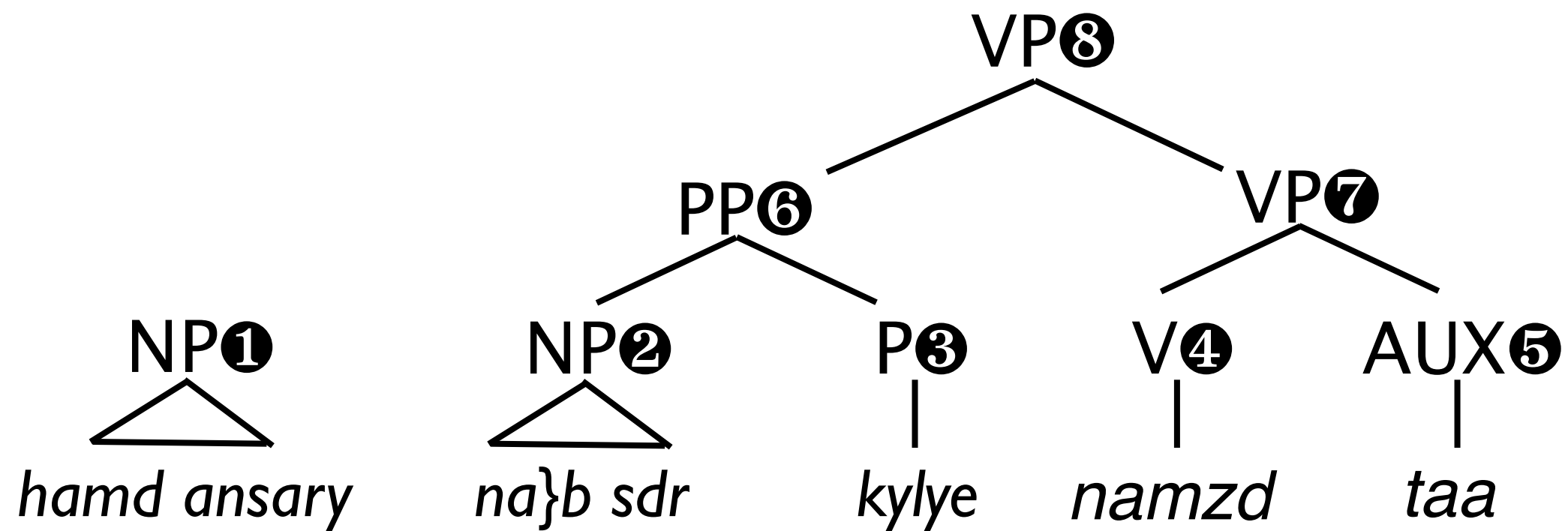
P**3**
|
for

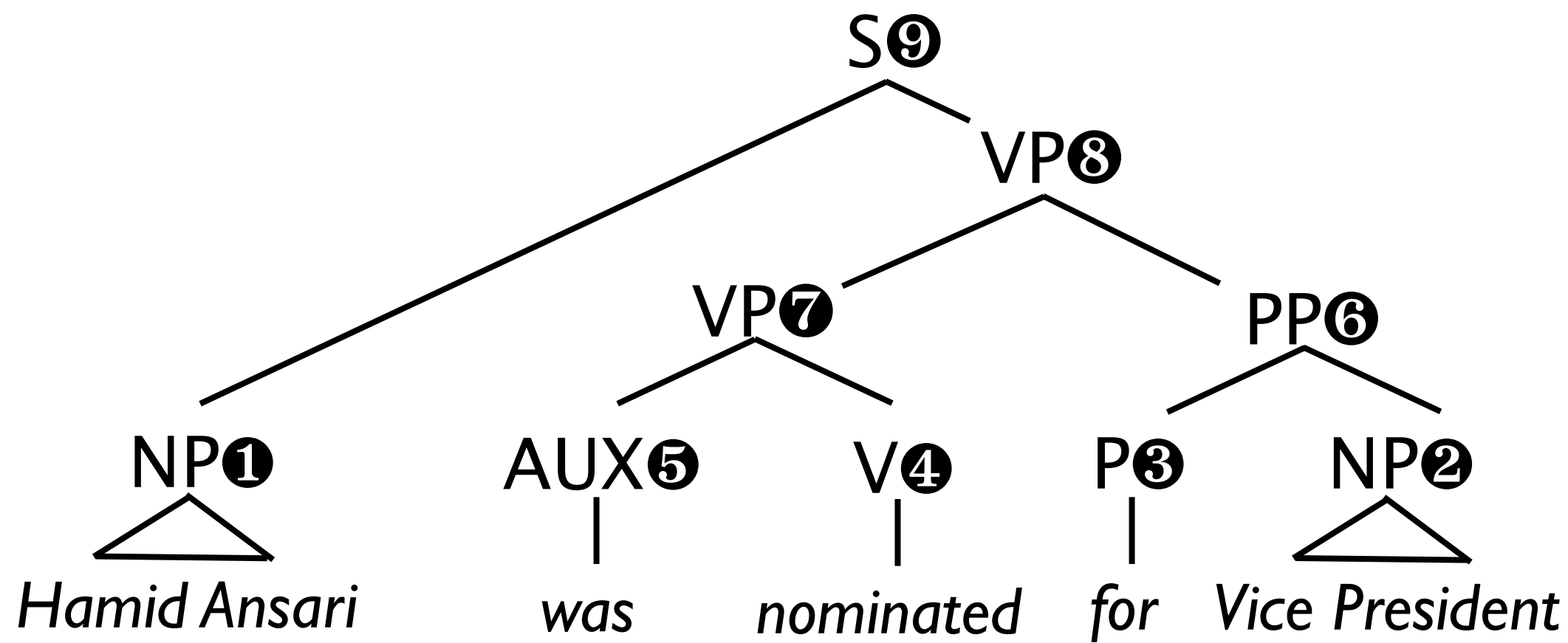
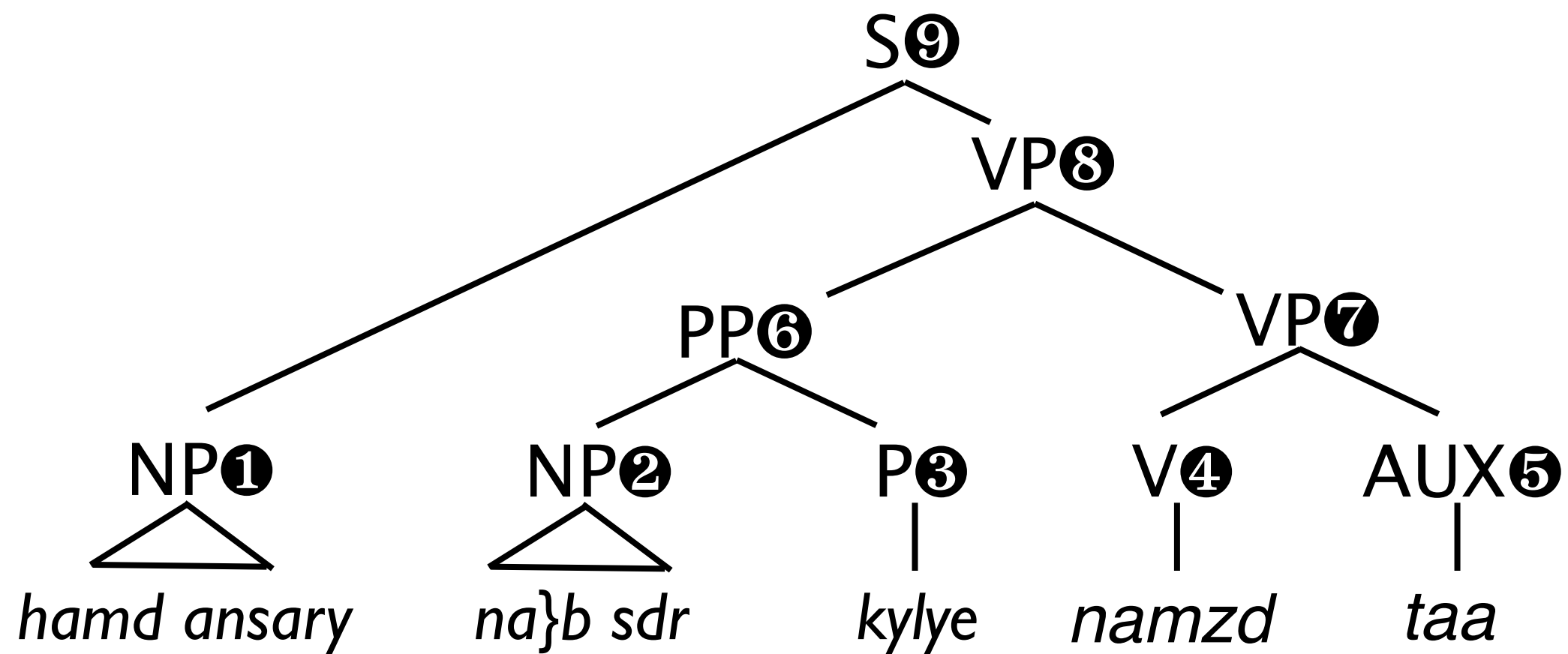
V**4**
|
nominated

AUX**5**
|
was

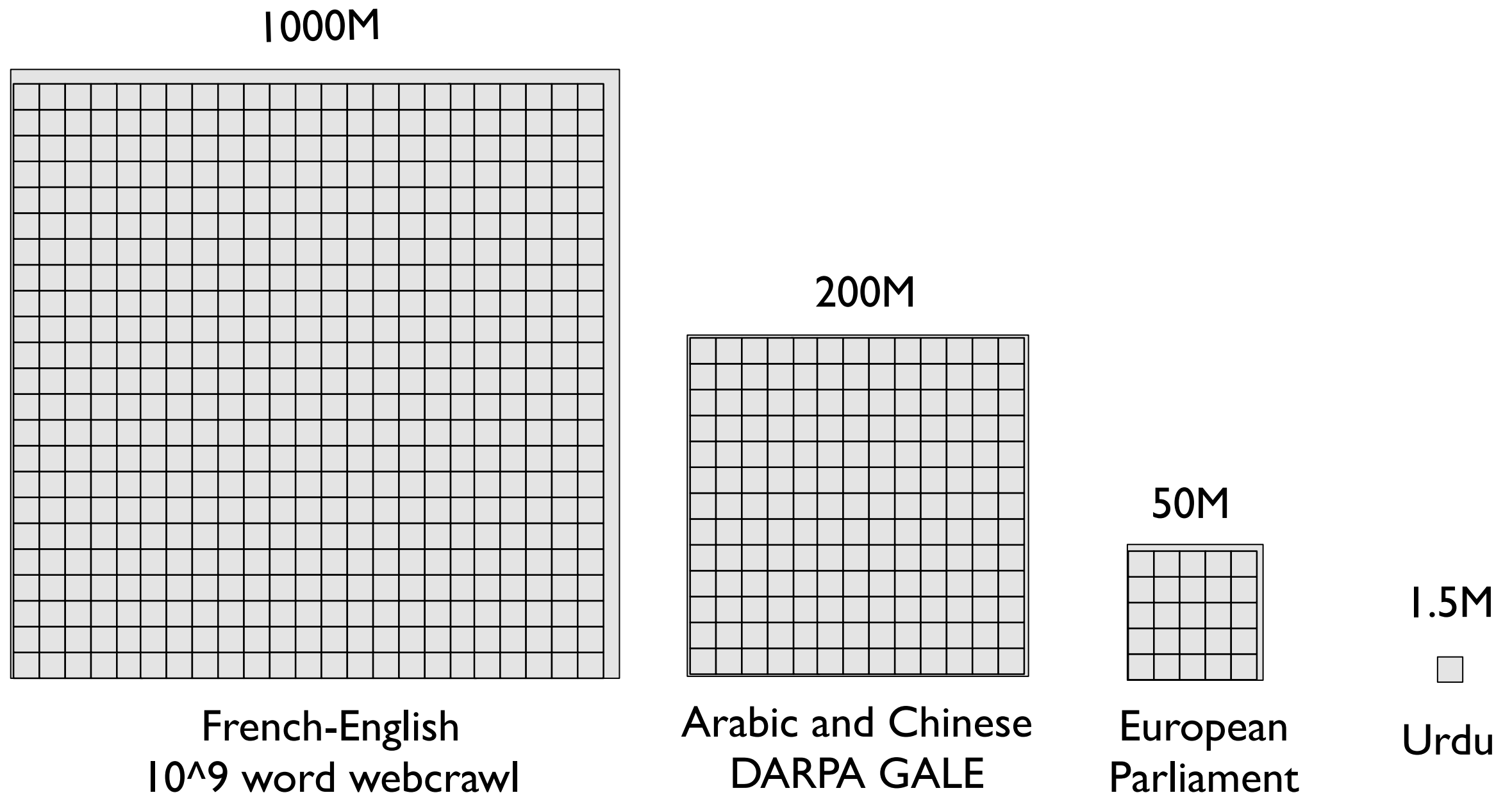




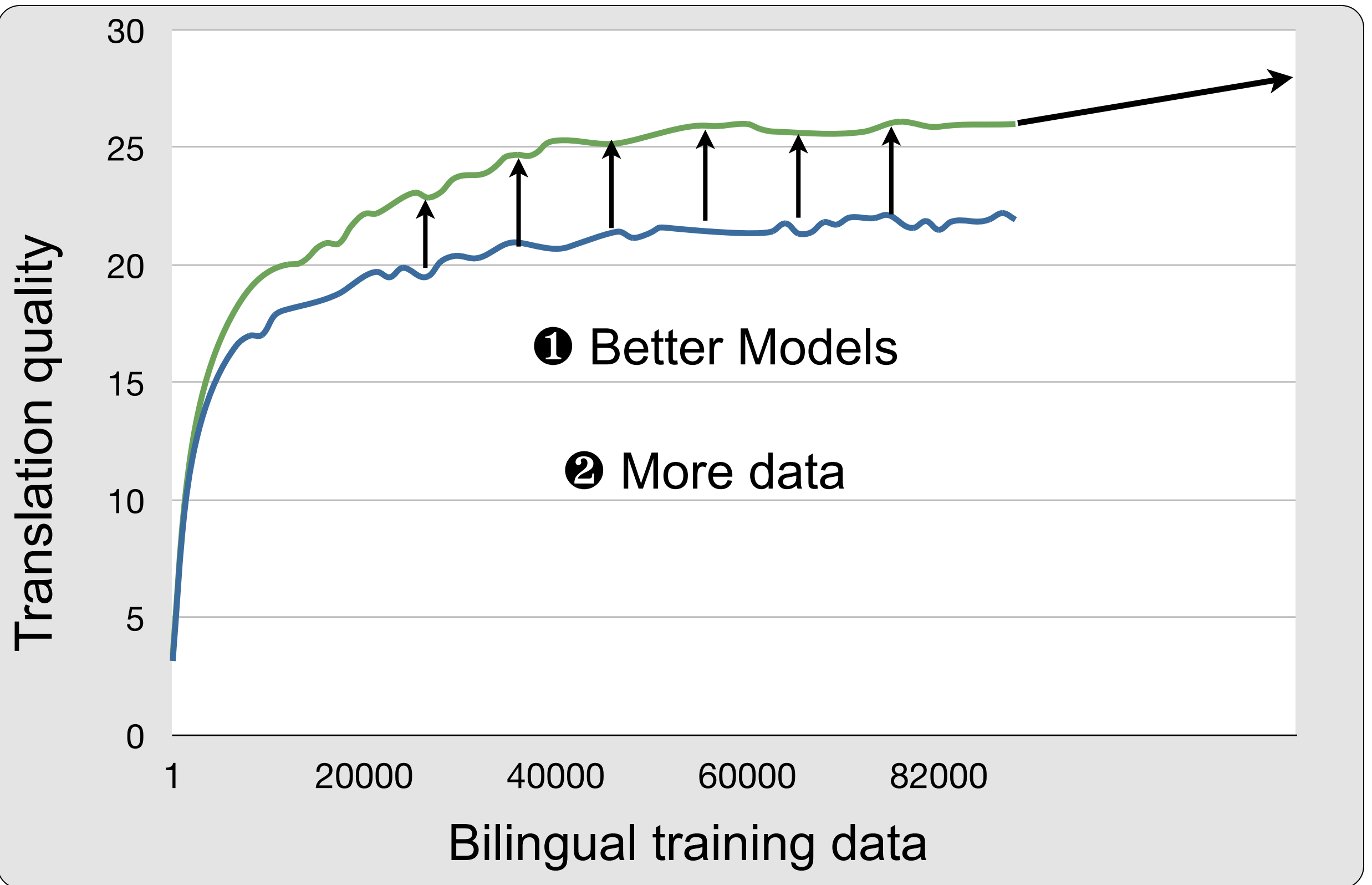




Training data varies by language



How to Improve MT



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
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

