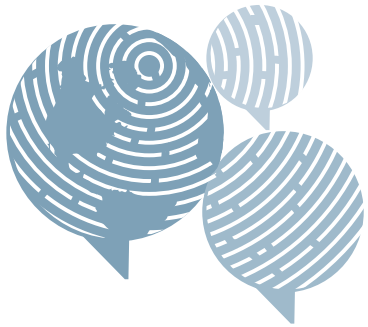
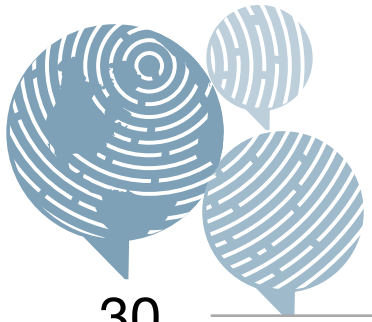


Creative Data Collection: Crowdsourcing Translation

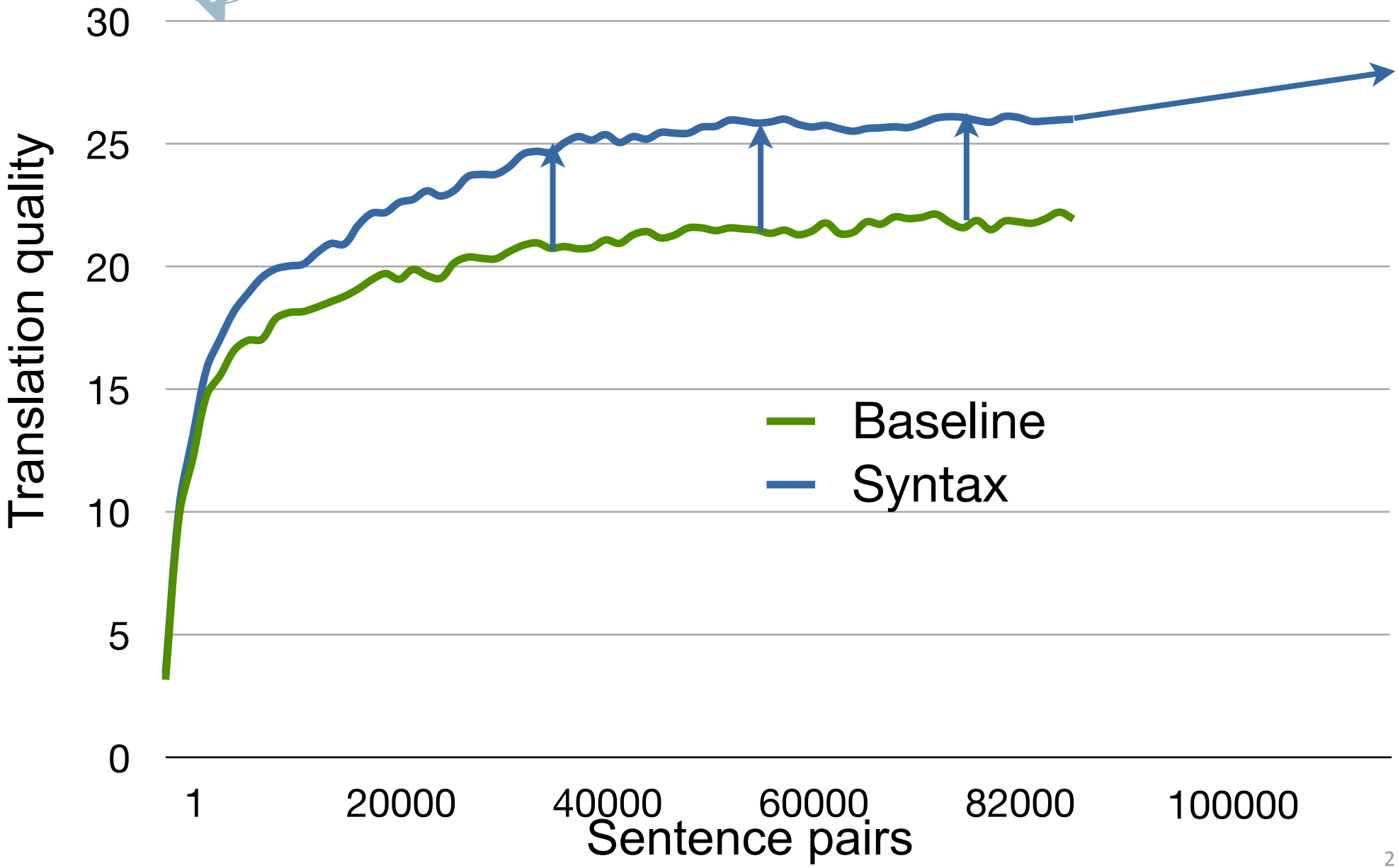
April 3, 2012

human language technology
center of excellence

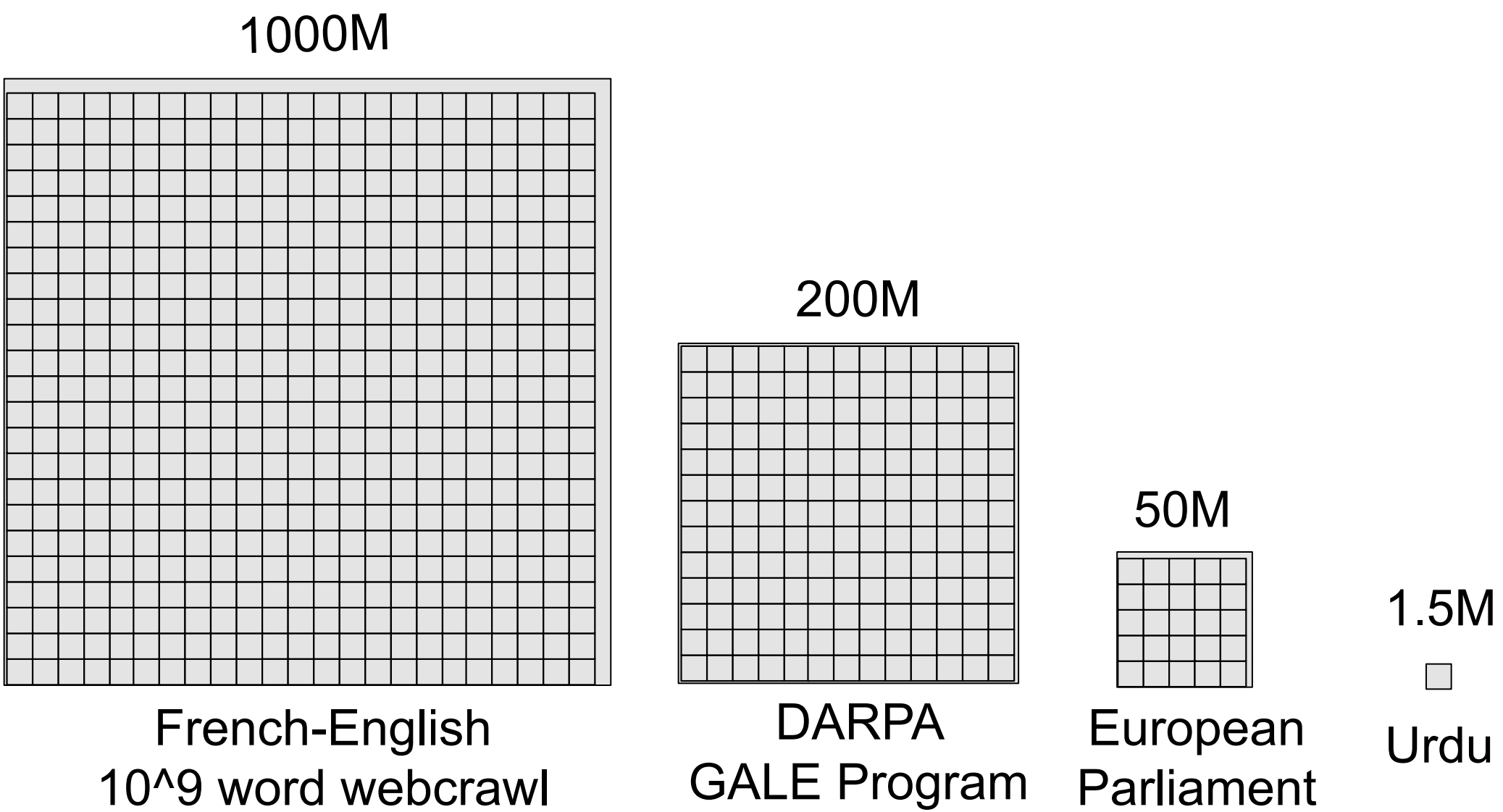
JOHNS HOPKINS
UNIVERSITY

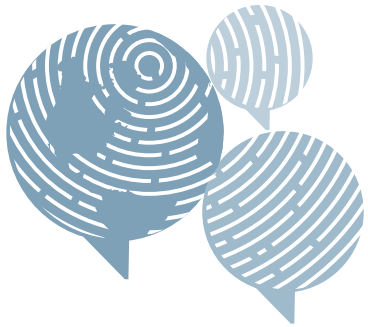


More data is better



Few languages have sufficient amounts of data





Goals for this lecture

- Introduce you to crowdsourcing via Amazon's Mechanical Turk
- Can we create training data for SMT?
  - How good are non-professional translators?
  - How much would it cost to create a parallel corpus?
  - Can we find speakers for low resource languages?
  - Can we train an SMT system from this data?
- What would low cost, high quality translations enable us to do?

amazonmechanical turk
Artificial Artificial Intelligence

Your Account | HITs | Qualifications

Already have an account?
Sign in as a Worker | Requester

Introduction | **Dashboard** | **Status** | **Account Settings**

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. View them now.

Make Money
by working on HITs

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

As a Mechanical Turk Worker you:

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

Find HITs Now

or learn more about being a **Worker**

Get Results
from Mechanical Turk Workers

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

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

Get Started

or learn more about being a **Requester**

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) (month & day)

Postmark Year: (Ex: 1909)

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

Postmark City: Barre

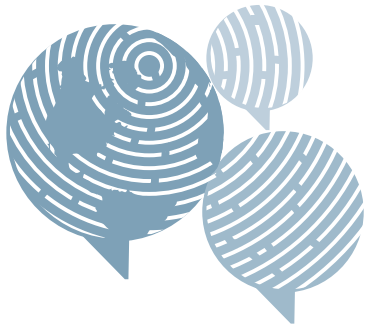
Postmark State: MA

Postmark Date: Oct-11

Postmark Year: 1886

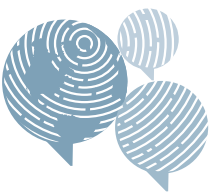
Stamp: 1c

$ 0.01



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



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 Shared Task

NAACL HLT 2010 LOS ANGELES

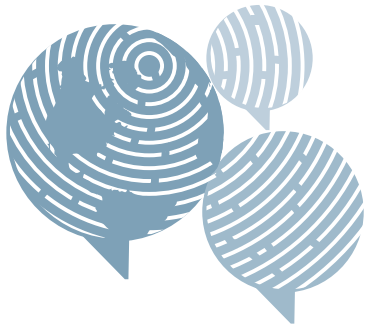
- Traditional NLP tasks
  - WSD, RTE, NLG, common sense knowledge
- Speech and Vision
  - Transcribed speech, accented speech, handwriting OCR
- Sentiment, Polarity, Bias
  - Cross language, blogs
- Information Retrieval
  - TREC style annotations
- Information Extraction
  - Relation extraction, NER
- Machine Translation
  - Paraphrases, alignments, training and eval sets, rule cleaning

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



Translation on Mechanical Turk

- Can Turkers do tasks as complex as translation?
- How do we design the task to thwart cheating?
- What mechanisms can we use to ensure high quality translations?