LDC Translation

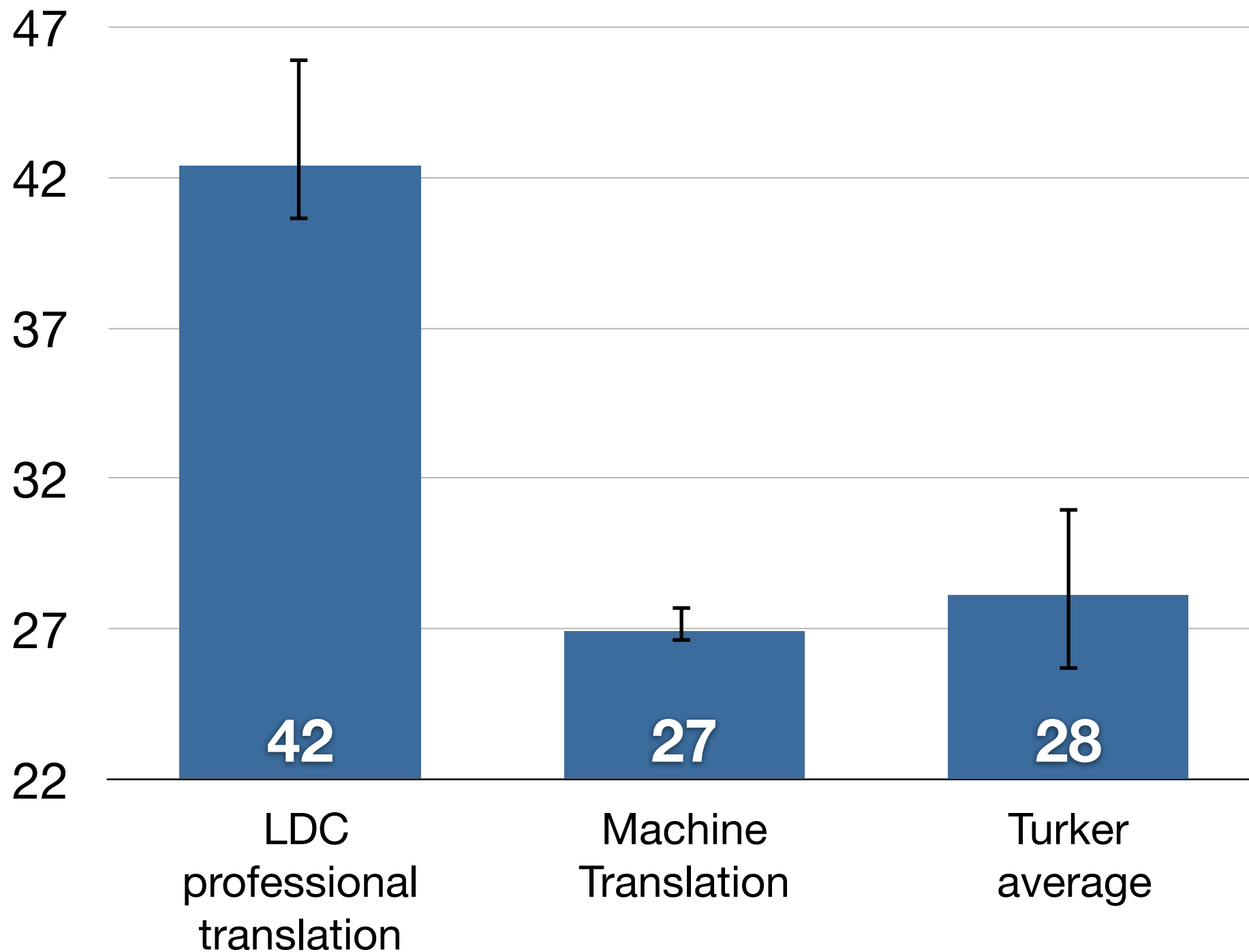
In 1994, the remains of
early human beings who
were believed to be eight
hundred thousand years
old were discovered who
were given the name
homo antecessor meaning
the 'founder man'. Prior to
this, the six hundred
thousand years old man,
called homo
heidelebergensis in
scientific terms, was
believed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



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 mice.	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 mice 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 mice.	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 mice 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 mice.	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 mice 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۔

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۔

۱. ملک بالت تب اب انہوں نے بنائے، عربوں نے کھنڈے منگولوں نے

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.

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.

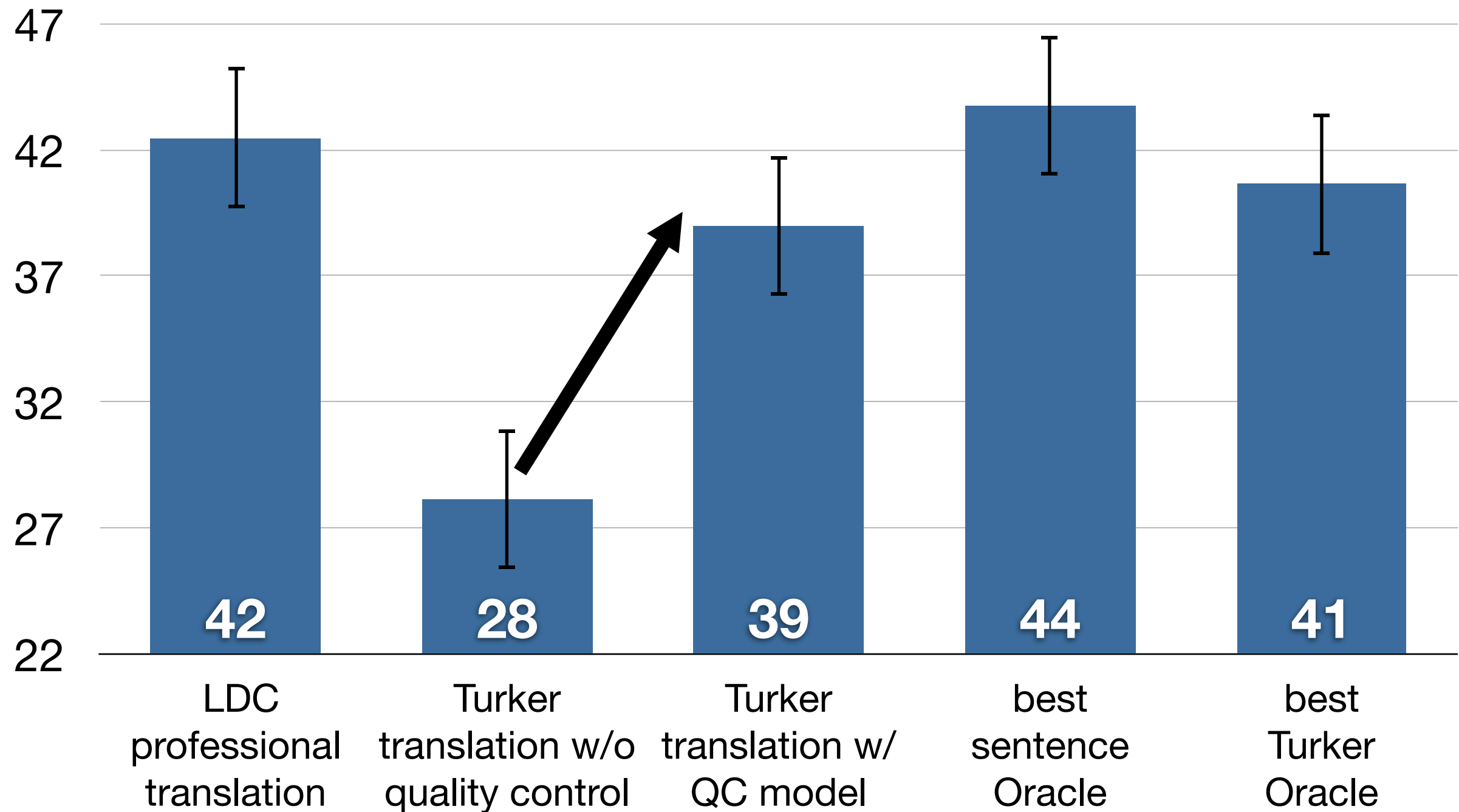
<input type="radio"/>	The research proved this old talk that decrease eating is useful in fever.
<input type="radio"/>	Research disproved the old axiom that " It is better to fast during fever"
<input type="radio"/>	research has proven this old myth wrong that its better to fast during fever.
<input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

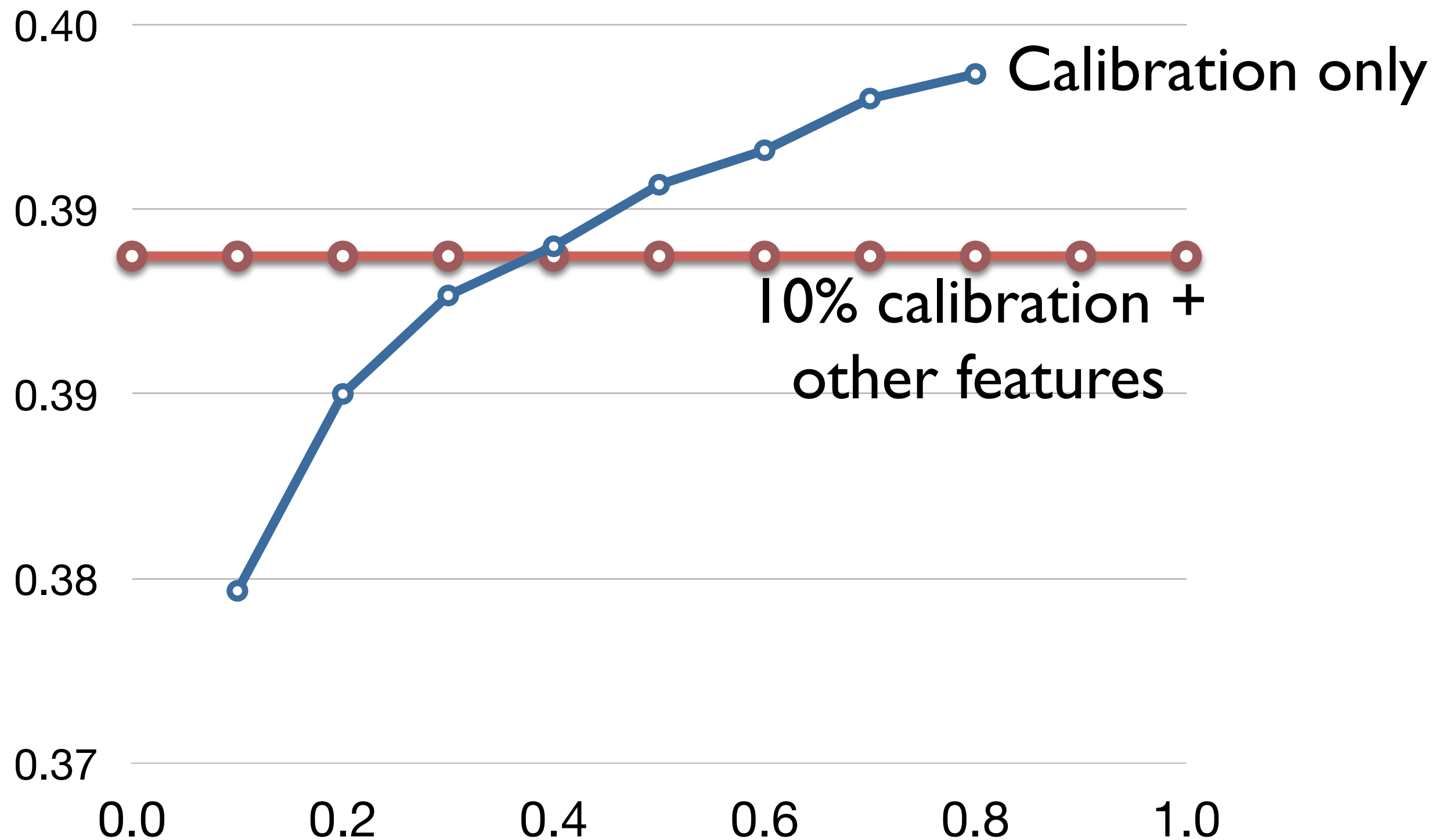
<input type="radio"/>	According to the Researchers of the State of Michigan University diet taking rats have less strength to fight with the infection and also the chances of death increased whereas earlier those rats on common diet had more strength to fight with the flu virus.
-----------------------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

Professional Quality from Non-Professionals

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



% of gold standard data used for calibration



ESL editing HIT

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

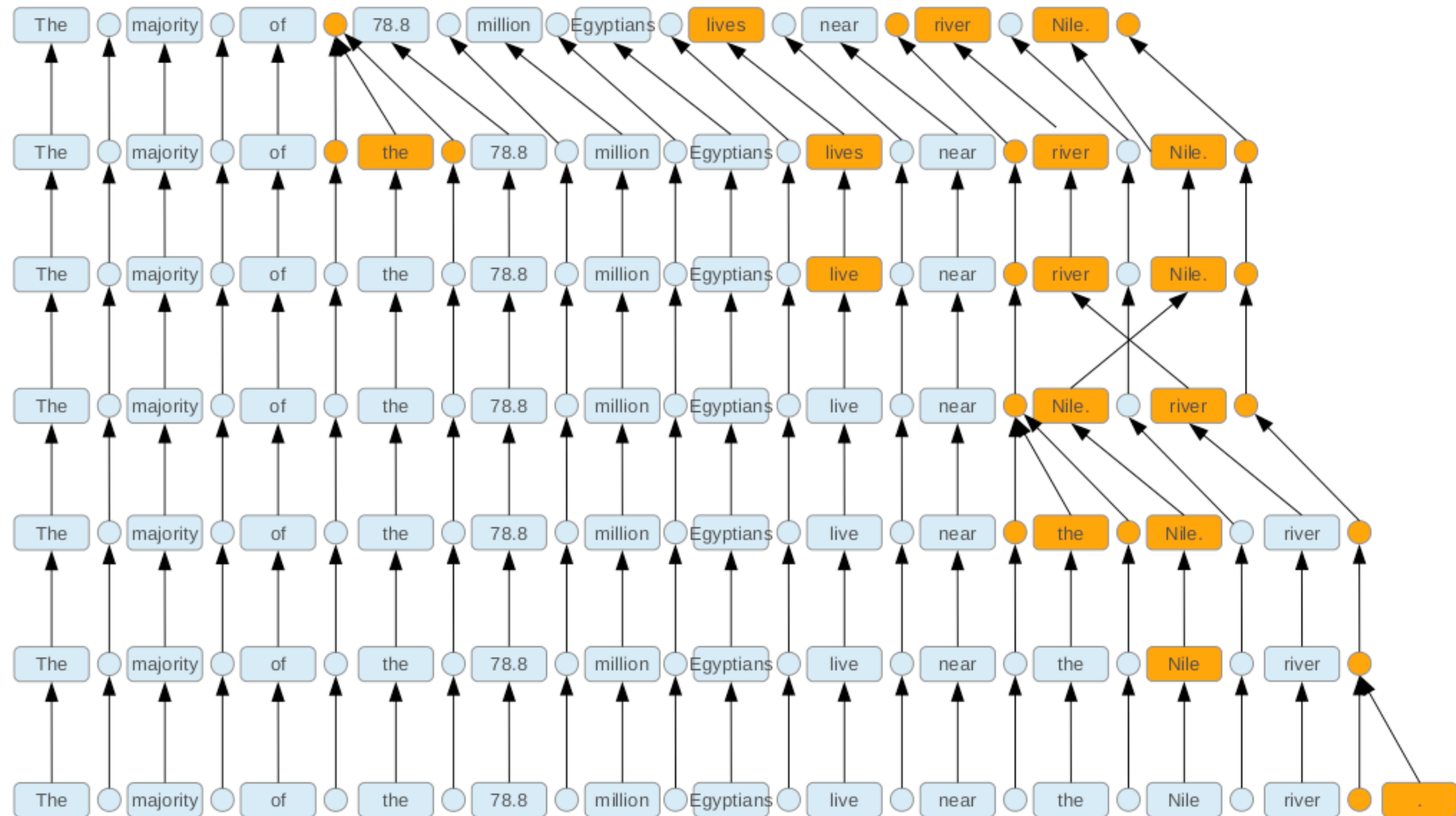
are thousands ^v the
There is ~~thousand~~ of animals living in Sri Lanka
~~which includes several Sri Lanka originated animals~~ .
including native species

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

ESL editing HIT

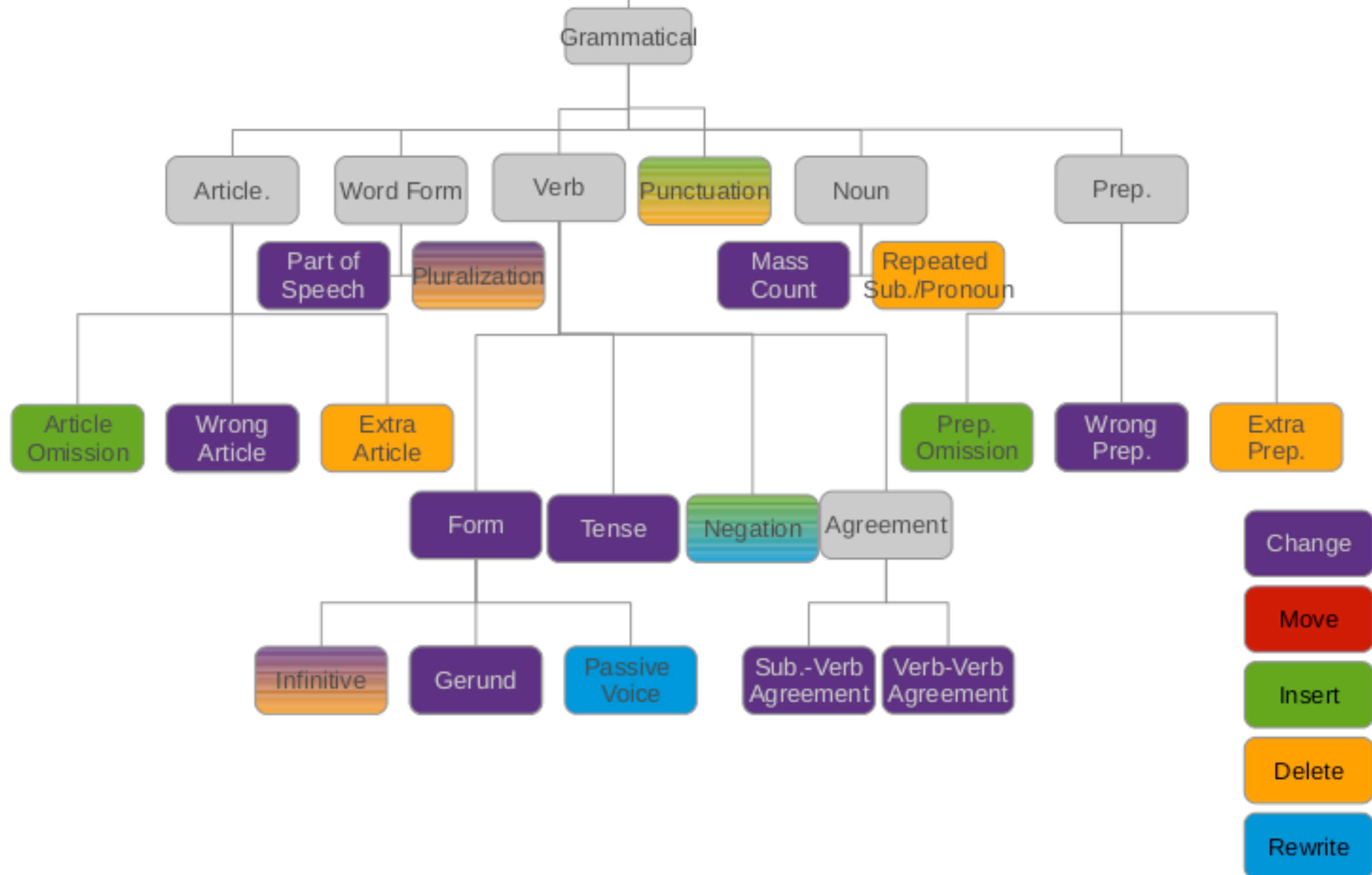
- Agreement no longer works for QC
 - Most agreement happens between lazy workers!
- Create a sentence with a known set of corrections:
 - Start with a grammatical English sentence
 - Make several transformations to it
 - Break subject-verb agreement / change prepositions / etc
 - Measure how many transformations are fixed
- Easier if we require structured corrections