Translate Urdu into English

Help us
Commo
and Eng
Google

- D
- H
- D
- Is

First, plc

Is Urdu
How m
Is Engli
How m

```
<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
man languages. This cou
searching, or new user

study.

ating at any time withou
or closing your browser

have adequate knowledg
tractions, or your answe

on kept about you will b
This information may be

to the principle
ou have been treated
niversity Institutional

s 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 ہیں۔

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۔

Machine Translate

A1B6VQGOT2G4TL			Lahore	Pakistan		two of the other respectable career .	Two were connected with misscellenous honorable professions.	march 1782_OOV amin_OOV british cabinet accepte...
A1B6VQGOT2G4TL	60	60	Lahore	Pakistan		florida initial european nawabadion_OOV center of which in pansacola_OOV , fort carolaven_OOV and senate agstien_OOV .	Florida became the center of the initial European settlements,which included Pensacola,Fort Caroline and Saint Augustine.	but clear goal in front of revolutionary leaders .
A1B6VQGOT2G4TL	60	60	Lahore	Pakistan		you people of business for the trade with the detainees .	The prophet would also carry the merchantdise of other people for purposes of trading	hazrat talib said that if syria assuaged or chr...
A1B6VQGOT2G4TL	60	60	Woodbury	United States		the increase in employment and construction on a large scale .	The rate of employment increased rapidly and construction work commenced on a huge scale.	isi_OOV near tokyo , yokosuka_OOV area of w...
A1WYSSW33M2FZ2	27	20	Riyadh	Saudi Arabia		palestinian muslim ._OOV	Palestinian Muslims,	britain has the second world war from the works...
A1WYSSW33M2FZ2	27	20	Riyadh	Saudi Arabia		which palestinian muslims are dying , but israil_OOV especially fayed_OOV not . _OOV	From which Palestinian Muslims continued to die but Israel did not get a special benefit.	moshe_OOV dean_OOV terrorists in spite of the b...
A1WYSSW33M2FZ2	27	20	Riyadh	Saudi Arabia		the german citizenship .	Gained German Nationality once again.	1906a_OOV in zurek_OOV univisti_OOV has .

Reject

Worker ID	Years Speaking Hindi	Years Speaking English	City	Country	Input.Machine Translation10	Translation10	Input.Machine Translation1
A356LXGP5Z6D75	16	16	Chennai	India	Amravati Division - Vidarbha's top Hind...		Daily Hindi News - First Antarjaliy lot of Bund...
A6SJI1FPZZT7I	26	20	Taipei	Taiwan	According to area in the world Chauntisven place.	sdfuhgjigvc jashcfiu sdfuhgjigvc jashcfiu	According to the Pakistani government this time...
A3KBG0N9VNAXK3	10	10			Introduction	parichay	Literary texts and the world's longest ...
A6SJI1FPZZT7I	26	20	Taipei	Taiwan	Introduction	it athu but anal what enna say sol sir iya it a...	Literary texts and the world's longest ...
A2IZPICYIBNNXL	7	7	Cochin	India	Delhi's power - change	to India, thus forming the base for the Indo-Persian culture	Mughal state then the spread was only from Kabu...