Structured Edits



ESL Error Types

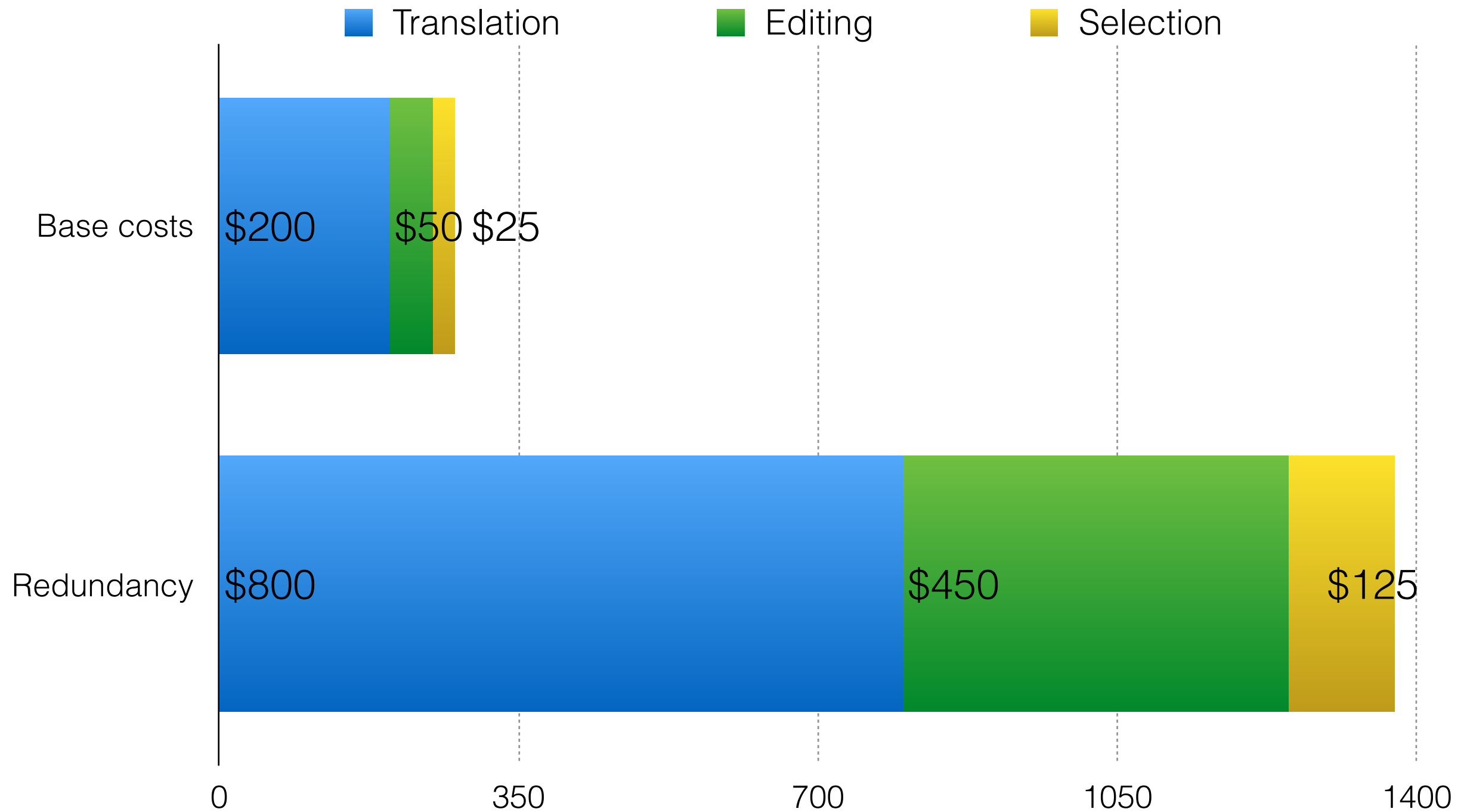
ESL Errors



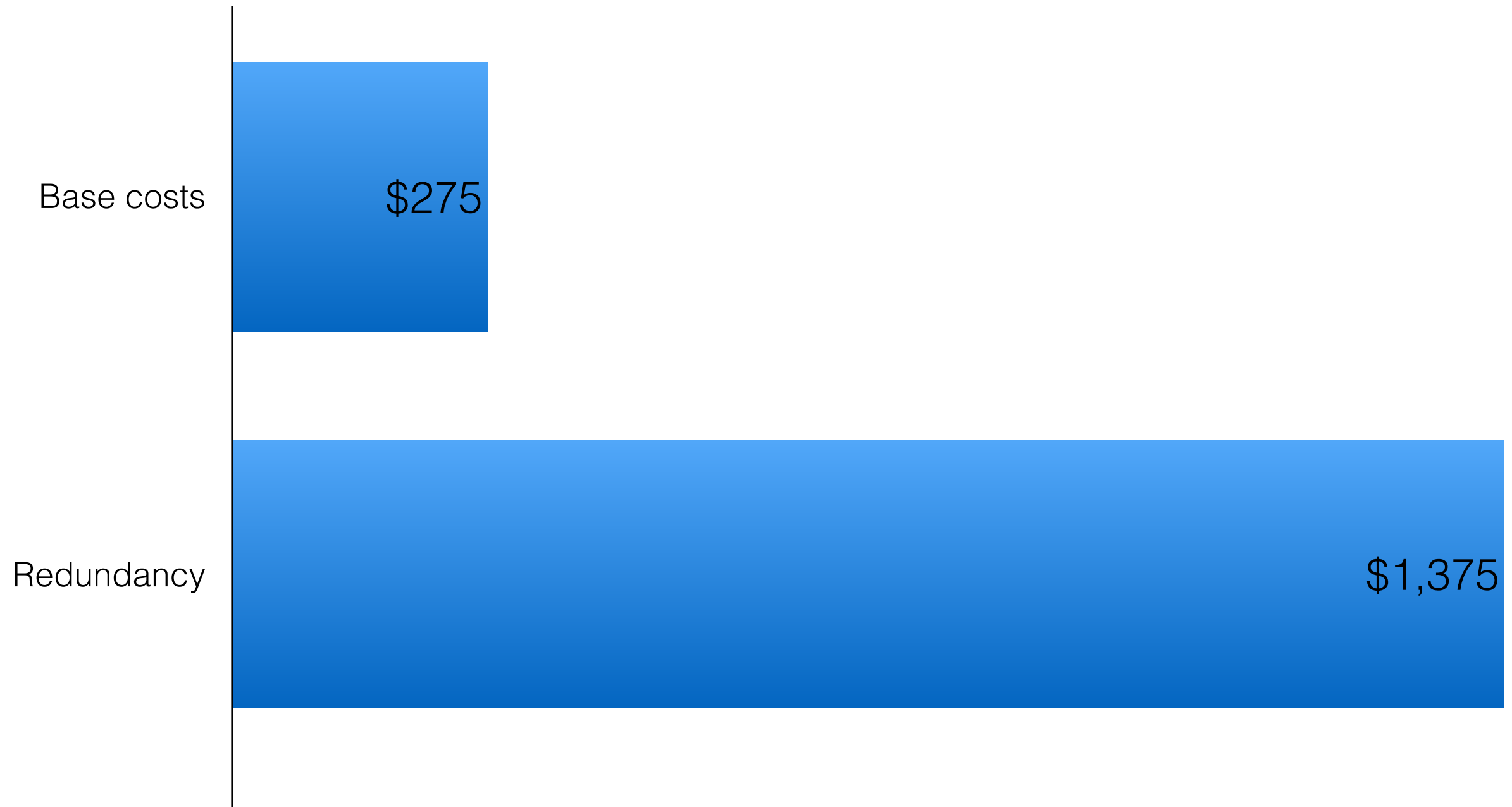
Quality wrap up

- 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

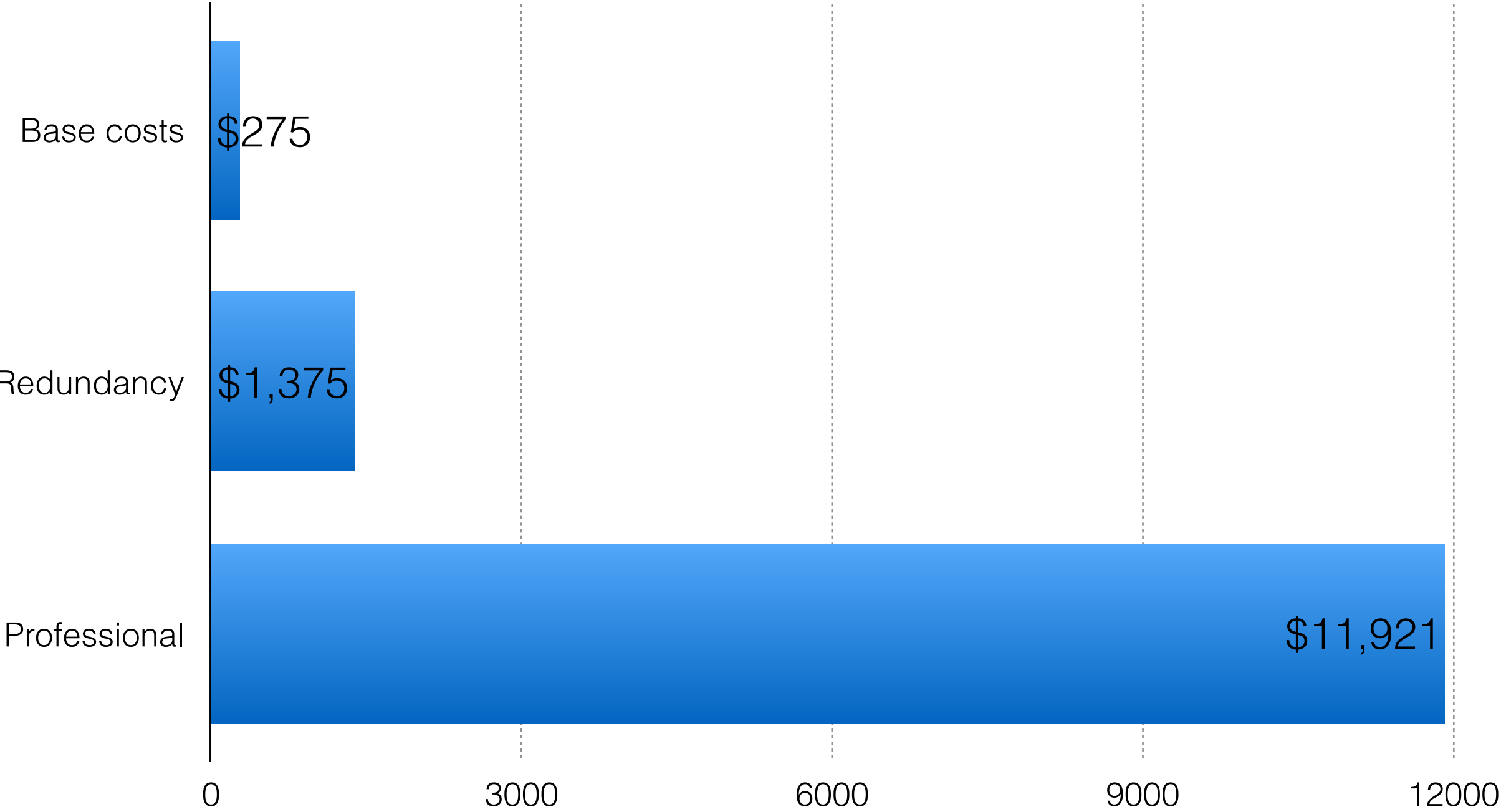


Cost of Translation on MTurk



Profe

Cost of Translation versus Professionals

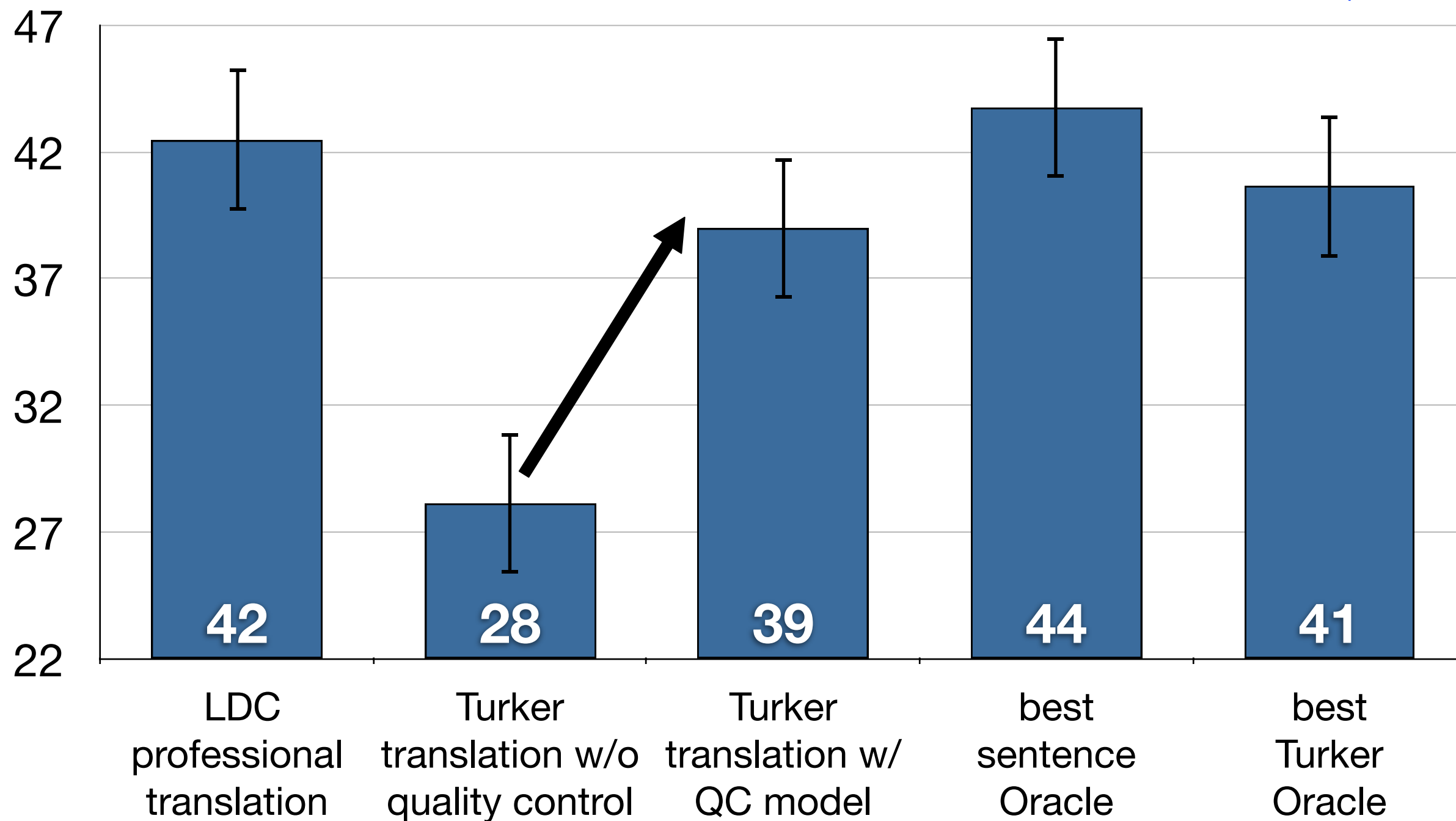


Cost of Translation on MTurk

- 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

Professional Quality from Non-Professionals

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

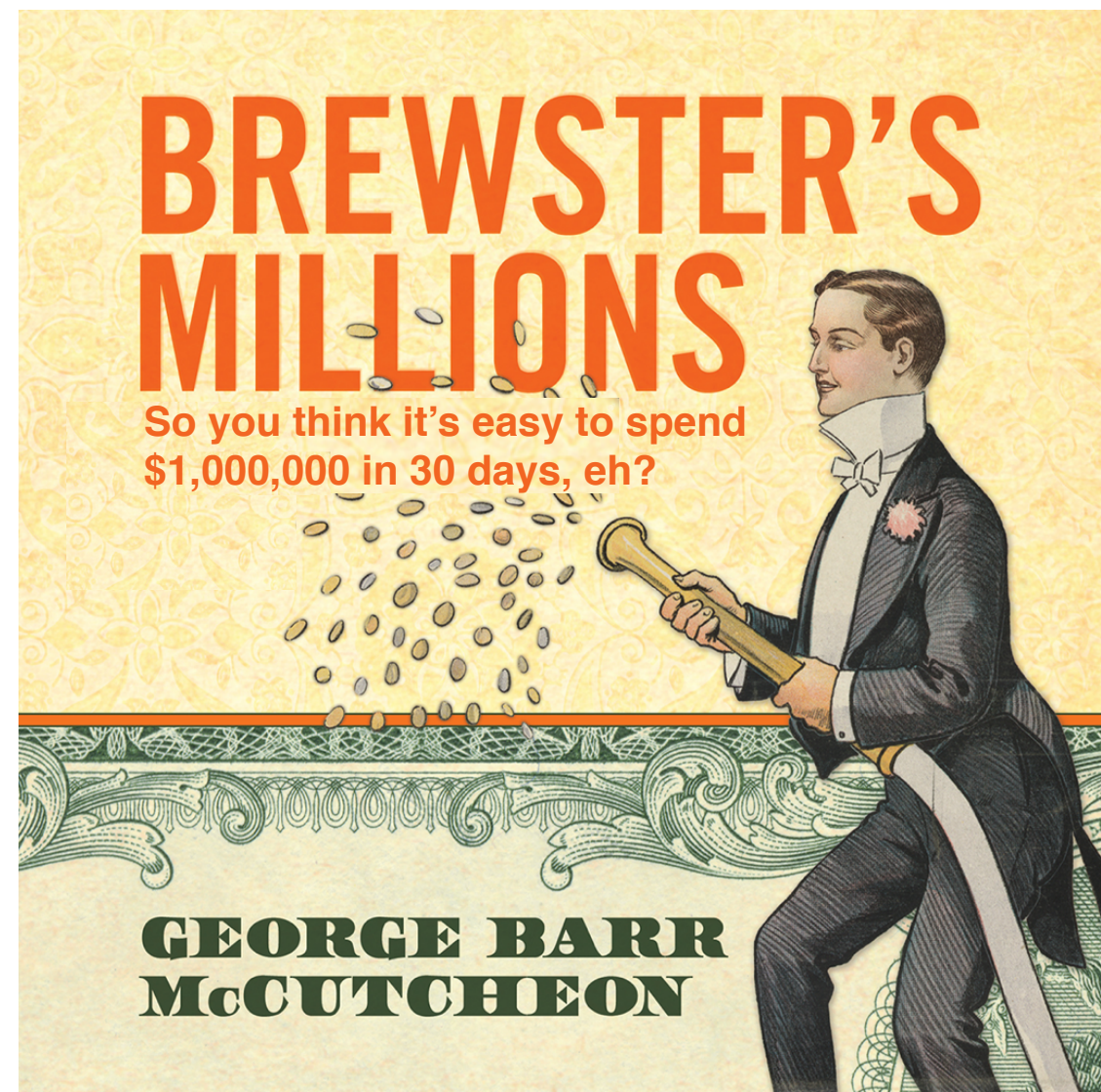


\$100,000 challenge

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



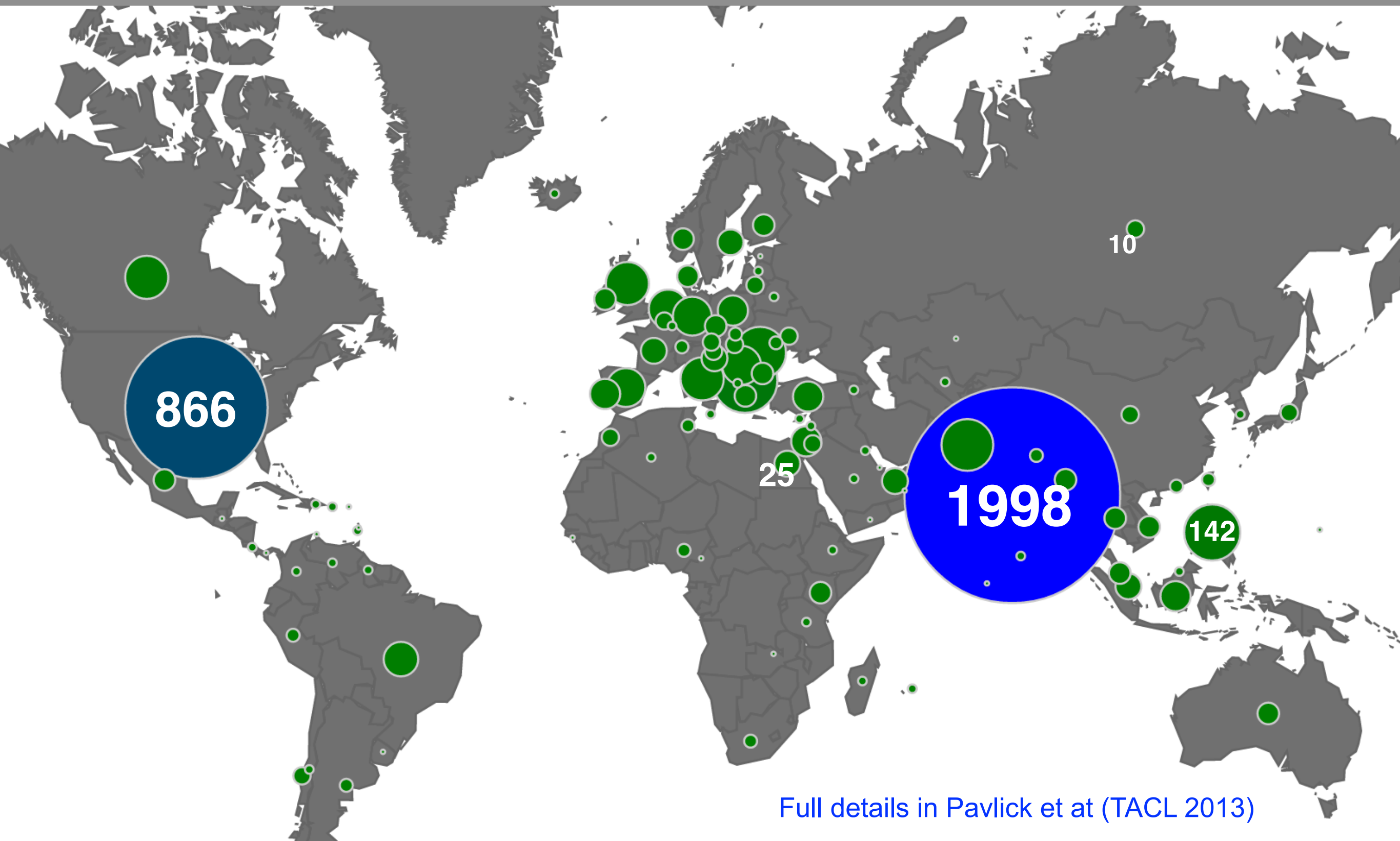
Leaderboard

Mechanical Turk Monitor: Top-1000 Recent Requesters

General Top requesters Arrivals Completed Search About

	Requester ID	Requester	#Task	#HITs	Rewards
1	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 al (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



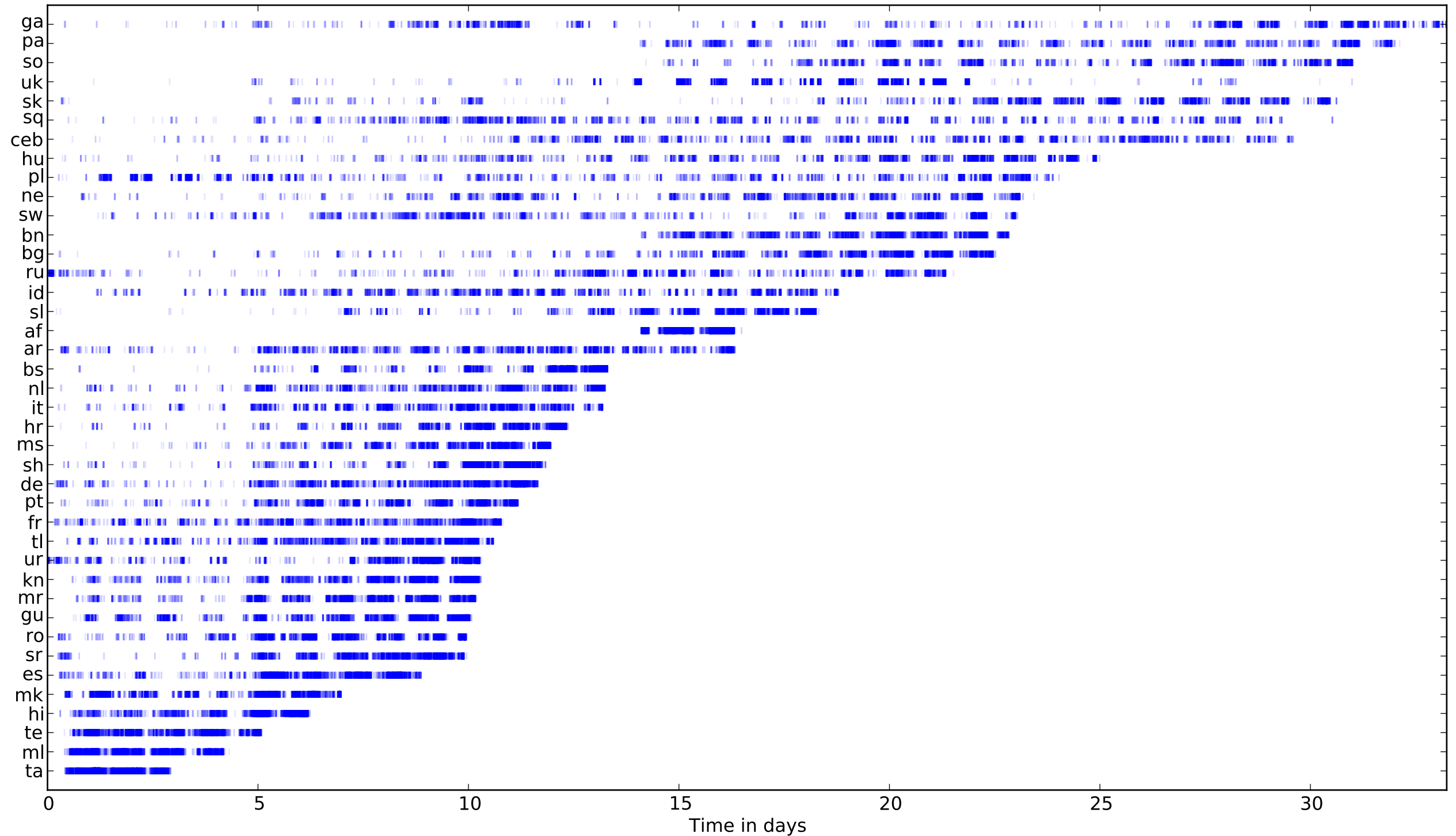
WIKIPEDIA

- 100 languages with the most articles
- 1,000 most viewed articles
- 10,000 most frequent words

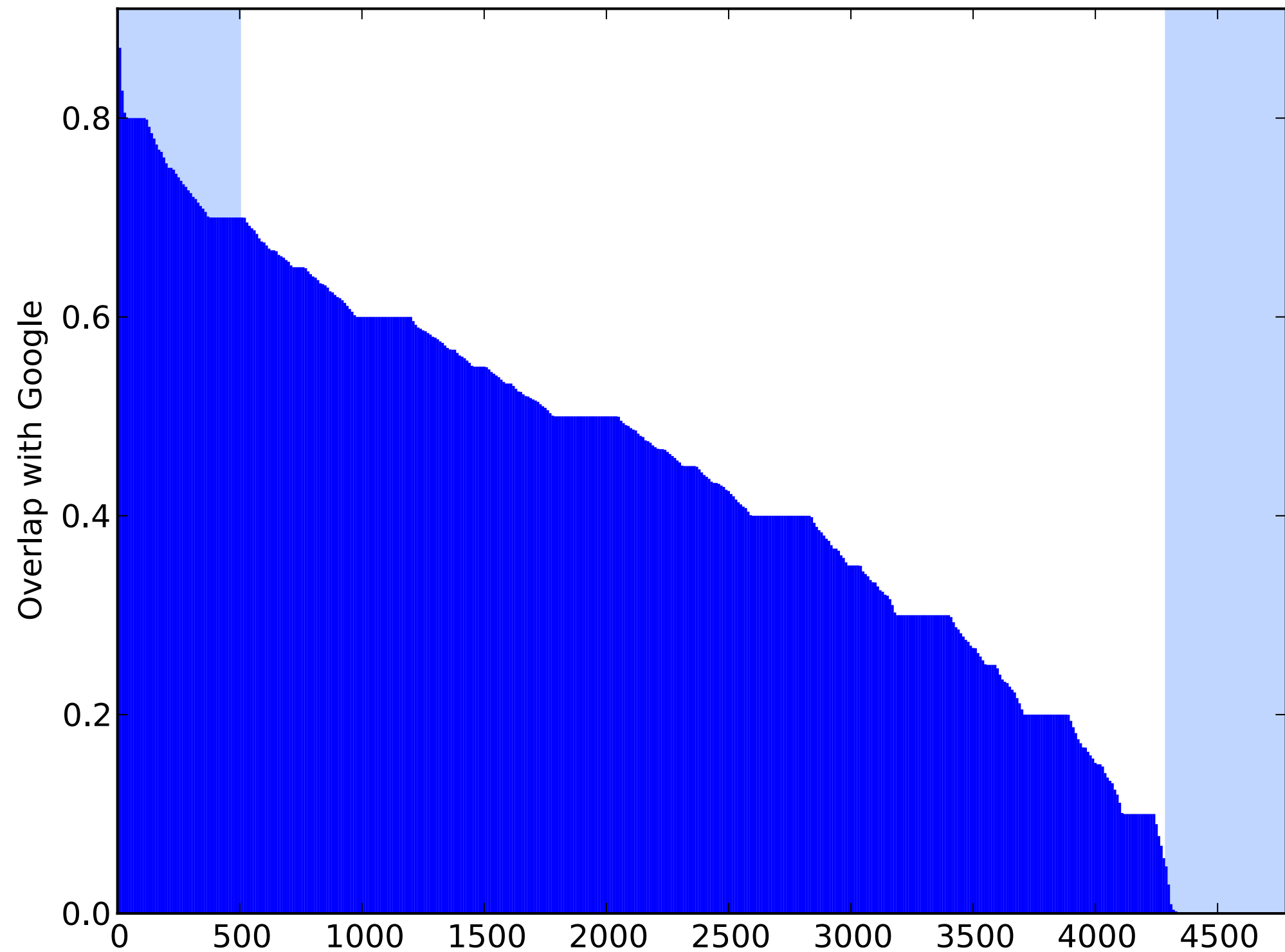


Afrikaans	Czech	Italian	Norwegian (bokmal)	Swahili
Albanian	Danish	Japanese	Norwegian (nynorsk)	Swedish
Amharic	Dutch	Javanese	Pashto	Tagalog
Arabic	Esperanto	Kannada	Persian	Tamil
Aragonese	Finnish	Kapampangan	Piedmontese	Tatar
Armenian	French	Kazakh	Polish	Telugu
Asturian	Galician	Korean	Portuguese	Thai
Azerbaijani	Georgian	Kurdish	Punjabi	Tibetan
Basque	German	Latvian	Quechua	Turkish
Belarusian	Greek	Lithuanian	Romanian	Ukrainian
Bengali	Gujarati	Low saxon	Russian	Urdu
Bishnupriya	Haitian	Luxembourgish	Serbian	Uzbek
Bosnian	Hebrew	Macedonian	Serbo-croatian	Vietnamese
Breton	Hindi	Malagasy	Sicilian	Walloon
Bulgarian	Hungarian	Malayalam	Sindhi	Waray-waray
Catalan	Icelandic	Malay	Slovak	Welsh
Cebuano	Ido	Marathi	Slovenian	West frisian
Central_bicolano	Ilokano	Neapolitan	Somali	Wolof
Chinese	Indonesian	Nepali	Spanish	Yoruba
Croatian	Irish	Newar/Nepal Bhasa	Sundanese	Zazaki

Time to complete 10,000 translations



Cheating with Google Translate



Quality Control

Wikipedia page titles connected by inter-language links, removing pairs for which

- either title was longer than a single word
- English word didn't appear in Wordnet
- English Wikipedia page was a subcategory of person or place
- titles were identical or one was a substring of the other

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 \(plant\)](#)

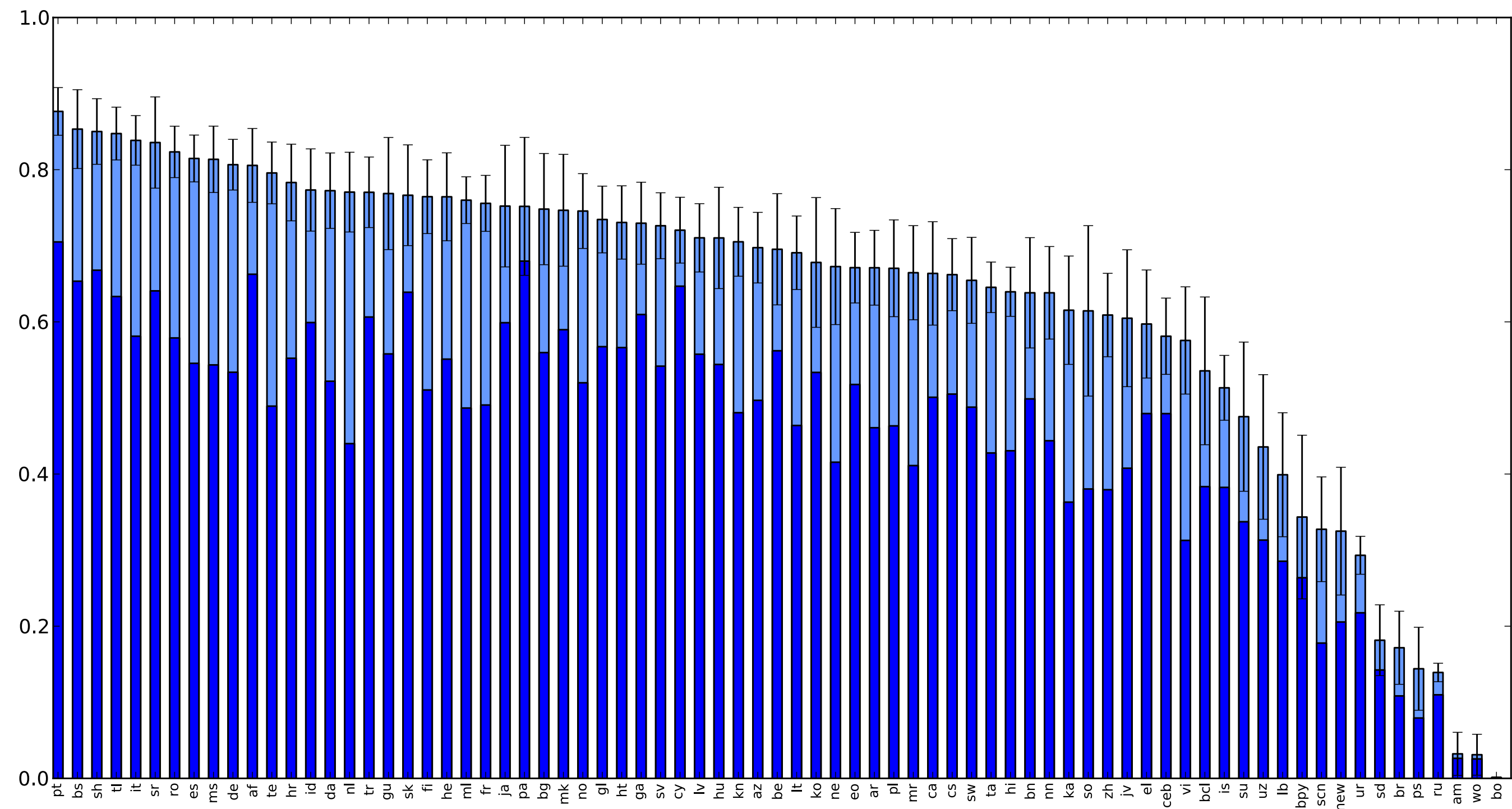
- ☐ Yes ☐ No ☐ Related but not synonymous
☐ Word is misspelled

[demo](#) and [show](#)

- ☐ Yes ☐ No ☐ Related but not synonymous
☐ Word is misspelled

Translation Quality

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



Translation Quality

	Avg. Turker quality (# Ts)		Primary locations	Primary locations
	In region	Out of region	of Turkers in region	of Turkers out of region
Hindi	0.63 (296)	0.69 (7)	India (284) UAE (5) UK (3)	Saudi Arabia (2) Russia (1) Oman (1)
Tamil	0.65 (273) **	0.25 (2)	India (266) US (3) Canada (2)	Tunisia (1) Egypt (1)
Malayalam	0.76 (234)	0.83 (2)	India (223) UAE (6) US (3)	Saudi Arabia (1) Maldives (1)
Spanish	0.81 (191)	0.84 (18)	US (122) Mexico (16) Spain (14)	India (15) New Zealand (1) Brazil (1)
French	0.75 (170)	0.82 (11)	India (62) US (45) France (23)	Greece (2) Netherlands (1) Japan (1)
Chinese	0.60 (116)	0.55 (21)	US (75) Singapore (13) China (9)	Hong Kong (6) Australia (3) Germany (2)
German	0.82 (91)	0.77 (41)	Germany (48) US (25) Austria (7)	India (34) Netherlands (1) Greece (1)
Italian	0.86 (90) *	0.80 (42)	Italy (42) US (29) Romania (7)	India (33) Ireland (2) Spain (2)
Amharic	0.14 (16) **	0.01 (99)	US (14) Ethiopia (2)	India (70) Georgia (9) Macedonia (5)
Kannada	0.70 (105)	NA (0)	India (105)	
Arabic	0.74 (60) **	0.60 (45)	Egypt (19) Jordan (16) Morocco (9)	US (19) India (11) Canada (3)
Sindhi	0.19 (96)	0.06 (9)	India (58) Pakistan (37) US (1)	Macedonia (4) Georgia (2) Indonesia (2)
Portuguese	0.87 (101)	0.96 (3)	Brazil (44) Portugal (31) US (15)	Romania (1) Japan (1) Israel (1)
Turkish	0.76 (76)	0.80 (27)	Turkey (38) US (18) Macedonia (8)	India (19) Pakistan (4) Taiwan (1)
Telugu	0.80 (102)	0.50 (1)	India (98) US (3) UAE (1)	Saudi Arabia (1)
Irish	0.74 (54)	0.71 (47)	US (39) Ireland (13) UK (2)	India (36) Romania (5) Macedonia (2)
Swedish	0.73 (54)	0.71 (45)	US (25) Sweden (22) Finland (3)	India (23) Macedonia (6) Croatia (2)
Czech	0.71 (45) *	0.61 (50)	US (17) Czech Republic (14) Serbia (5)	Macedonia (22) India (10) UK (5)
Russian	0.15 (67) *	0.12 (27)	US (36) Moldova (7) Russia (6)	India (14) Macedonia (4) UK (3)
Breton	0.17 (3)	0.18 (89)	US (3)	India (83) Macedonia (2) China (1)