A6SJI1FPZZT7I	26	20	Taipei	Taiwan	Devanagari intensive Clerk - 100% pure typing t...	sdfuhgjigvc jashcfiu sdfuhgjigvc jashcfiu sdfuh...	Hindiraeatar the Aeme (growth stop)
A6SJI1FPZZT7I	26	20	Hong Kong	Hong Kong (SAR)	So it called India's financial capital.	sdfhug jsbjg ajksbhjig mkasbnjigh amksfbjigmn j...	It produced lava formed seven small - small isl...
AK32OHKIJHRLC			Hyderabad	India	So it called India's financial capital.		It produced lava formed seven small - small isl...
			Hong	Hong	Retikaalen poets of the	sdfhug jsbjg	But according to

A2IZPICYIBNNXL	8	8	Cochin	India	Was the last Mughal King Bahadur Shah Zafar.	Urdu publications are the reason of this language staying alive	It is believed that today's modern Delh...	
A2IZPICYIBNNXL			Bangalore	India	State legislature	The war had profound economic consequences.	Also the governor is a constitutional head is h...	
A2IZPICYIBNNXL			Bangalore	India	Akĥj Ejan (search engine) to give Ssraĥ Kaĵad ...	The war had profound economic consequences.	Domain name problem	
A2T8QQ8P74APXE	10	15	Madurai	India	Akĥj Ejan (search engine) to give Ssraĥ Kaĵad ...	search engine	Domain name problem	
A6SJI1FPZZT7I	26	20	Taipei	Taiwan	India [T? Involved] word means or Avideka inner...	HASGDUG JAGFDGACJB AUGFGC JGADUG BAUGF HASGDUG ...	India is the world's tenth largest econ...	
A6SJI1FPZZT7I	26	20	Hong Kong	Hong Kong (SAR)	2. Supreme Court Nirhnyanusar ST 324 [1] in the...	IT ATHU BUT ANAL WAT ENNE SAY SOL SIR IYA ASFJI...	Election Commission's procedure / function	
A6SJI1FPZZT7I	26	20	Taipei	Taiwan	Category: Internet	To change which calendars you receive daily age...	Avihaywar top ten lists of blog sites	
ADOXDQPX17PHN	27	27	Mumbai	India	Divine Miracle (Maoajja)		For purposes of orders in the Qur'an ha...	
A2IZPICYIBNNXL	7	7	Cochin	India	Padoz are rows of other places in the garden an...	in other places in gardens one find trees, its not so over here	Is inscribed on the tomb of Shah Jahan;	
				Hong	Dungarpur and Udaipur	sdfbug ishig	Similarly,	

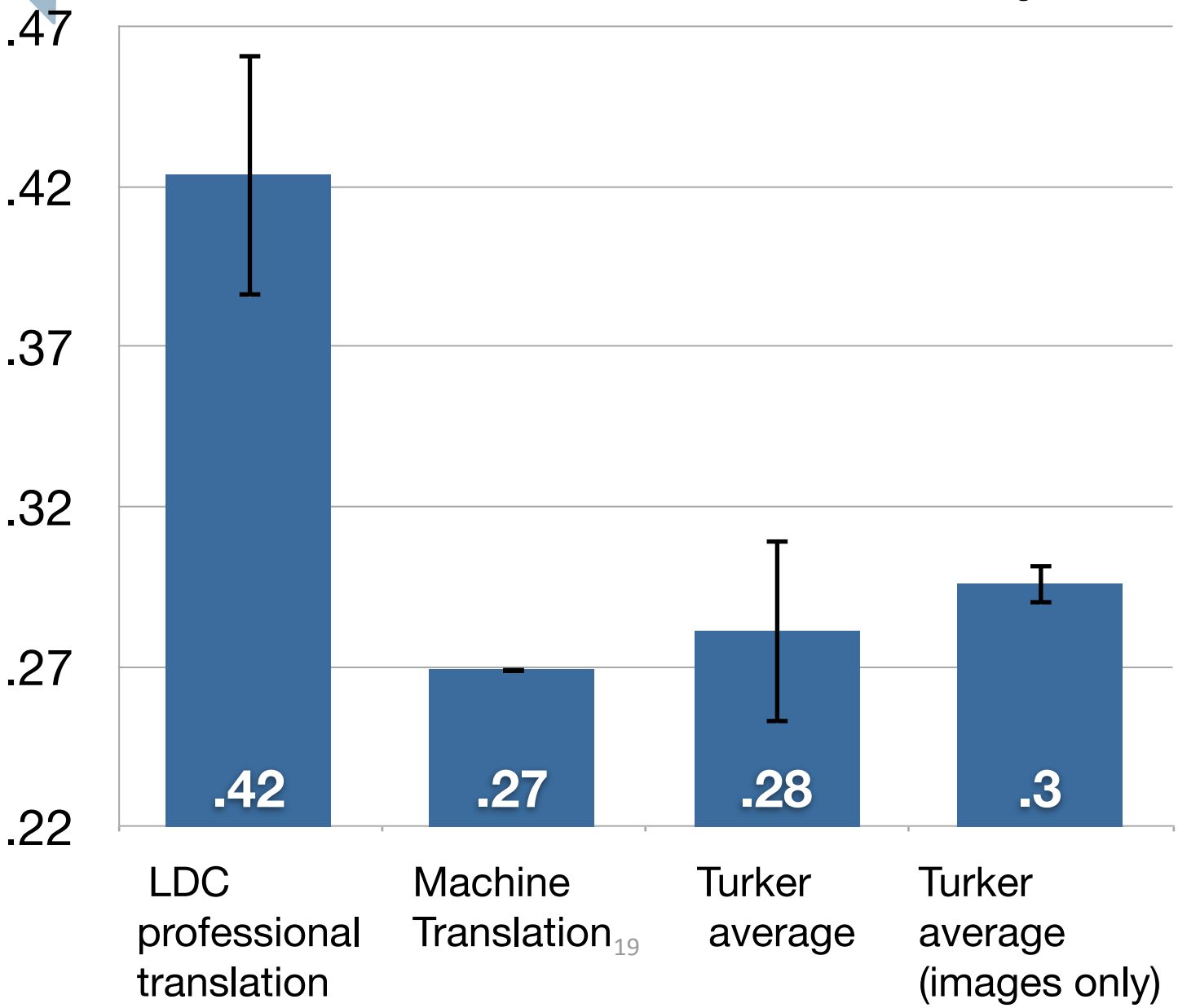
Professionals v. Turkers

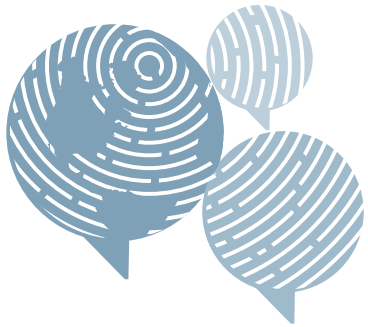
Urdu	LDC Translation	Turk Translation
1994 میں اس خطے میں ابتدائی انسانوں کی باقیات جو تقریباً 8 لاکھ سال پرانی مانی جاتی ہے ، دریافت کی گئیں جنہیں ہومو اینٹی سیسر یعنی 'بانی انسان' کا نام دیا گیا۔ اس سے قبل 6 لاکھ پرانے انسان جنہیں سائنسی اصطلاح میں ہومو ہیڈلبرجینسس کہا جاتا ہے ، اس خطے کے قدیم ترین	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.

Measuring Translation Quality

- We'd like to quantify how good (or bad) the translations are
- Automatic metrics (like Bleu) compare a translation against a set of reference translations
- Generally used to measure machine translation quality but we apply them to humans
- Measure one professional translator's translations against three other professionals
  - Gives us a range of what to expect in good translations

Bleu Scores (professional v Turk w/o QC)





Improving quality through redundancy

- We have collected multiple translations for each source segment
  - 4 translations, a total of 10 edits
- Selection strategies
  - Pick best on a sentence-by-sentence basis
  - try to pick the best Turker
- Goal: use machine learning to predict translation goodness

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.

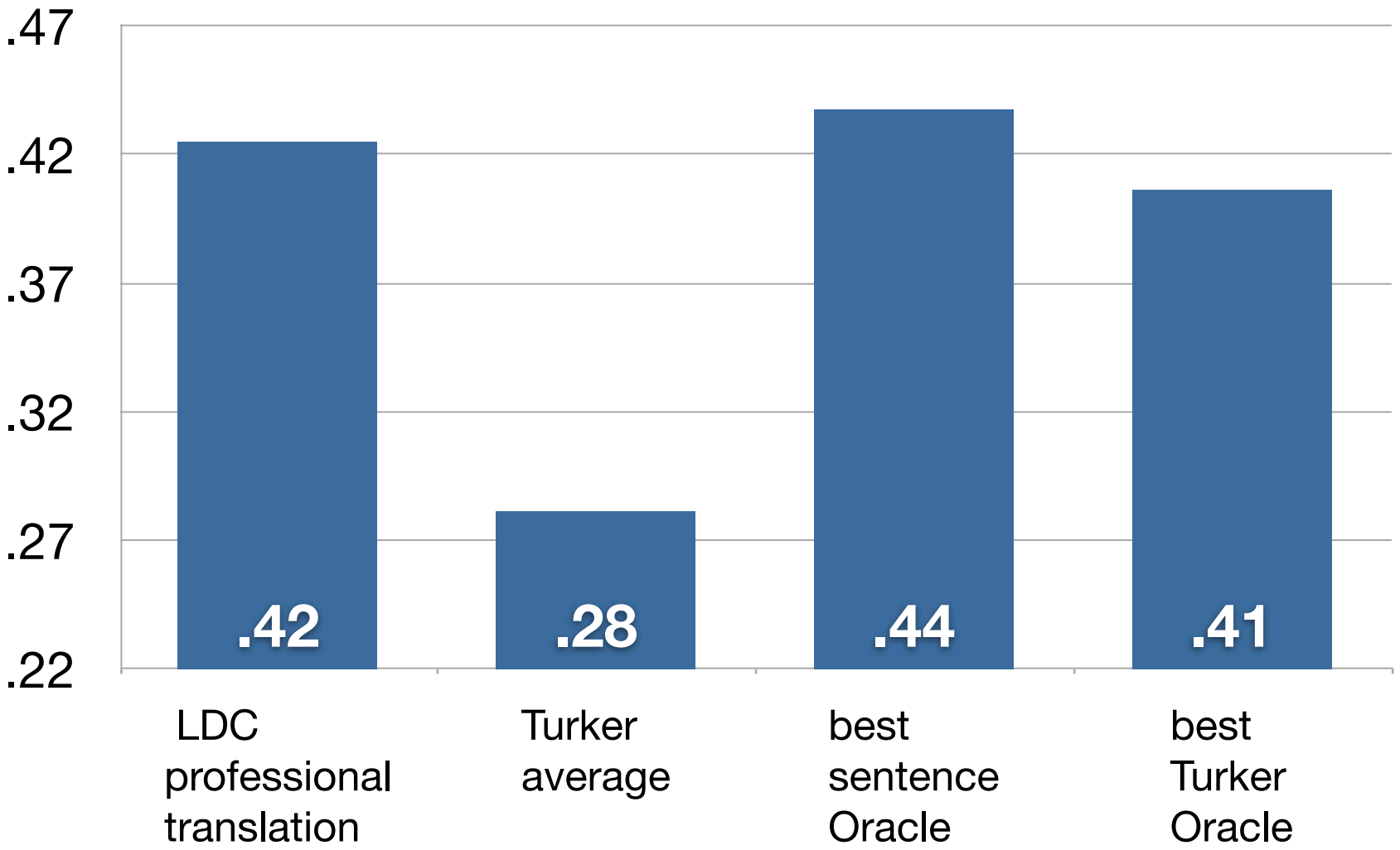
Avoiding dieting to prevent from flu	abstention from dieting in order to avoid Flu	Abstain from decrease eating in order to escape from flue	In order to be s from flu quit dietin

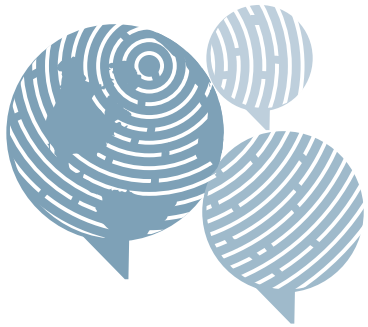
Avoid		dieting		to		prevent	from	flu
Abstention	from	dieting	in order	to		avoid		Flu
Abstain		decrease eating	in order	to		escape		flue
quit		dieting		to	be	safer	from	flu

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 … 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

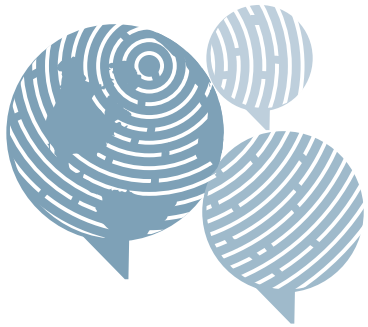
Bleu Scores (ORACLE selection)





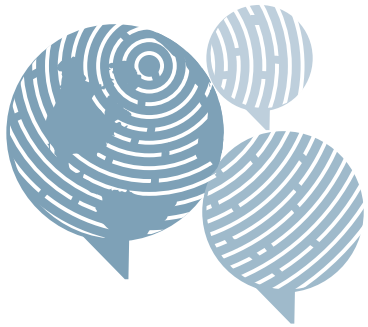
Quality Control Model

- Model assigns a score to translations based on weighted features
- 4 groups of features
  - Sentence features
  - Turker features
  - Ranking featured (based on second pass vote)
  - Calibration feature (Bleu against professionals)
- Weights are set on a dev set of professional translations



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



Quality Control Model

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

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:

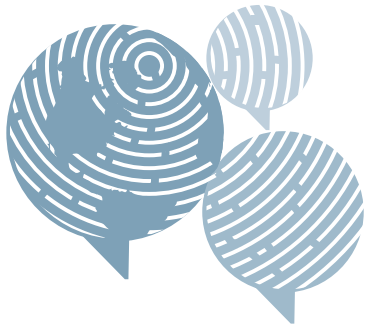
- Does it make more sense than the others?
- Is the English reasonably good?
- Do the grammar and spelling require only minimal correction?

○	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.

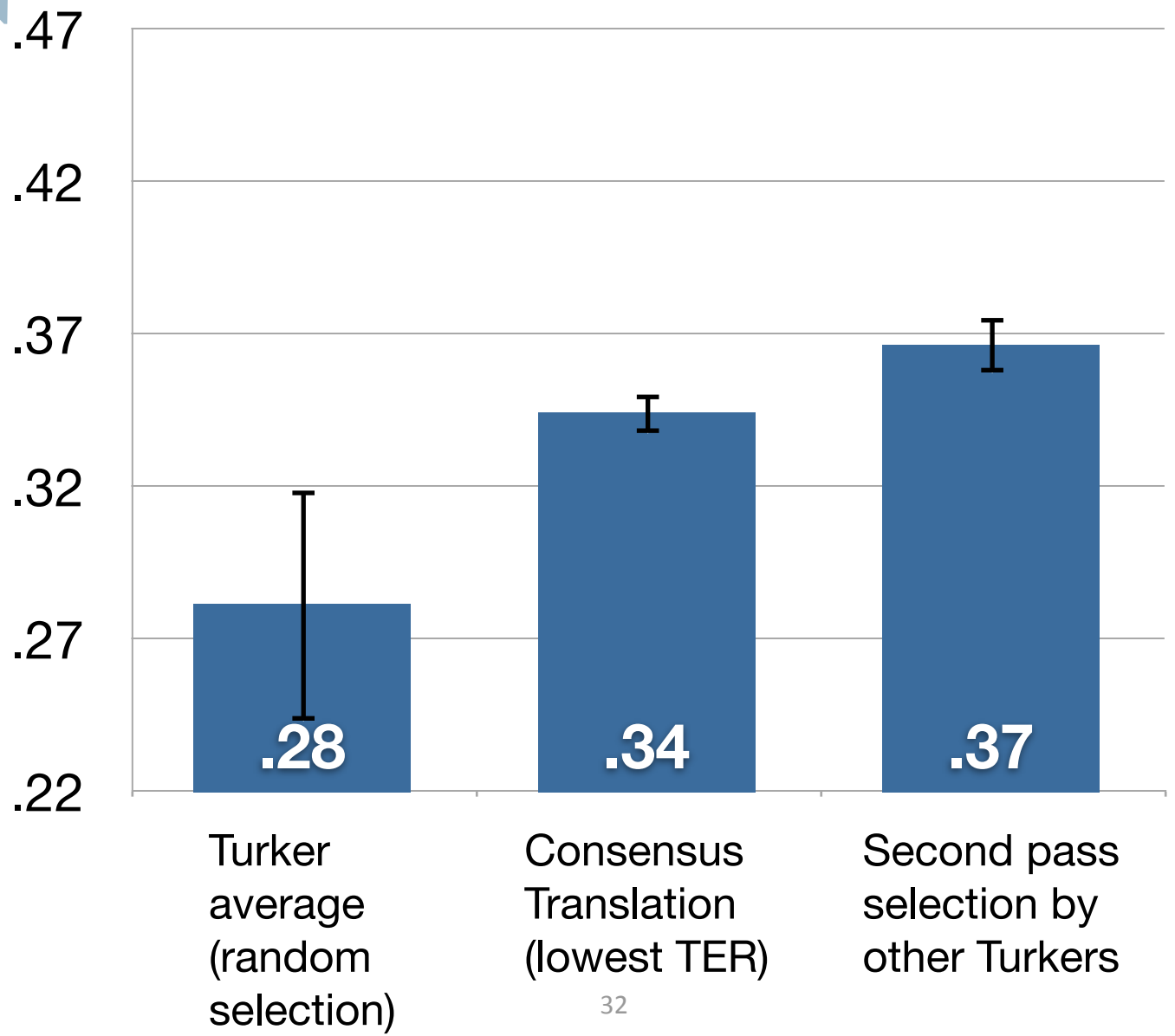
○	According to the Researchers of the State of Michigan University diet taking rats have less strength to fight with the infection and also the chances of death increased whereas earlier those rats on common diet had more strength to fight with the flu virus.

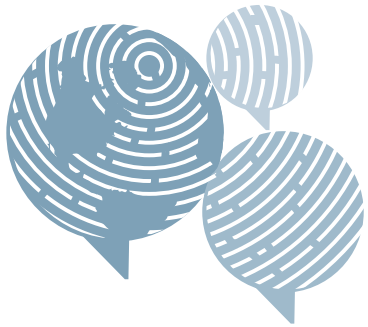


Quality Control Model

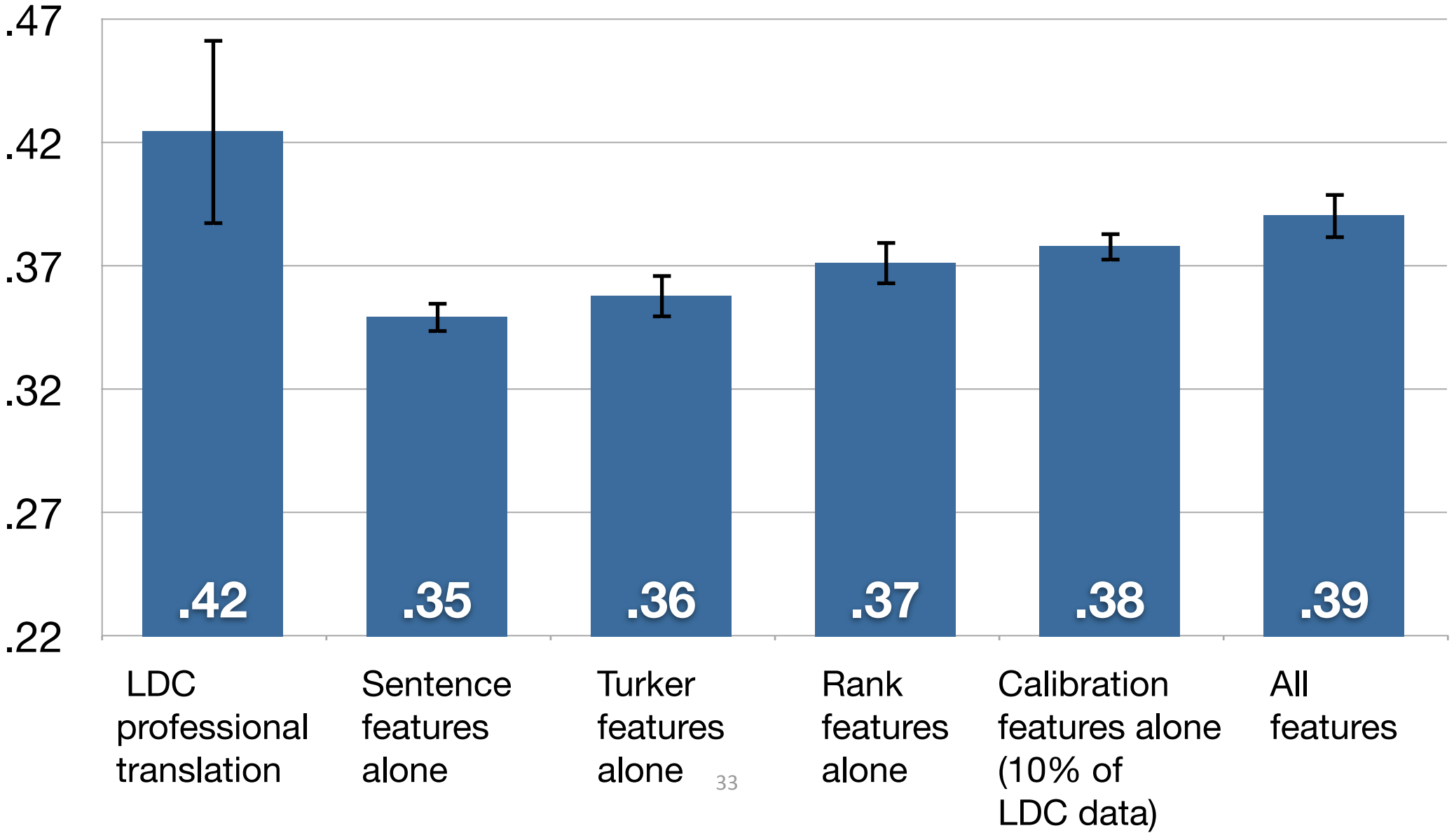
- Ranking features
  - Average rank: averaged over the five rank labels for this translation
  - Is-Best percentage
  - Is-Better percentage

Baseline selection method

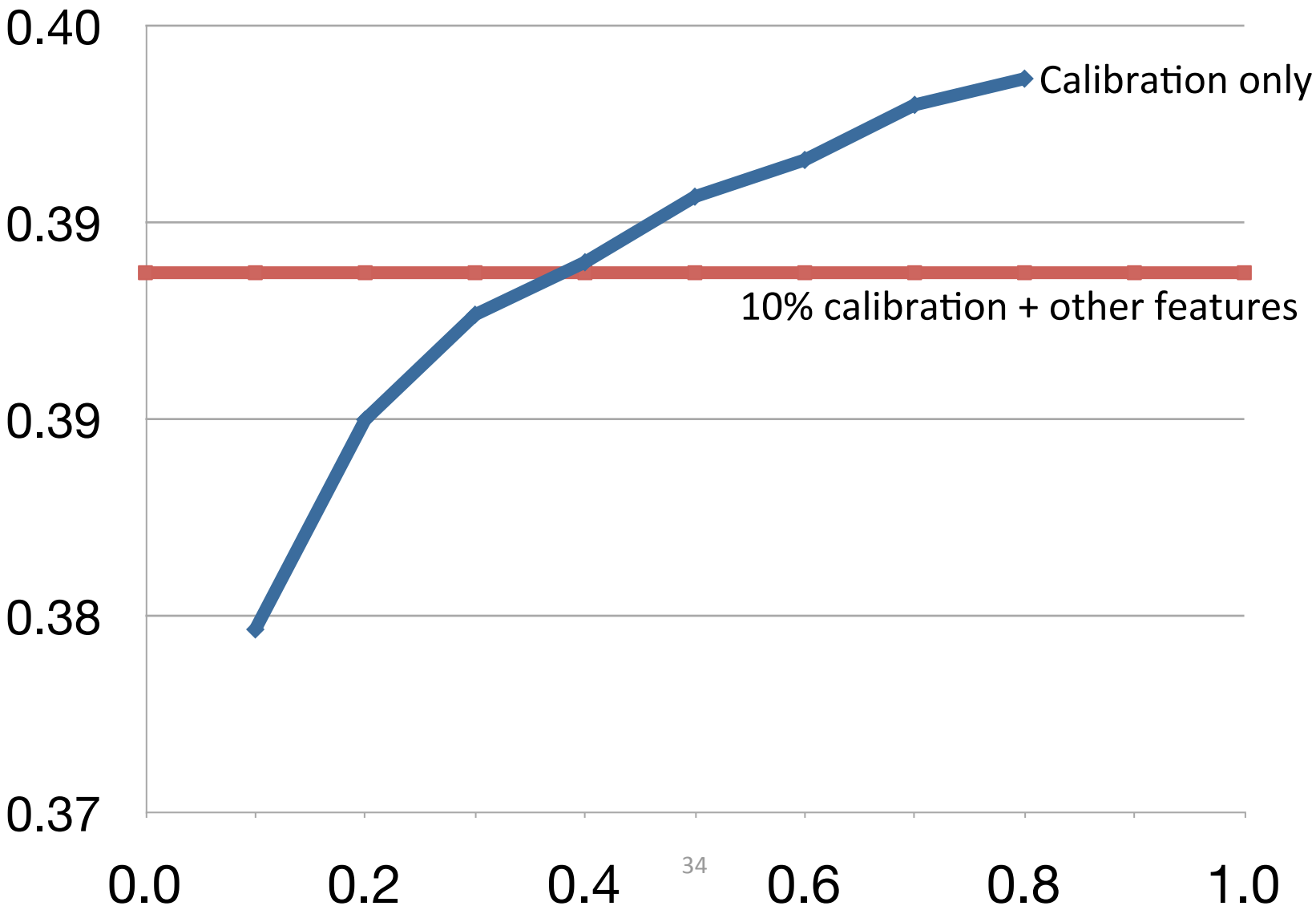


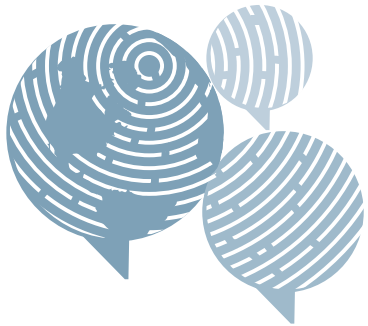


Quality control model



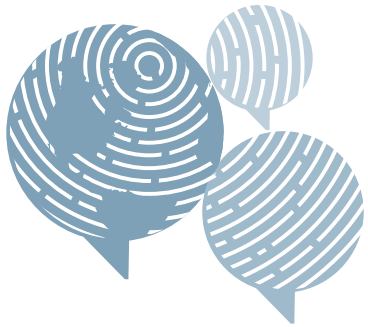
% of LDC data used for calibration



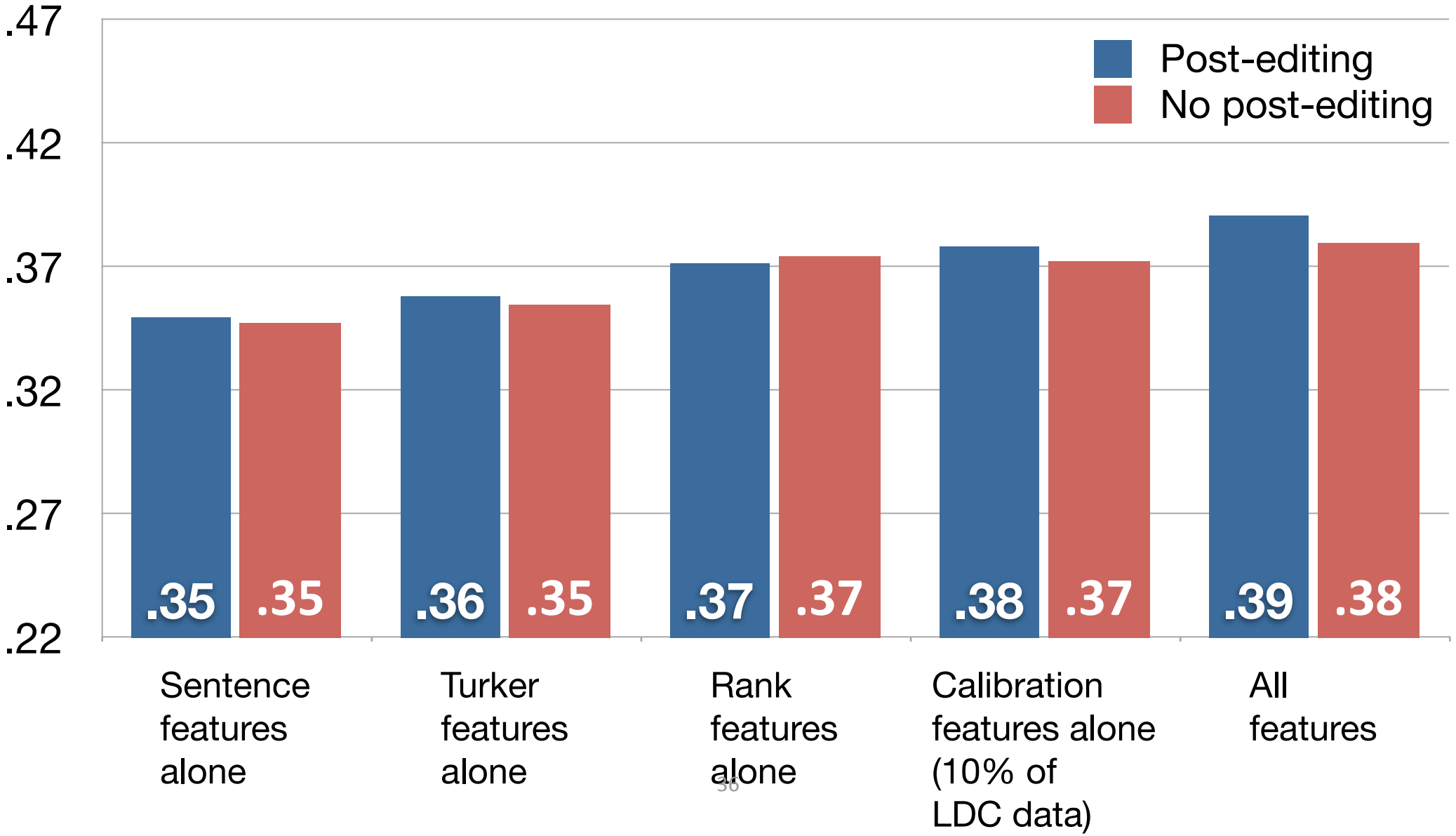


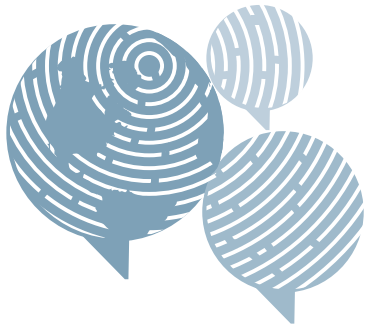
Post-Editing

- The translations were produced by non-native English speakers
- Pretty obvious that they have spelling and grammar errors
  - US-based Turkers to the rescue!
- Each translations was post-edited by 3 other Turkers
- How much of a difference did it make?



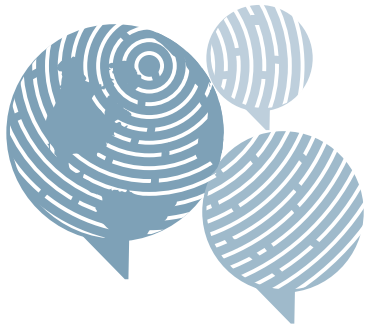
Post editing





Why didn't editing help more?

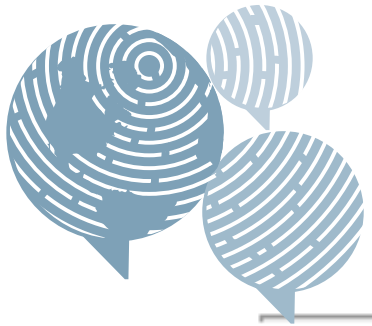