Language Feasibility

workers	quality	speed	Full details in Pavlick et al (TACL 2013)
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, Neapolitan, Norwegian (Nynorsk), Pashto, Piedmontese, Somali, Sudanese, Swahili, Tatar, Vietnamese, Walloon, Welsh
none			Esperanto, Ido, Kurdish, Persian, Quechua, Wolof, Zazaki

Full Sentence Translations

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

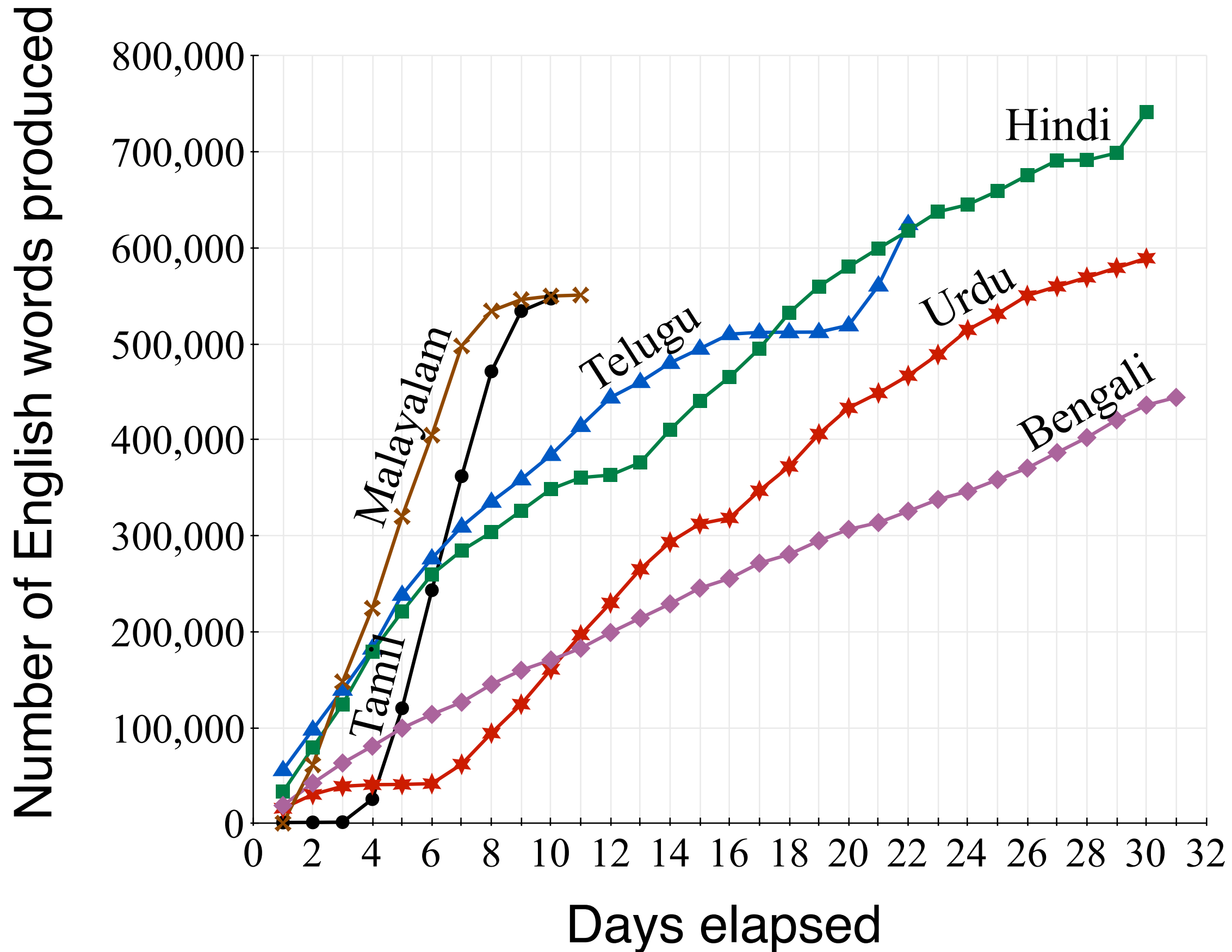
In retribution pakistan also did six nuclear tests on 28 may 1998.

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.

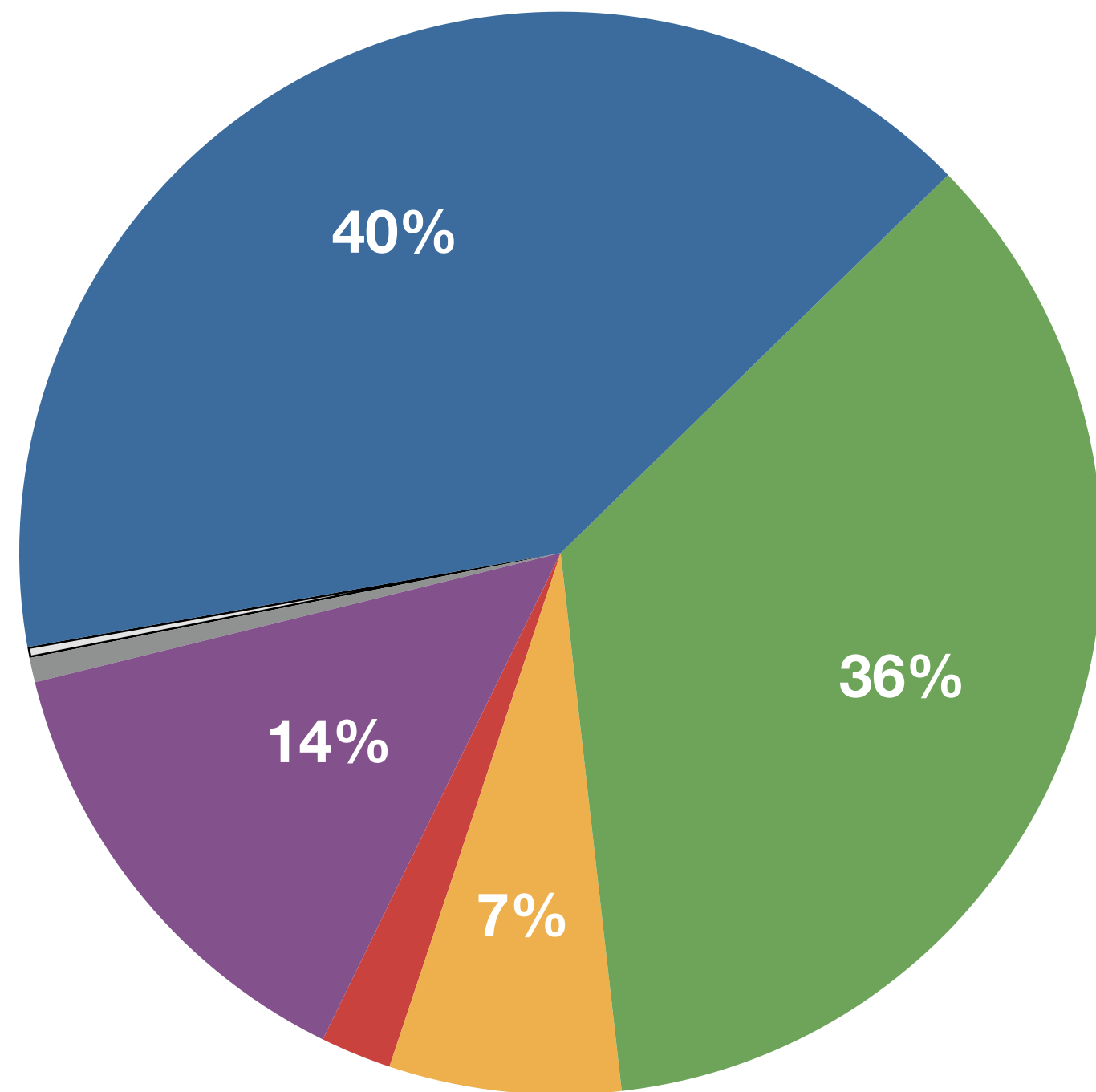
pakistan also did 6 nuclear test in retribution on 28 may, 1998

Rate of Translation

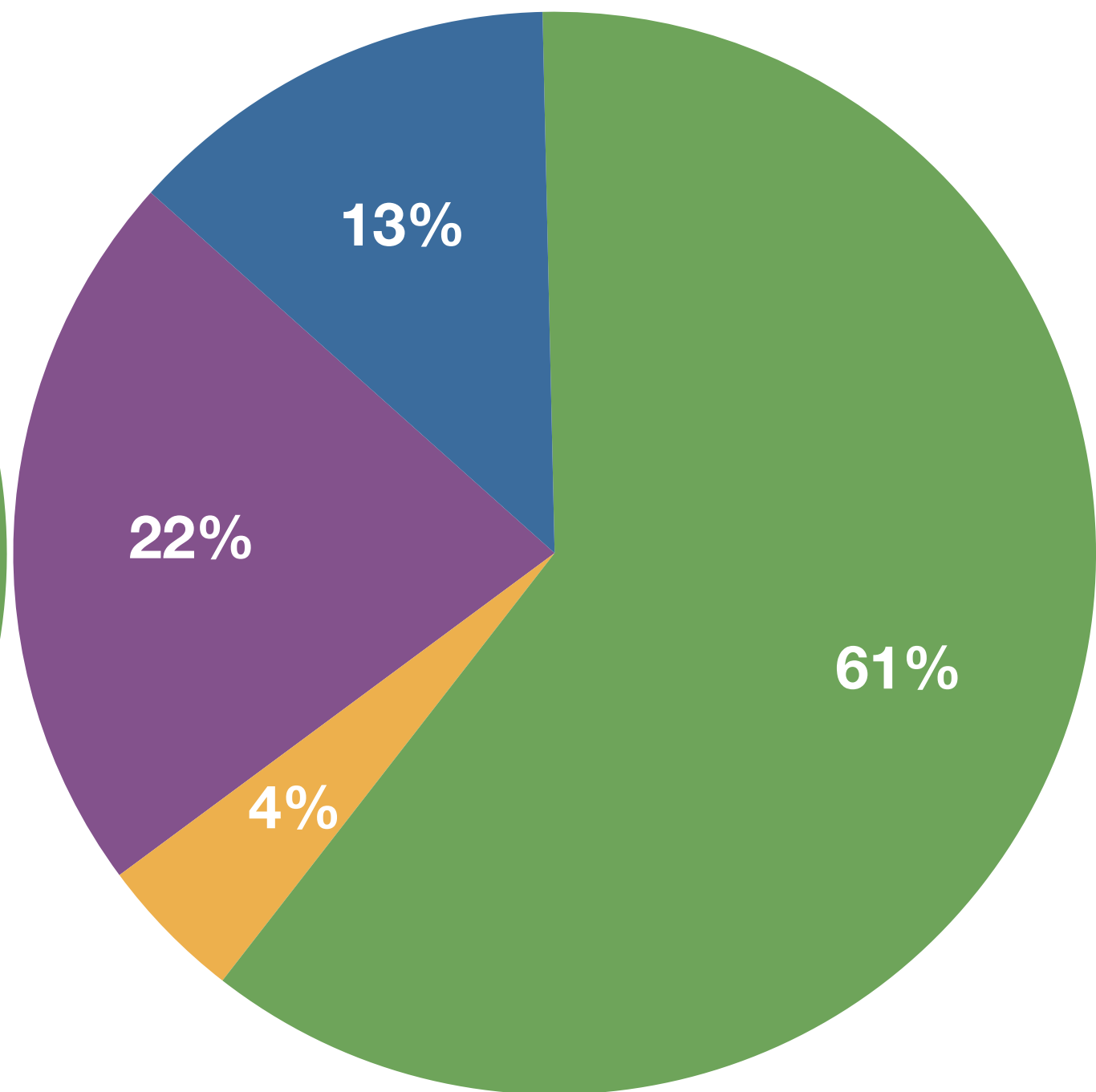


Better Representation of Languages

All Languages



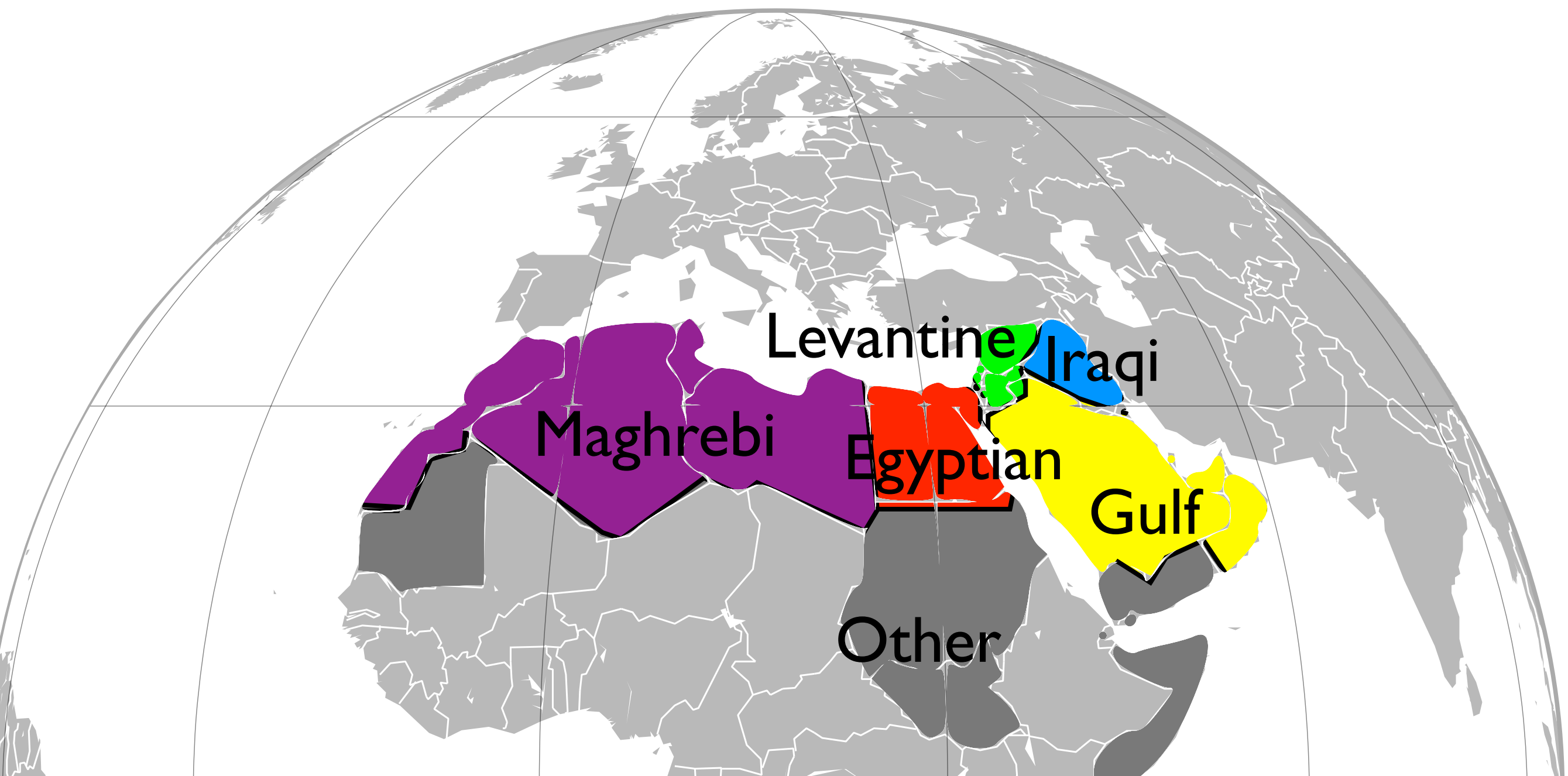
SMT Languages



● 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 التلة من المجرمين تخضع للمحاكمة ؟

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

Levantine: ايمتى رح نشوف هالشلة من المجرمين بتتحاكم ؟

AymtY 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.)



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)

خليك فى حالك وابعد عنا

بواسطة: bebooo

بتاريخ: الخميس، 16 يونيو 2011 - 09:39

EGY

خليك فى حالك انت وابعدوا عنا وانتم اخر ناس تتكلم عن التقارب والعروبة

الجزائر إخوة 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 بها الوطن العربي ، أهلاً ومرحباً بأي تقارب عربي عربي بين الأشقاء في الوطن والمصير ، يسقط الاستعمار الجديد المتمثل في أمريكا والنااتو ومن ورائة آل صهيون

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

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

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 لمنتجع الحكام المخلوعين بالمملكة السعودية وسيبوا الشعوب تبحث في قيام الولايات العربية المتحدة و عاصمتها القدس إنشاء الله (معلش حبيبي هو حلم يقطر يمكن يتحقق في يوم من الأيام من منا يعلم الغيب؟)

MSA

EGY

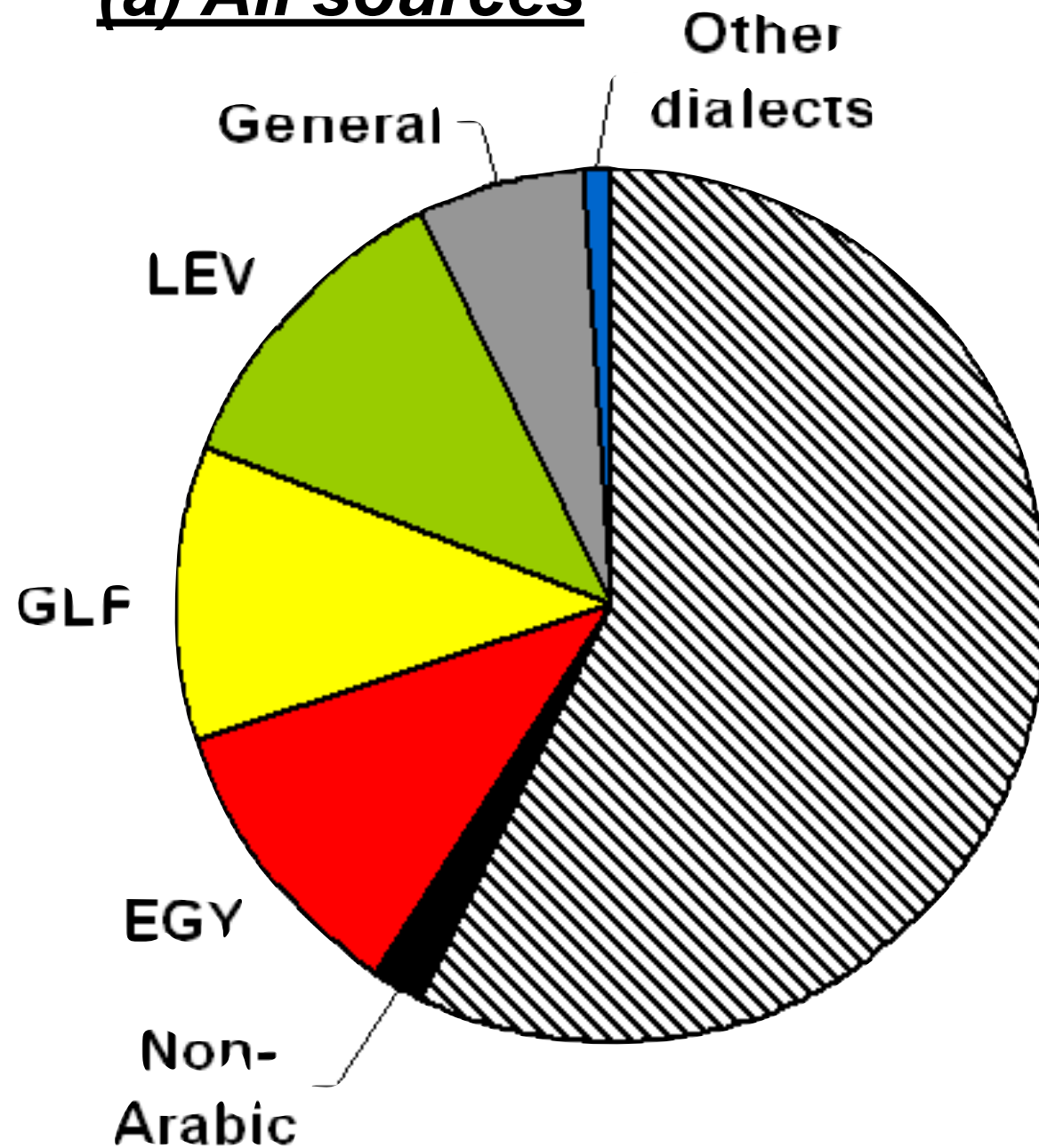
Crowdsourcing Dialect ID

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

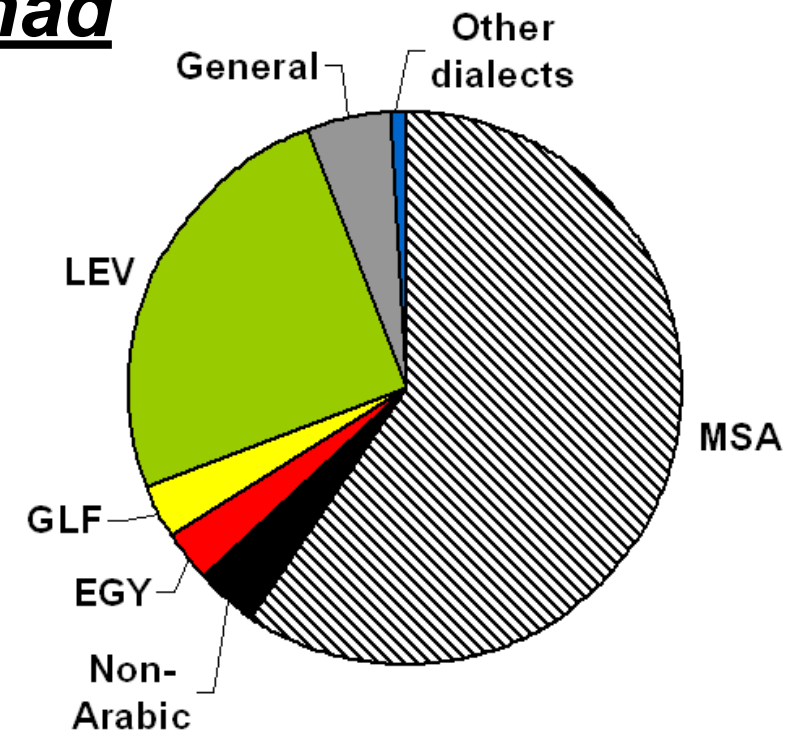
- **Labeled 142k comments** gathered from three online newspapers:
 - Al-Ghad (الغد), a Jordanian newspaper **LEV** **MSA**
 - Al-Riyadh (الرياض), a Saudi newspaper **GLF** **MSA**
 - Al-Youm Al-Sabe' (اليوم السابع), an Egyptian newspaper **EGY** **MSA**
- 59% MSA, **41% dialect**
- Trained classifiers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