- Post-editing improves results by ~1 Bleu point
- Because there are many easy-to-correct grammatical, spelling errors, we should get more
- One problem is the design of the editing HIT
  - Free form text box
- We have no good mechanism for quality control
  - Turkers tend to be lazy and make few/no corrects
  - So taking a consensus of the edits doesn't work



COMING SOON

A better editing HIT

- Include controls in the editing HIT to distinguish good/bad editors
- 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



ESL HIT

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

several commissions where created to protect the remaining forest region

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

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

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 .

COMING SOON

ESL HIT

Sri Lanka 's forest region was destroyed by agriculture , forestry , animal grazing , etc . ,

wooden works: forestry

vetinary feeds: animal grazing

Word choice/ Awk phrases

Several commissions were created to protect the remaining forest region

several: Several

where: were

Spelling and capitalization

Sri Lanka is considered to be a bird 's sanctionary x .

as: to be

the: a

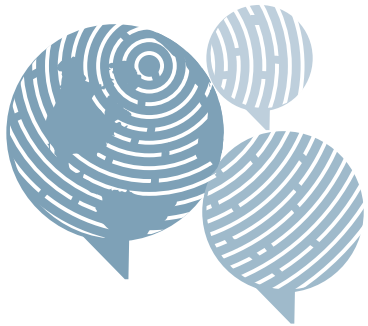
place:

Wrong/missing determiners

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

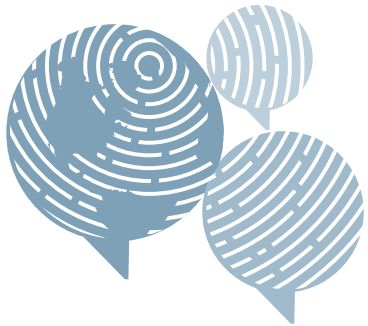
Indian Subcontinent: the Indian subcontinent

There are thousands of animals living in Sri Lanka which includes several Sri Lanka native



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

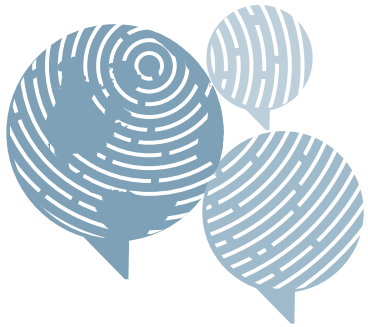
- Single-pass Mechanical Turk translation is 60 times less expensive than professional translation
- I pay ~$1 per 10 sentences = $0.005 / per word
- Additional costs:
  - 4x redundancy on translation
  - Second-pass selection
  - Editing by native English speakers
- Mechanical Turk is still an order of magnitude cheaper than professionals

Cost

- Paid $1 to translate 10 sentences, $0.25 to edit them, and $0.15 to rank them

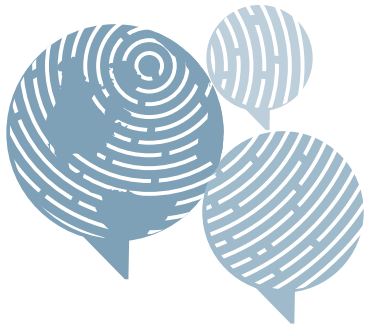
	Done once	w/ redundancy
Translation	$179.20	$716.80
Editing	$44.75	$447.50
Ranking	$26.88	$134.40

- With Amazon's 10% fee our total <$1,500 for 7k translations, 17k edited sentences, 35k ranks
- For 10% calibration data, cost increases by $1k



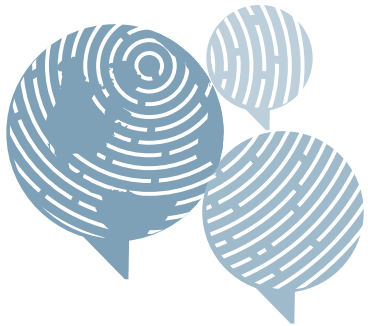
Cost Reduction

- Current pipeline is dominated by cost of professional translation used for calibration
  - Can we reduce this by using a small % of that data?
  - Can we further reduce by using 1 ref instead of 4?
  - Can we get away with no calibration at all?
  - Does calibration data have to come from the current test set? Or can we just have one calibration set per language for all time?



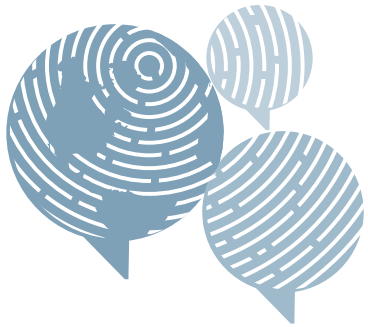
Cost Reduction

- In terms of the crowdsourced components, the most costly part is redundancy in translation