Label Distribution

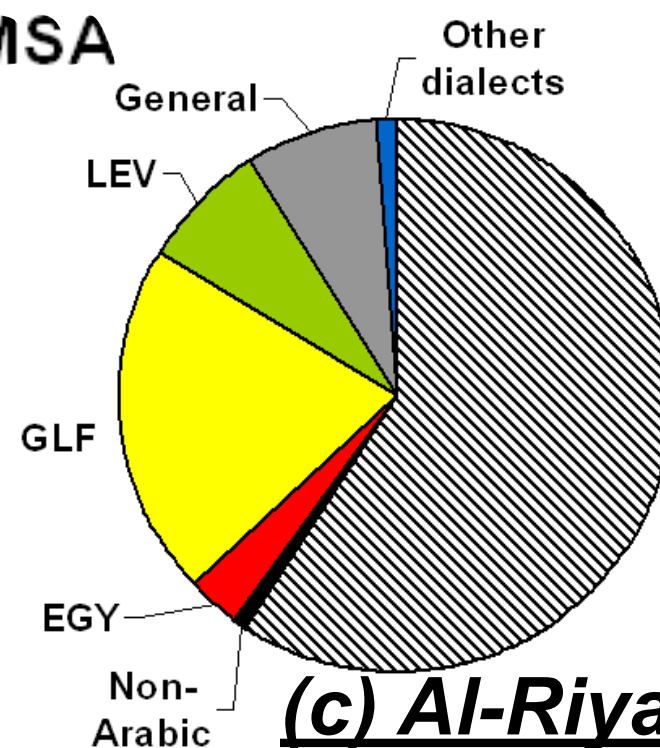
(a) All sources



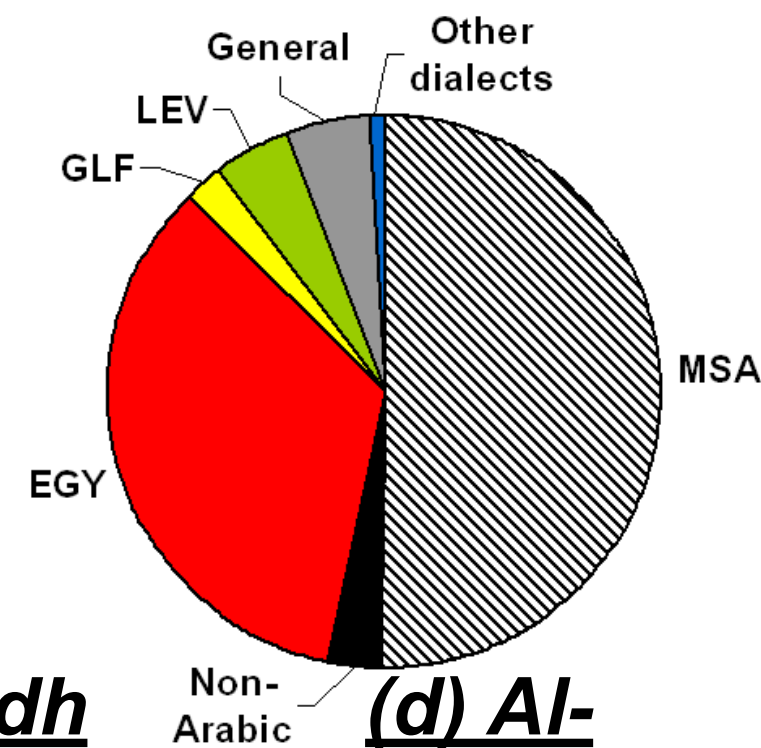
(b) Al-Ghad



MSA

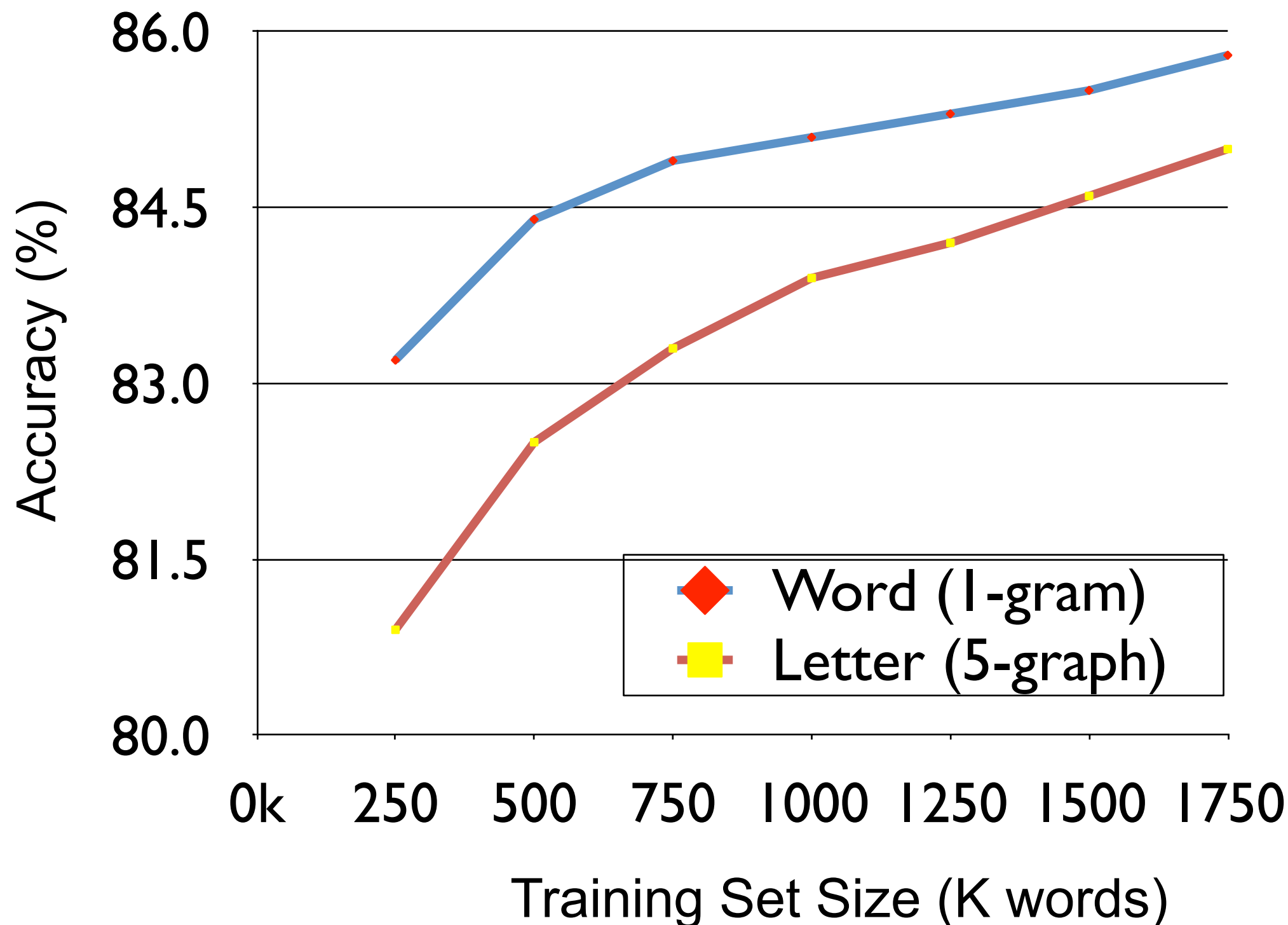


(c) Al-Riyadh



(d) Al-Youm

Automatic dialect ID



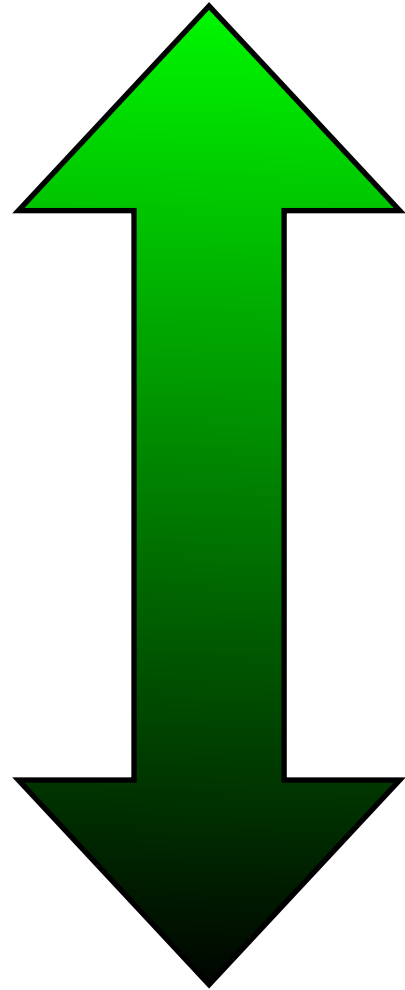
Baselines
Majority
class: **59%**

OOV wrt
Arabic
GigaWord:
66%

1 letter model
trained on our
data: **68%**

“Dialecttness” quantified

Extremely dialectal
Levantine words



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

<i>Al-Ghad</i>			
<i>w</i>		Gloss	<i>DF(w)</i>
\$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
Allnp	اللجنة	the committee	0.095

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

“Dialecttness” 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

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

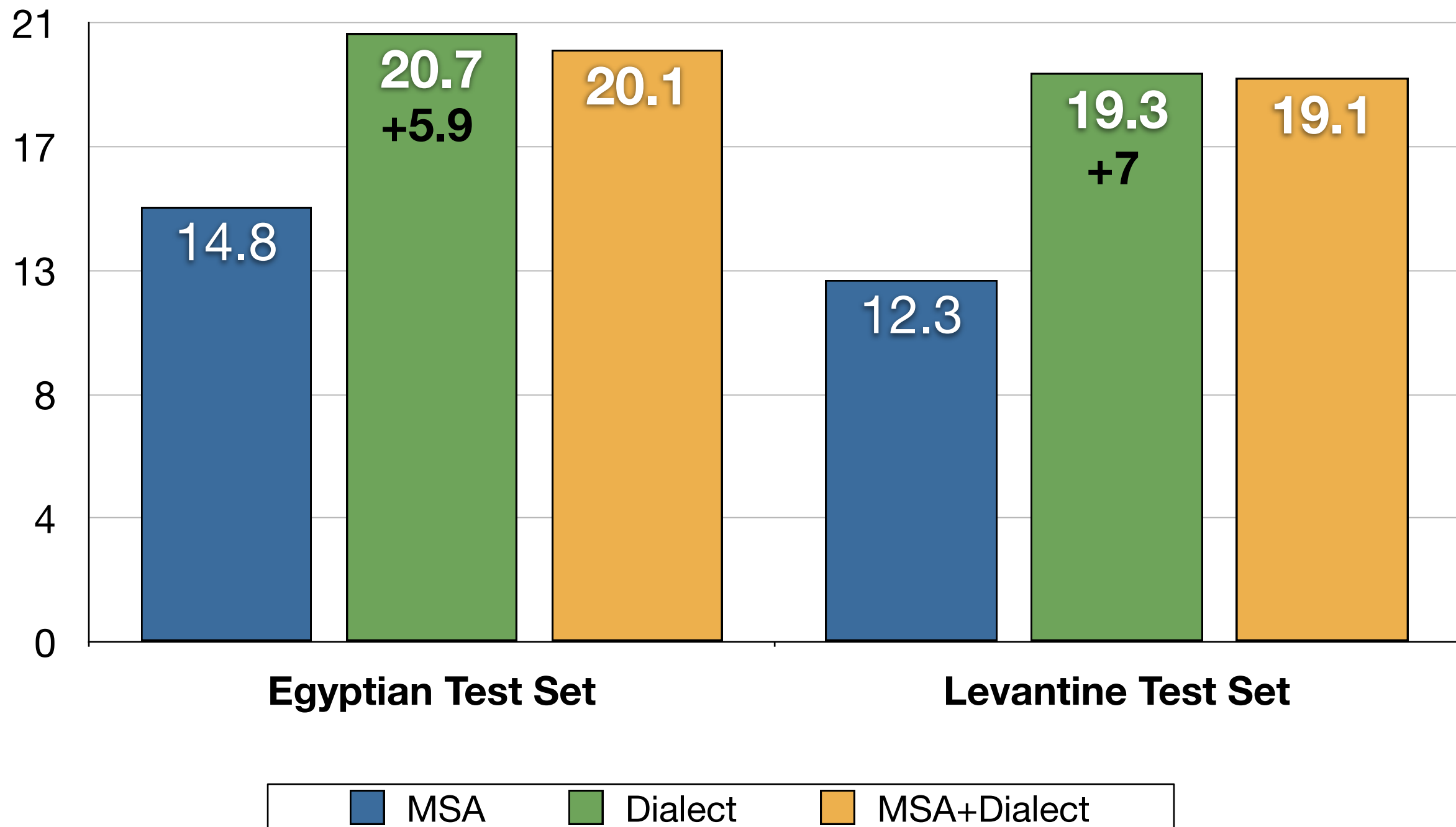
- **121 workers** completed 20+ assignments
- **200,000 words** translated **per week**
- **Trained** BBN's machine translation system

Examples of Dialect Translation

Dialect Input	MSA system	Dialect system	Reference
EGY انت 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?!!
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 كتييير 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

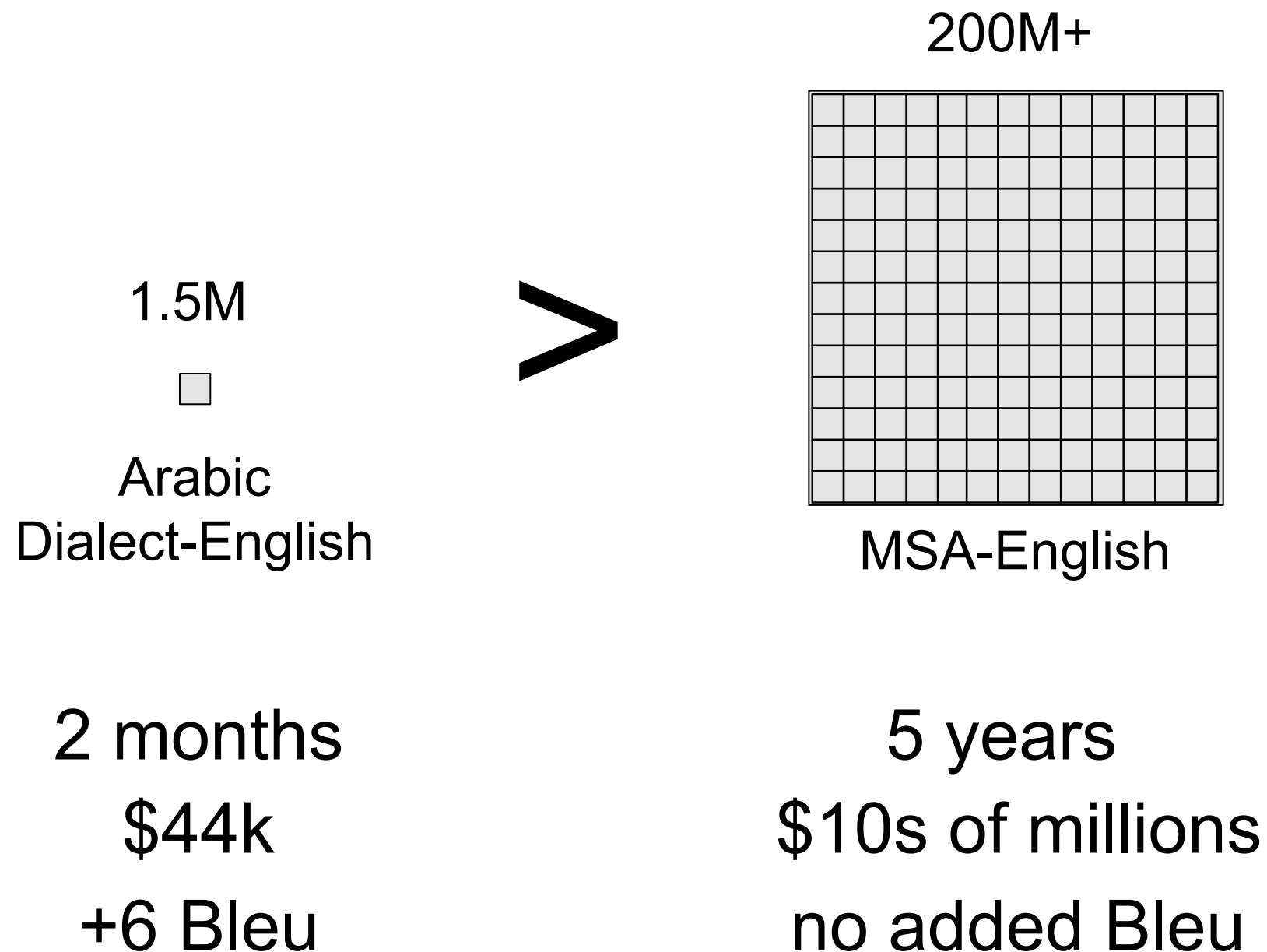
Dialect versus MSA

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



Dialect v. MSA data for MT

For machine translation of Arabic dialects



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

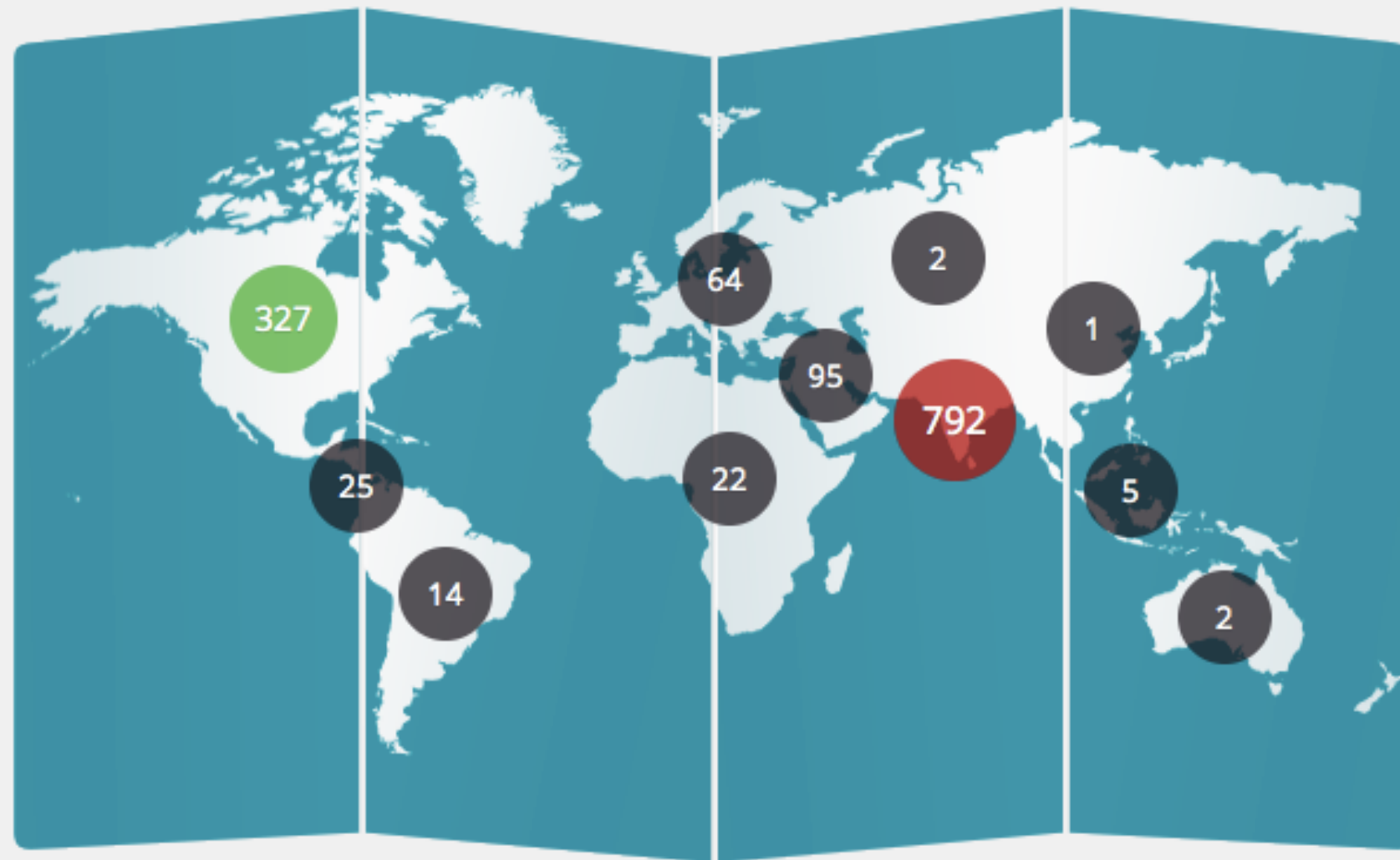
- Microsoft translator does 35 languages
- Google does 57 languages
- The DoD's Center for Applied MT does 64
- There is not enough data to reach acceptably high quality in new languages
- I want us to do 100 languages



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 corpus
 - ✓ Bilingual dictionaries for ~100 languages
- ✓ English translation of 180 hrs of spoken Spanish
 - Corpus annotating and correcting ESL errors
 - Speech collection and transcription for new lang.
 - De-romanization / text normalization corpus

Translator activity



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 >

Thanks!

Chris Callison-Burch

ccb@cis.upenn.edu

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>