  - Do we need redundant translations of every sentence?
  - Can we tell whether one translation is "good enough"?
  - I.e. can we predict if we solicit another translation whether it will be better than the ones we have?
  - Can we identify good Turkers and incentivize them to do more, so that we can reduce redundancy?



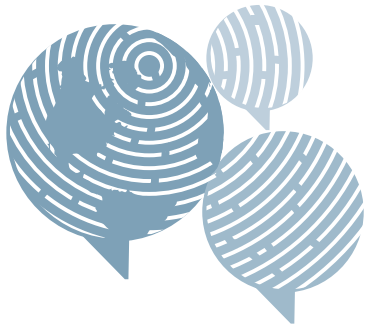
Cost Variance

- Some language pairs may be more costly than others
  - Some countries have higher costs of living than others
  - Some languages are not as well represented on Mechanical Turk (classic supply and demand)
- Speed / quality may be affected by payment



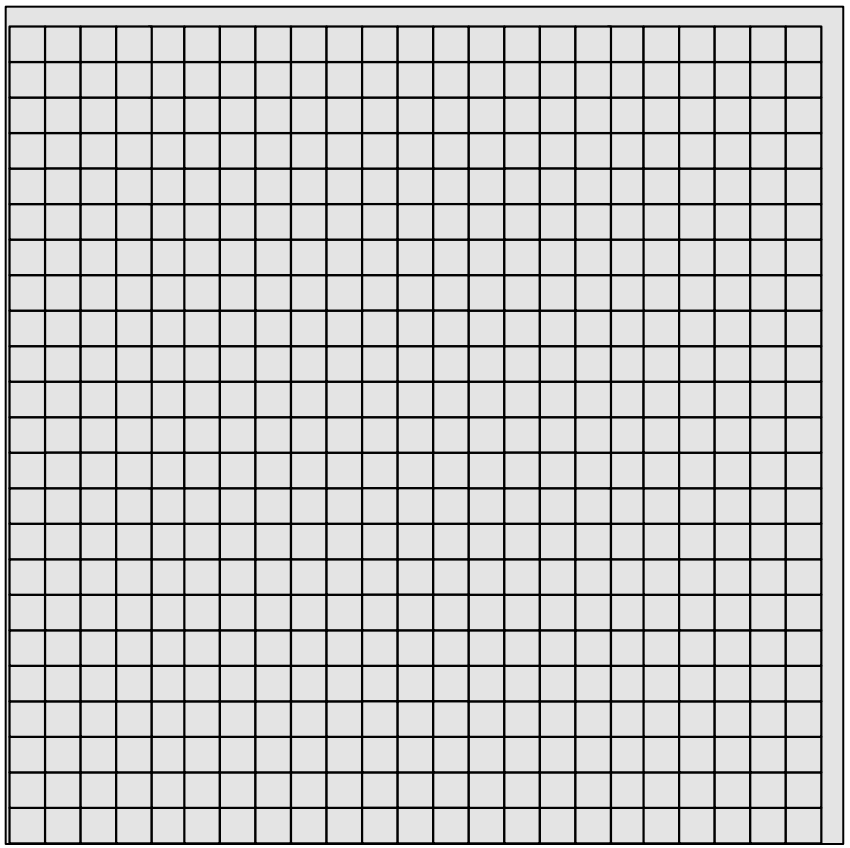
Less Commonly Taught / Low Resource Languages

- I would like to use this as a resource for creating data for low resource languages
- Low resource languages are ones where we do not have much parallel data
- Therefore translating statistical machine translation models is problematic



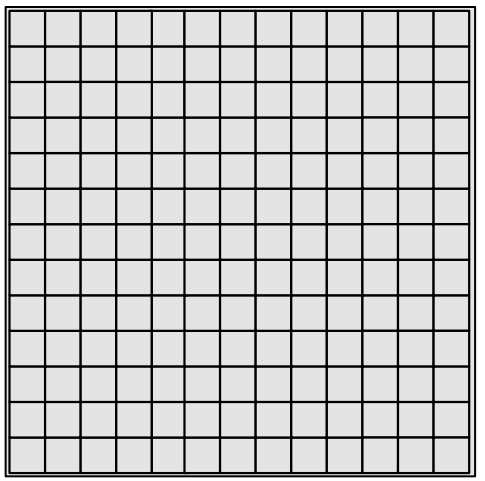
Volume of training data

1000M



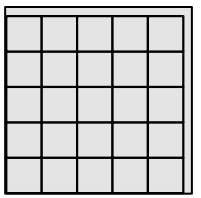
French-English
10^9 word webcrawl

200M



DARPA
GALE Program

50M



European
Parliament

1.5M

Urdu

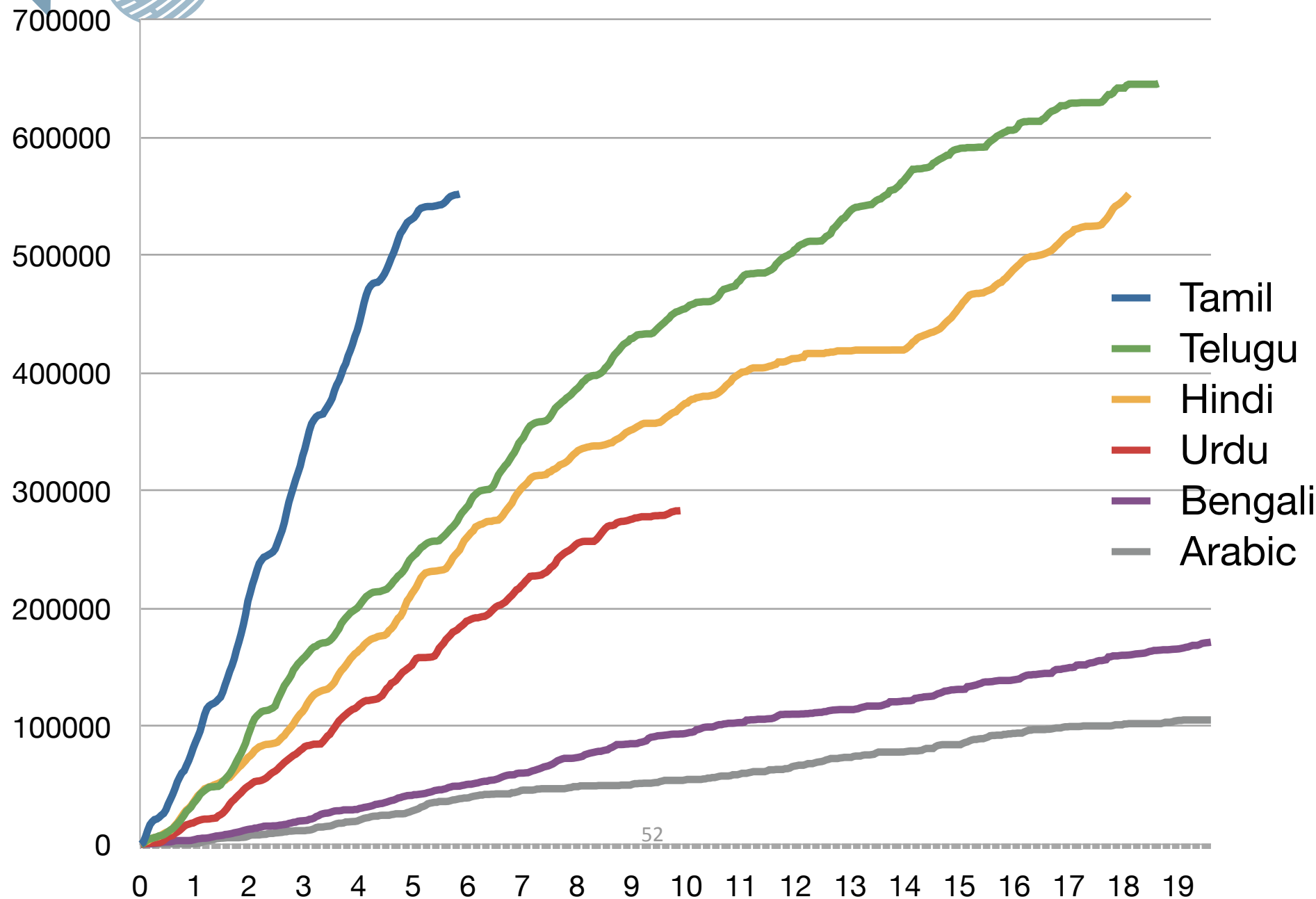
Feasibility study

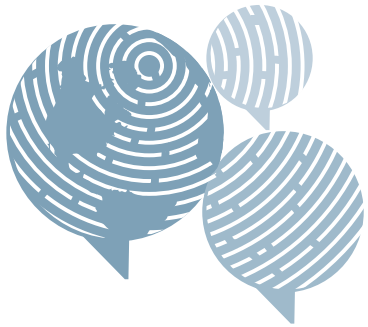
Feasibility study

Feasibility study

Basque
Armenian
Kurdish
Uzbek
Igbo

Volume Study

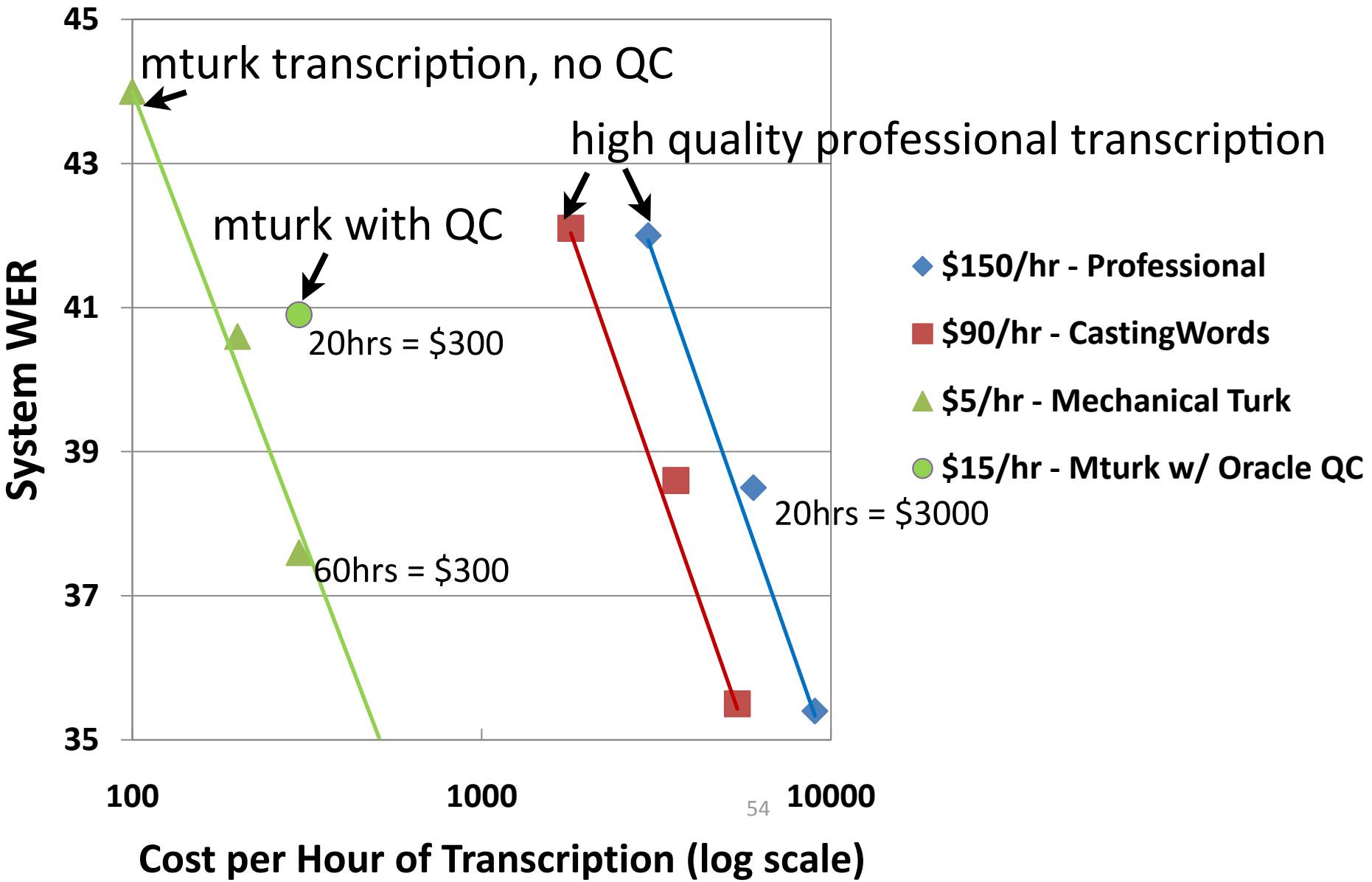




Quality v quantity when training statistical models

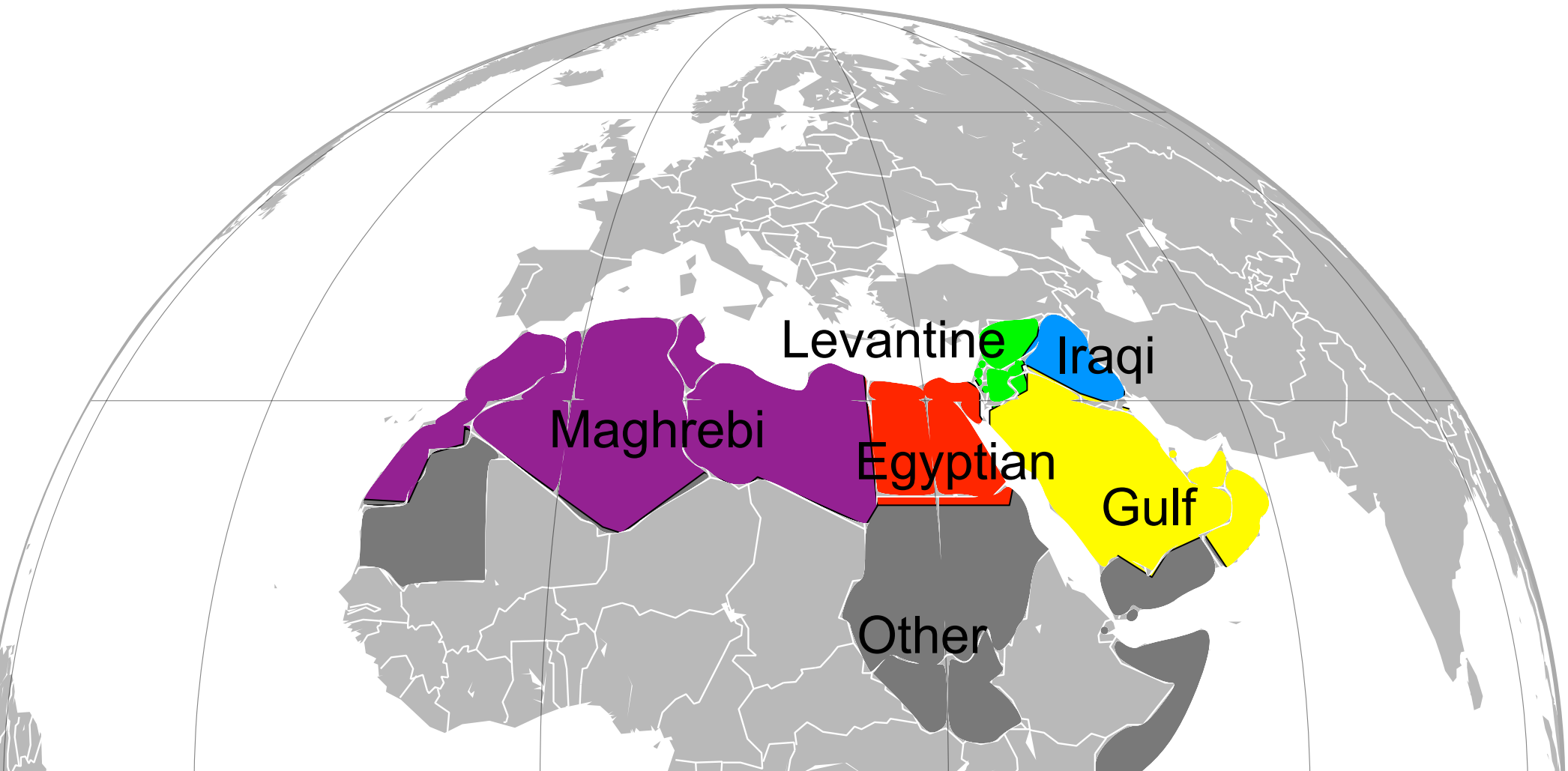
- Investing more money results in higher quality translation, but we get fewer items
- When training a statistical model, how much should we spend to ensure high quality translations?
- How robust to noise are our models?
- ASR experiments suggested not worth worrying too much about quality

Quality v quantity when training statistical models



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 ؟

AymtY 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**?

Arabic Dialect in Online Comments

Reader comments

بلخادم: ما يجمع مصر والجزائر يجعل التقارب أكثر من ضروري

الخميس 16 يونيو 2011 - 09:27

عبد العزيز بلخادم الممثل الشخصي للرئيس الجزائري

Article body (MSA)

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 فريقي البلدين في 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 الحكومي والجماهيري والمؤسساتي

تعليقات (55)

1 **خليك فى حالك وابعد عننا**

بواسطة: bebooo

بتاريخ: الخميس، 16 يونيو 2011 - 09:39

EGY

خليك فى حالك انت وابعدوا عنا وانتم 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 الراهنه التي يمر بيها الوطن العربي ، أهلاً ومرحباً بأي تقارب عربي عربي بين الأشقاء في الوطن والمصير ، يسقط الأستعمار الجديد المتمثل في أمريكا والناتو ومن ورائة آل 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 الجزائرى يعمل جاهدا من اجل السعي لتوجيد العالم العربي وكدالك 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 الأنضمة فاسدة و تابعة لأمريكا و إسرائيل، إن أردتم [illegible] العرب في [illegible] فما على الحكام إلا تكوين جمعية و الذهاب لمنتجع الحكام المخلوعين [illegible] وسيبوا [illegible] في قيام الولايات العربية المتحدة و عاصمتها القدس إنشاء الله ([illegible] حلم يقظة [illegible] يوم من الأيام من منا يعلم الغيب؟)

MSA

EGY

Identify The Arabic Dialect

Help us classify Arabic text into dialects! This HIT is for Arabic speakers who understand the different local Arabic dialects (اللهجات العامّية، أو الدّارجة), and can distinguish them from *Fusha* Arabic (الفصحى).

Below, you will see a table that contains several Arabic sentences. For each one:

1. Tell us how much dialect (عامّية) is in the sentence, and then
2. Tell us which Arabic dialect it is.

As you expect, **Egyptian** is اللهجة المصرية, and **Iraqi** is اللهجة العراقية. As for the other dialect names, **Levantine** (شامي) includes **all** dialects of بلاد الشام , **Gulf** (خليجي) includes **all** dialects of الخليج العربي , and **Maghrebi** (مغربي) includes **all** dialects of المغرب العربي.

This following map illustrates the dialect groups:

Informed Consent Form

Purpose of research study: We are collecting human annotations to improv other languages. These annotations might be class labels, judgments of out

Benefits: Although it will not directly benefit you, this study may benefit soc process human languages. This could lead to better translation software, im 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 pena button, or closing your browser window.

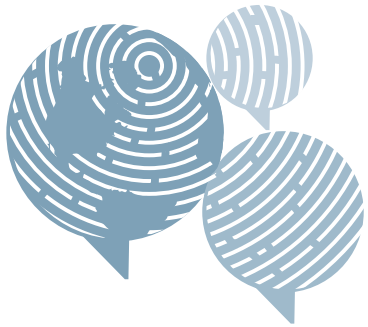
We may end your participation if you do not have adequate knowledge of t following the instructions, or your answers significantly deviate from known

Confidentiality: The only identifying information kept about you will be a Wo address. This information may be disclosed to other researchers.

Questions/concerns: You may e-mail questions to the principle investigator you have been treated unfairly you many contact the Johns Hopkins Univer

Clicking on the "Accept HIT" button indicates that you understand the info have not waived any legal rights you otherwise would have as a participant

Which Dialect? أى لهجة	Dialect Level كمّية اللهجة العامّية	Sentence الجملة
General (ليست لهجة معيّنة) Levantine (شامية) Gulf (خليجية) ✓ Egyptian (مصرية) Iraqi (عراقية) Maghrebi (مغربية) Other (أحد اللهجات الأخرى) I don't know! (لا أعرف) No dialect (فصحى فقط) Not Arabic (لغة أخرى أو رموز) Choose dialect... Choose level first	A bit of dialect (القليل من العامّية)	و لعب فيها فريق تل ابيب جنبا الى جنب مع المنتخبات الاسيويةّ على ارض قطر
	Choose level...	وكلف رئيس مفوضيةّ الاتحاد الافريقي جون بينج والأمين العام للجامعةّ العربيةّ عمرو موسى بإعداد النظام الاساسي لهذا الصندوق وتحديد اهدافه وادارته واوجه صرف هذه المساعدات وشروطها

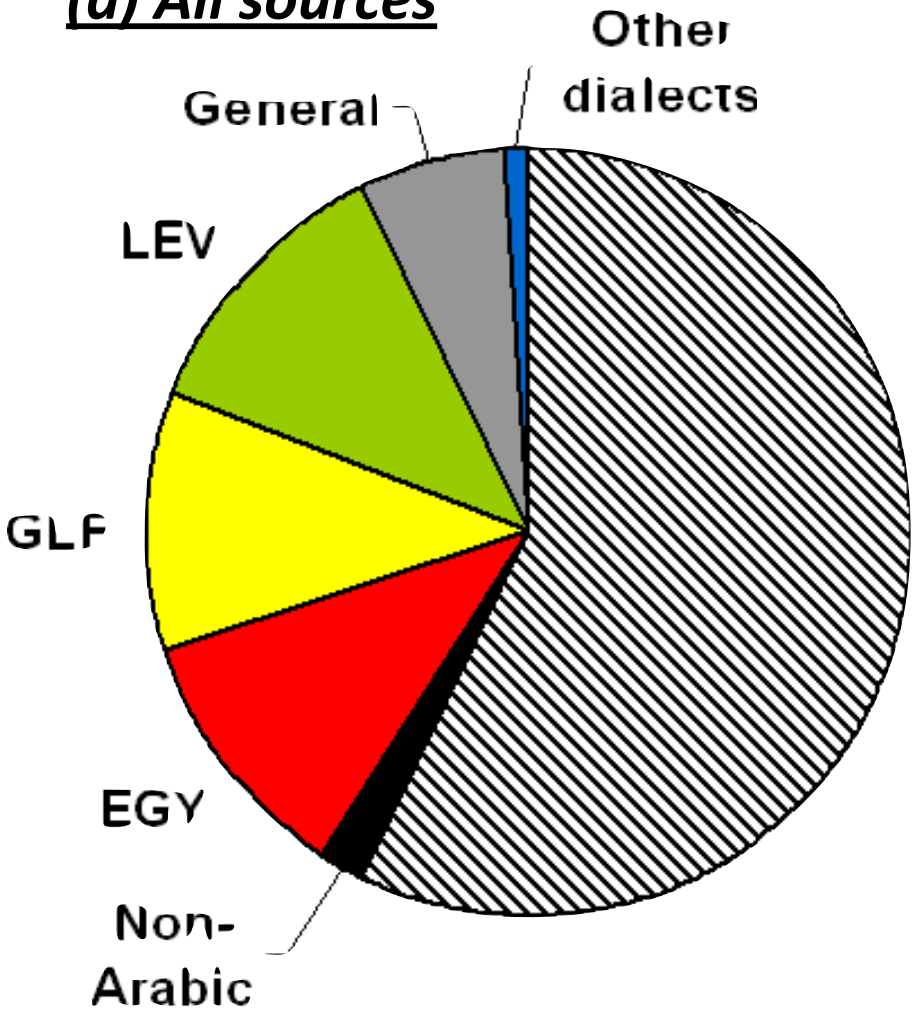


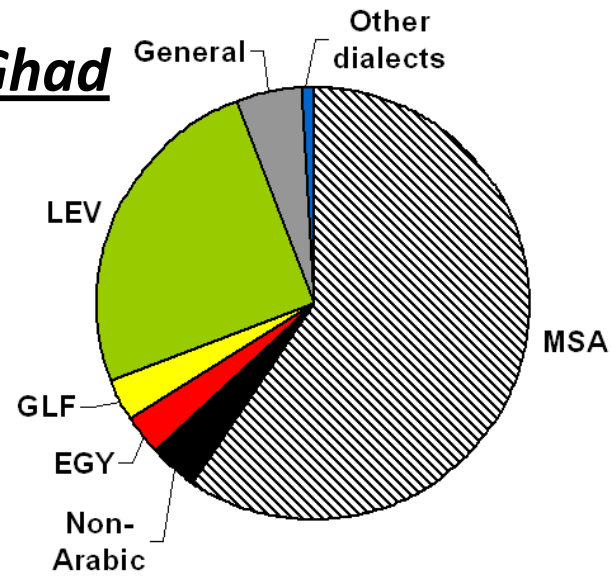
Crowdsourcing Dialect ID

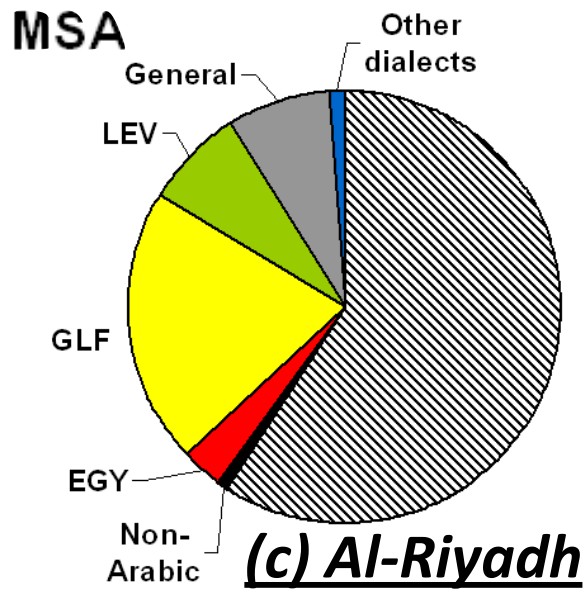
Full details in Zaidan and Callison-Burch (ACL 2011b)

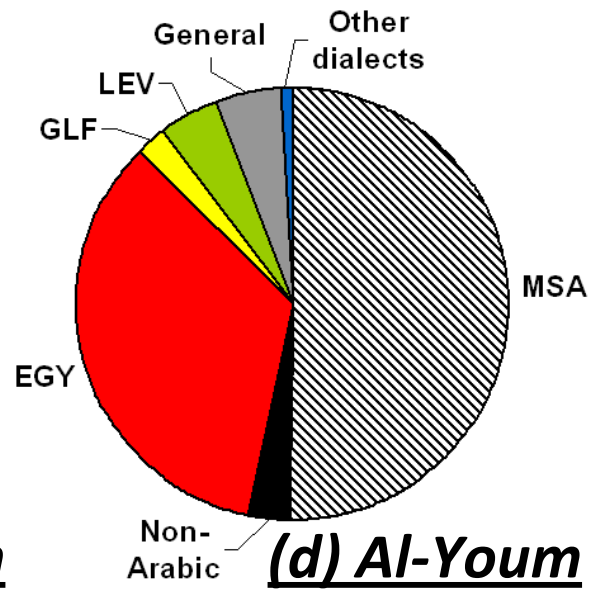
- Labeled 142k comments gathered from three online newspapers:
  - Al-Ghad (الغد), a Jordanian newspaper LEV MSA
  - Al-Riyadh (الرياض), a Saudi newspaper GLF MSA
  - Al-Youm Al-Sabe' (اليوم السابع), an Egyptian newspaper EGY MSA
- 59% MSA, 41% dialect
- Trained classifiers with 80-90% accuracy

Label Distribution

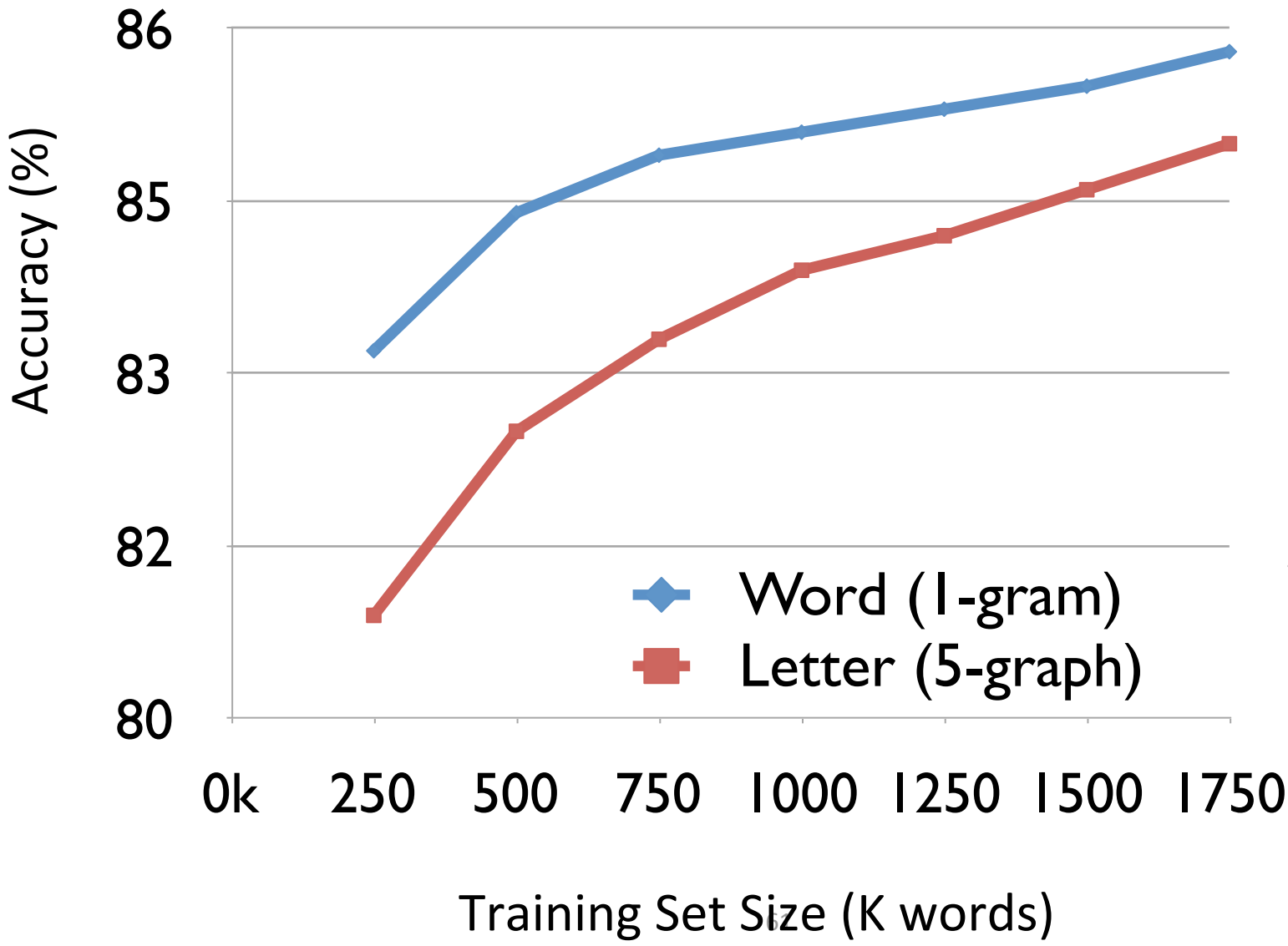








Automatic dialect ID

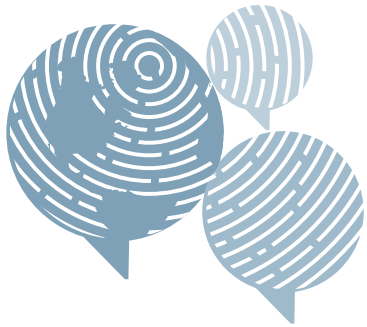


Baselines

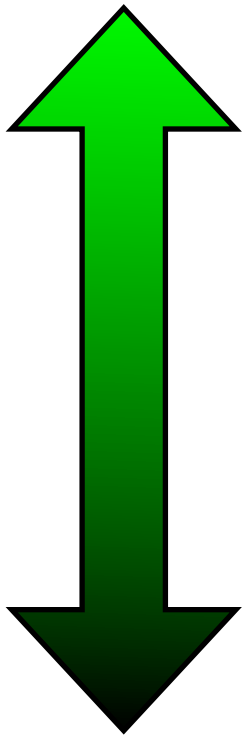
Majority class: **59%**

OOV wrt Arabic GigaWord: **66%**

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



Extremely dialectal Levantine words

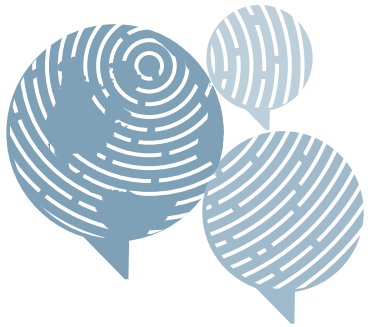


Extremely MSA words (i.e. very unlikely in dialectal context)

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
Alljnp	اللجنة	the-committee	0.095
f<n	فإن	(declarative)	0.062
AlmfAwDAt	المفاوضات	the-negotiations	0.038
AlmbA$rp	المباشرة	the-direct	0.029

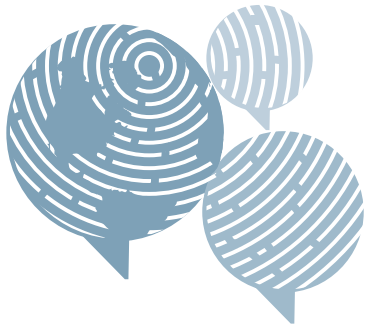
$$DF(w) \stackrel{\text{def}}{=} \frac{f(w|D)}{f(w|MSA)}$$

"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

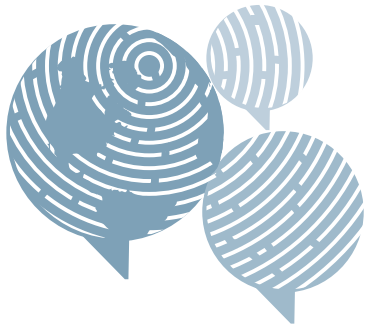


Crowdsourcing Translation

- Translated dialect-labeled segments EGY LEV

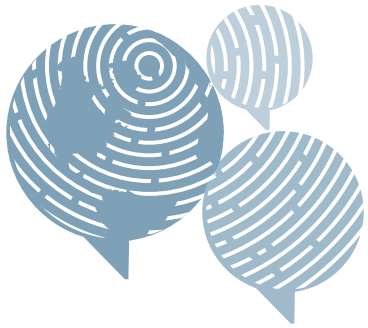
Dialect Classification HIT	$10,064
Sentence Segmentation HIT	$1,940
Translation HIT	$32,061
Total Cost	$44,065
Num words translated	1,516,856
Cost per word	2.9 cents/word

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



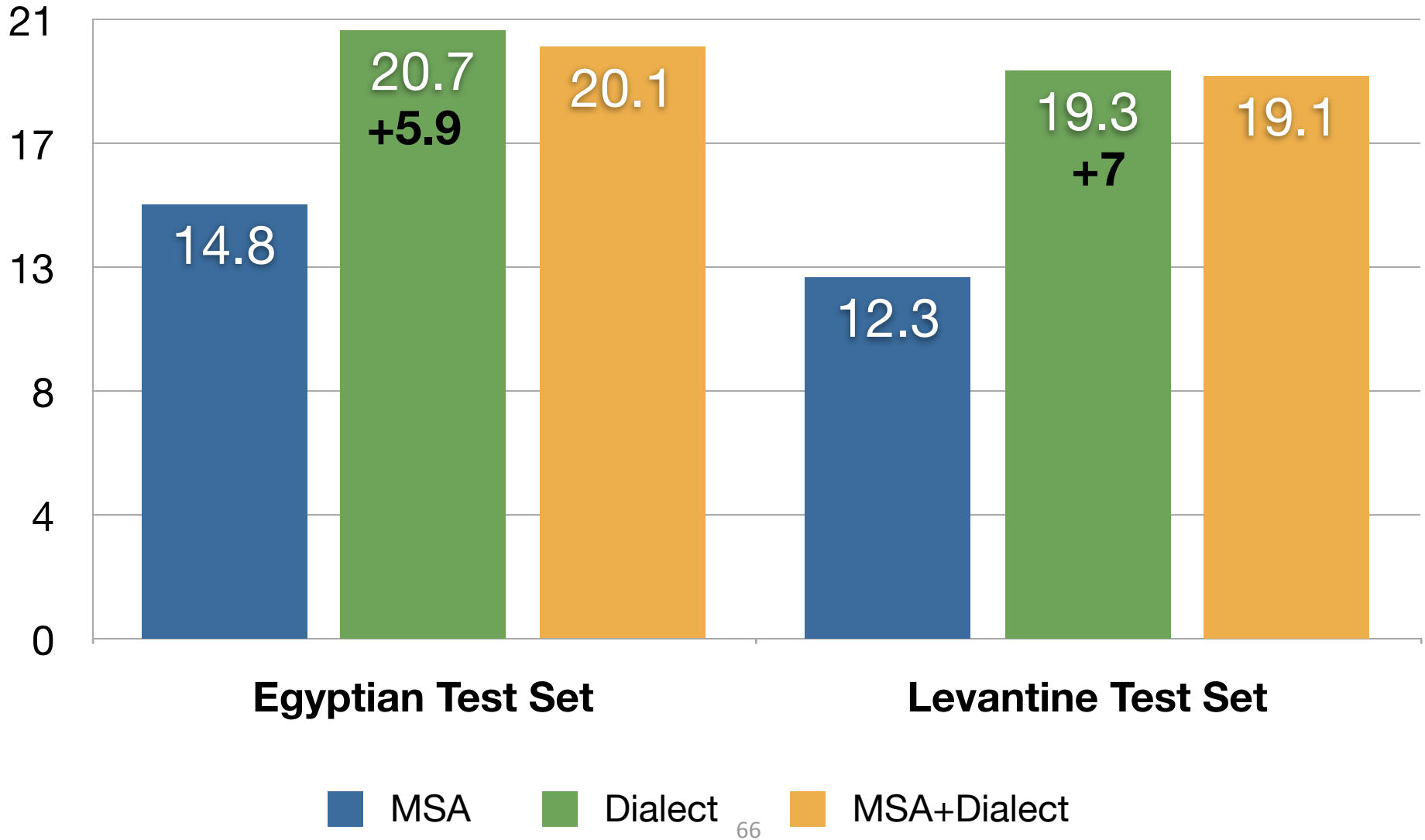
Examples of Dialect Translation

Dialect Input	MSA system	Dialect system	Reference
EGY لـمعتب تنا الو نالـعا هـل !! ؟ هـاي	You are working for a declaration and not?	You are making the advertisement for him or what?	Are you promoting it or what?!!
EGY يسفن ام دعب هيلع نئمطا دي هروصلا فاش	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 وجلا كيهل لووكك رييييتك	God you the atmosphere.	This is why the weather is so cool	This is why the weather is so cool
LEV مع كلاب لوط حزمن	Do you think about a joke long.	Calm down we are kidding	Calm down, we are only kidding

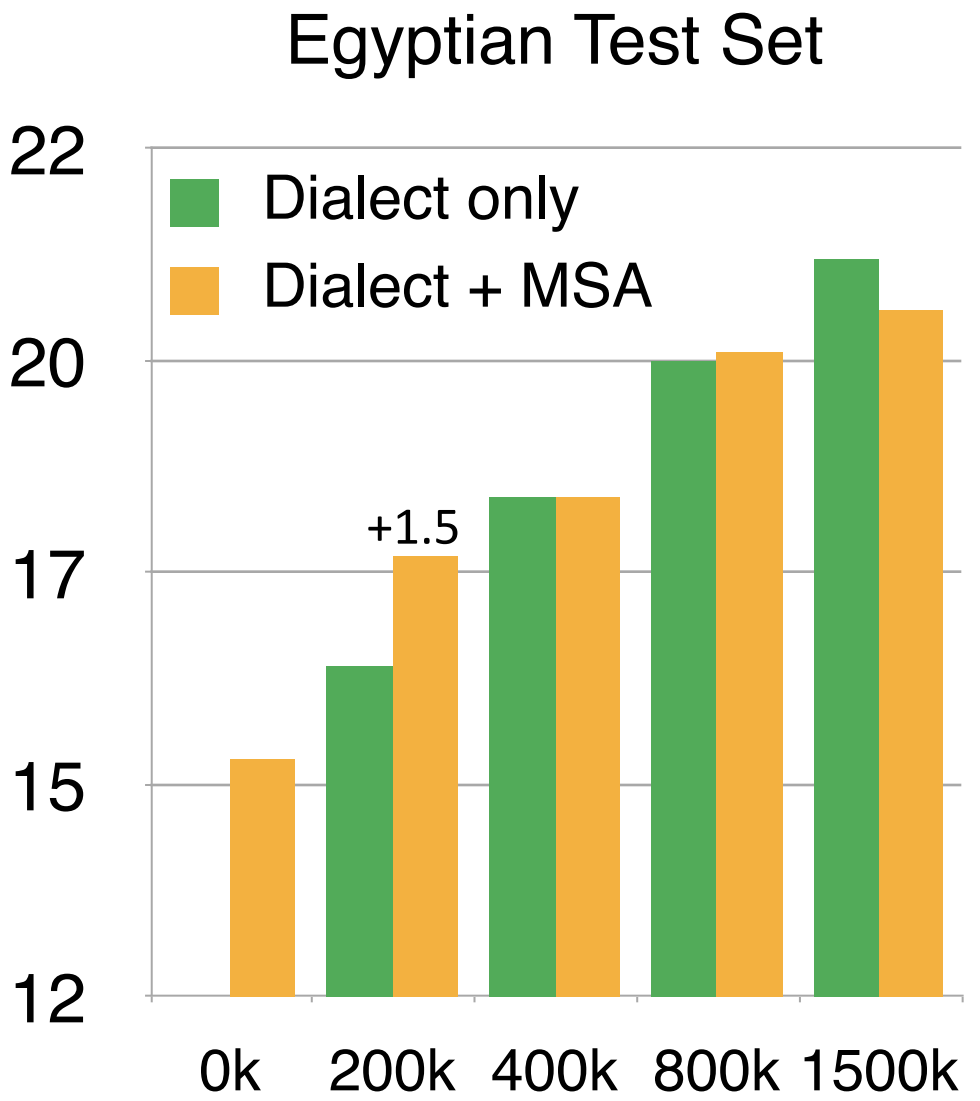


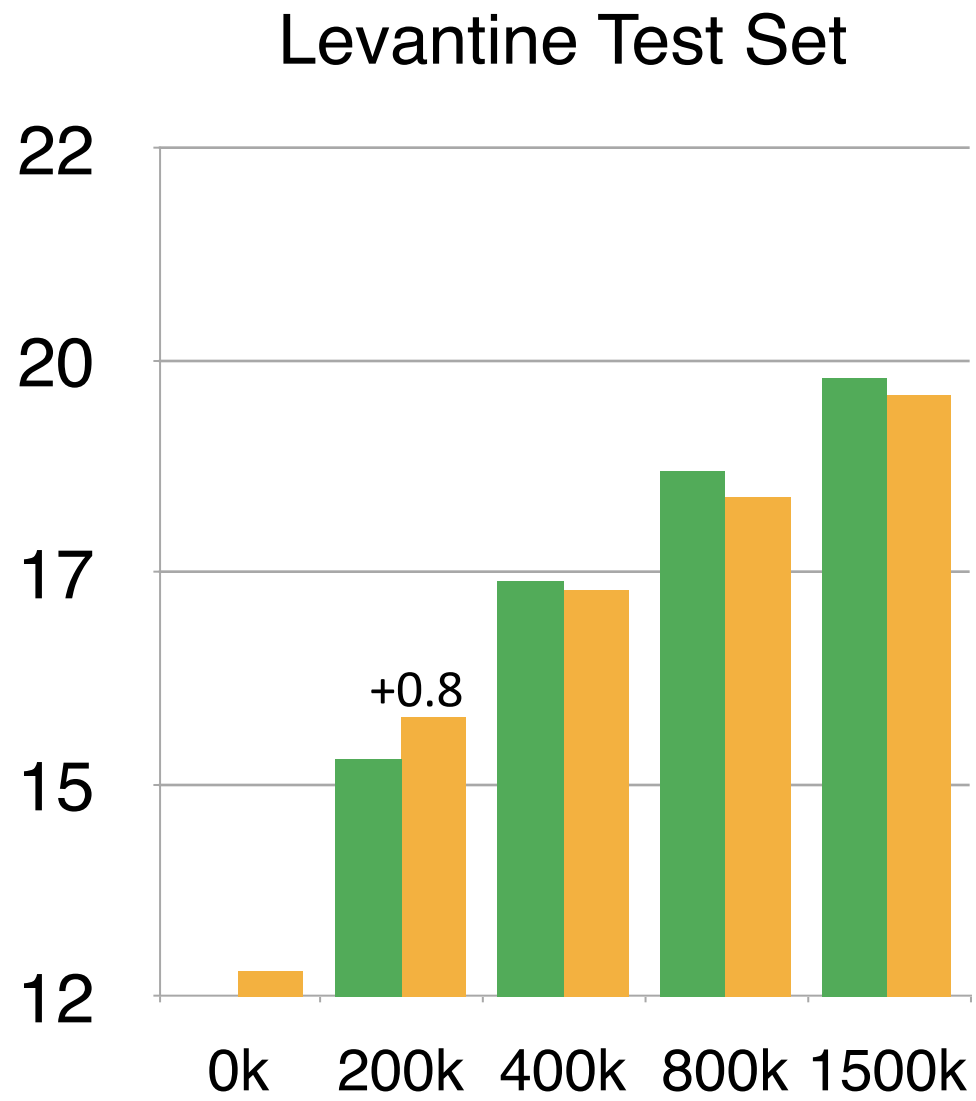
Dialect versus MSA

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



Value of MSA diminishes as dialect data increases





Dialect v. MSA takeaway

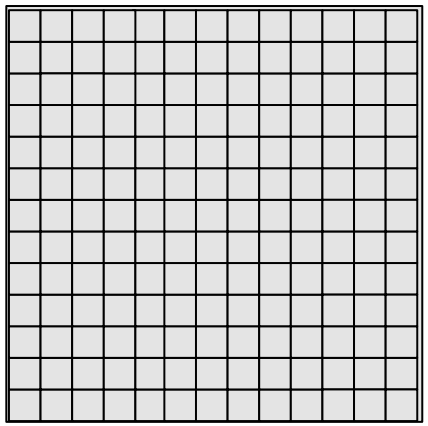
For dialect translation

1.5M

Arabic
Dialect-English

>

200M+



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.

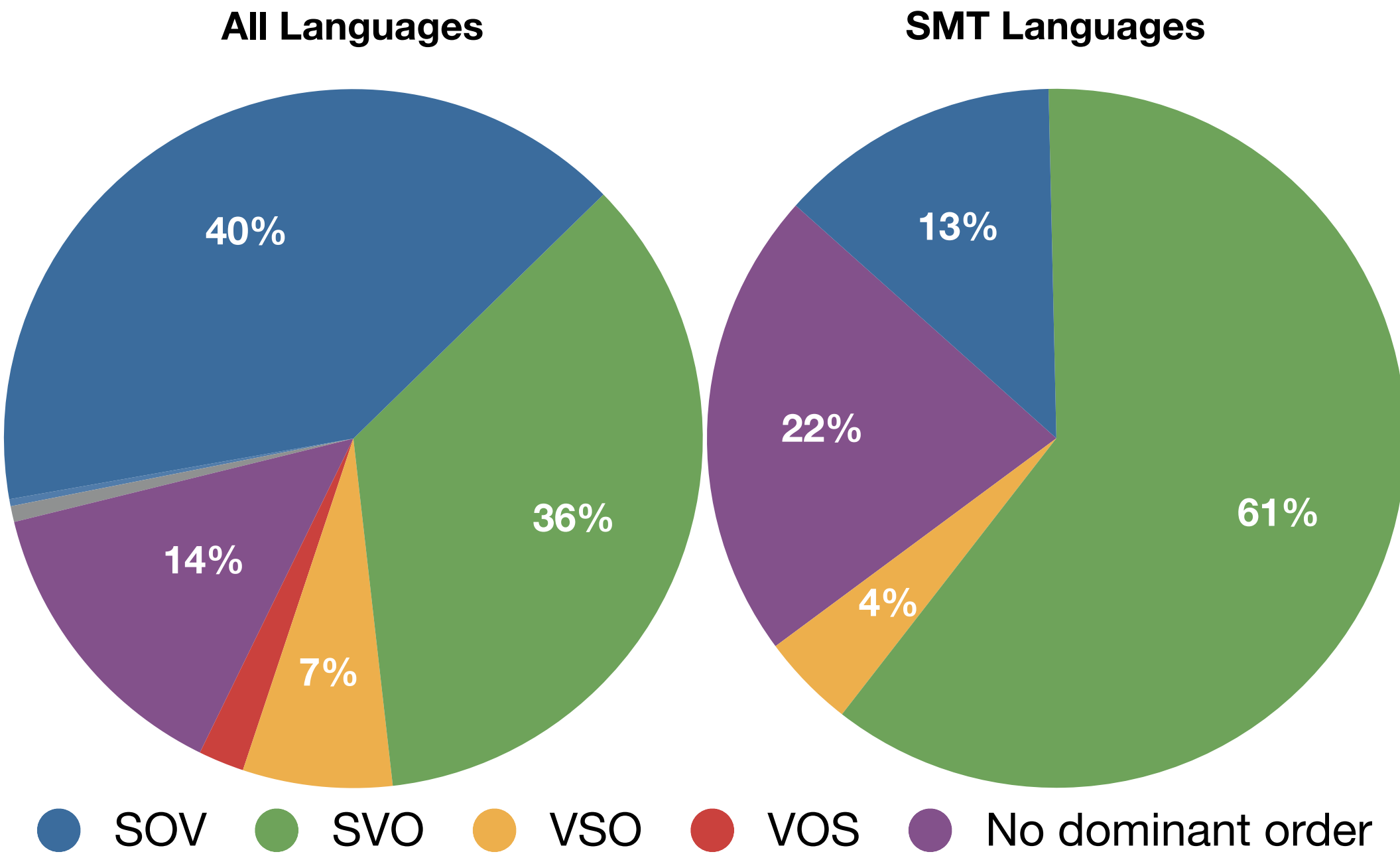
100 languages

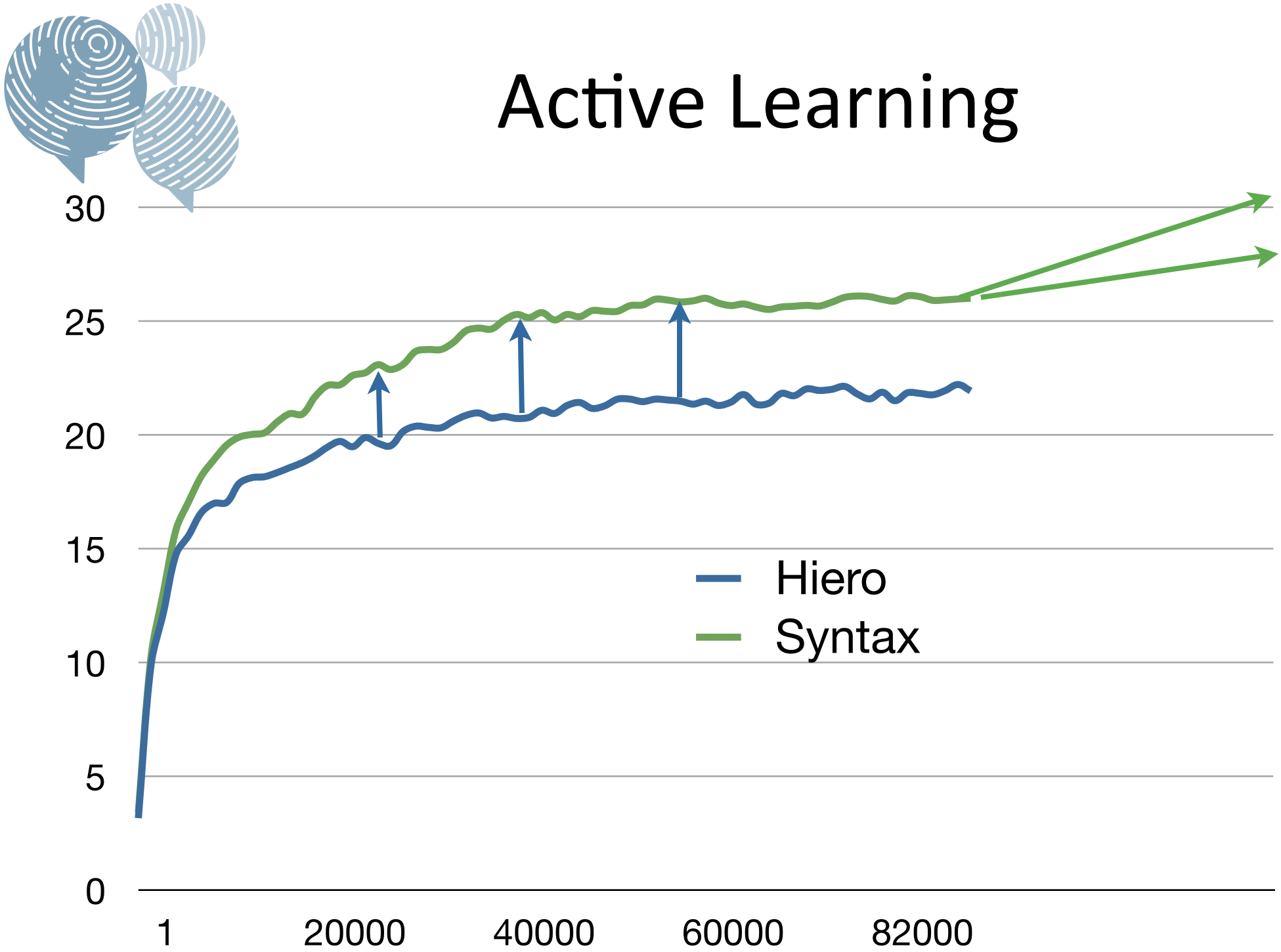
- Microsoft translator does 35 languages
- Google does 57 languages
- The DoD's Center for Applied MT does 64
- There is not enough data to reach acceptably high quality in new languages
- I want us to do 100 languages

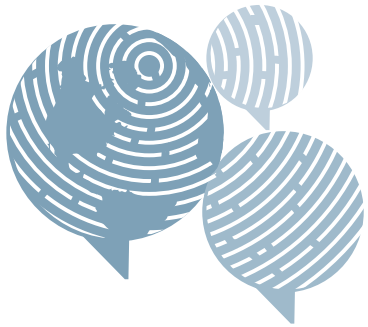
Google™

JOHNS HOPKINS UNIVERSITY

Better Representation of Languages




Active Learning
30
25
20
15
10
5
0
1
20000
40000
60000
82000
Hiero
Syntax



Questions?