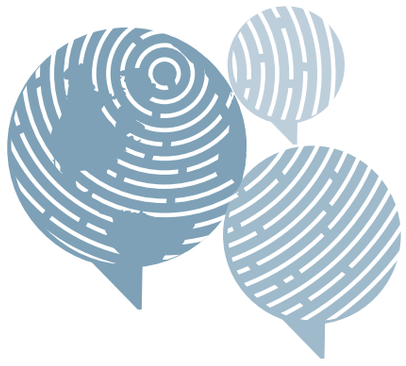
Constructing Parallel Corpora for Six Indian Languages via Crowdsourcing

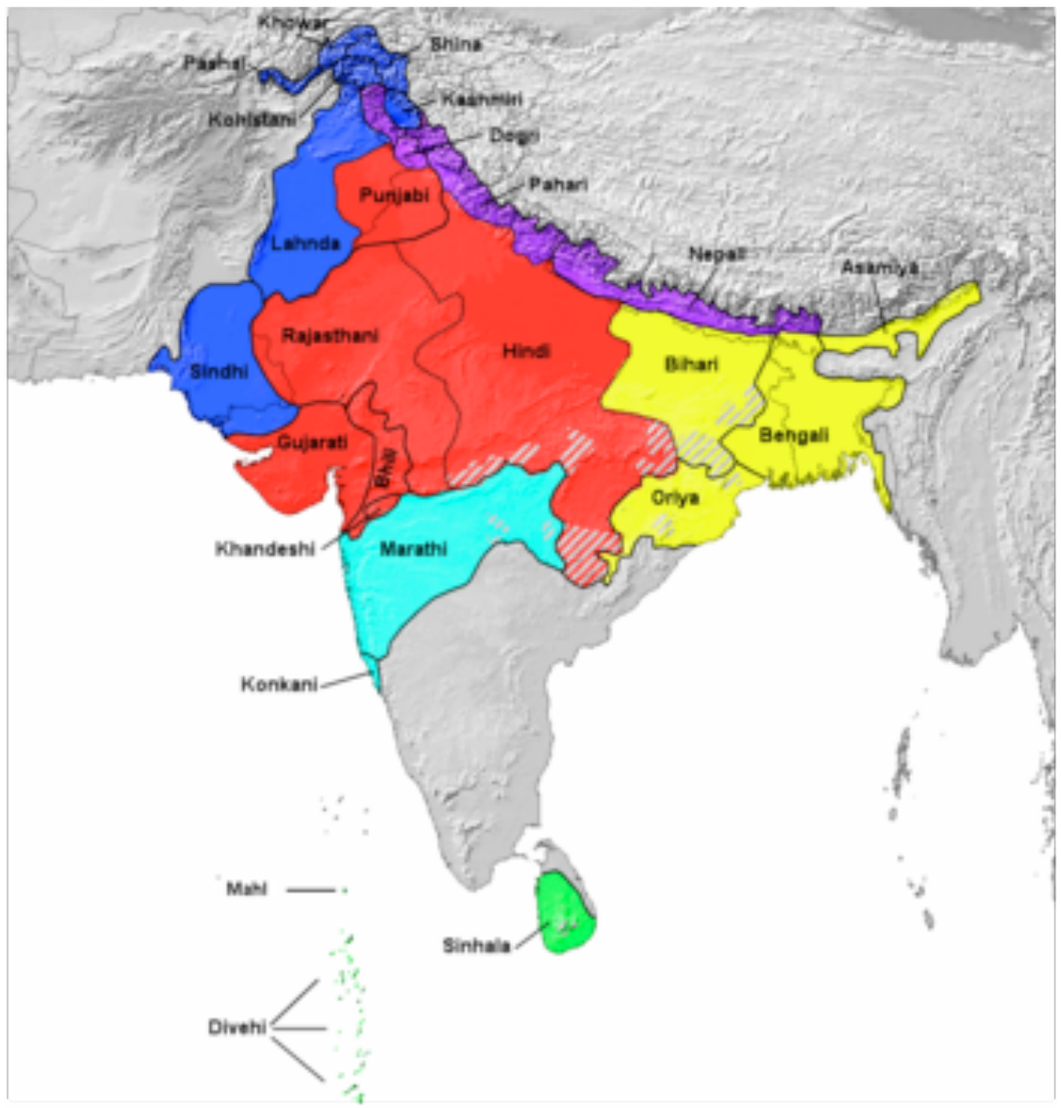
Matt Post, Chris Callison-Burch, and Miles Osborne
June 8, 2012



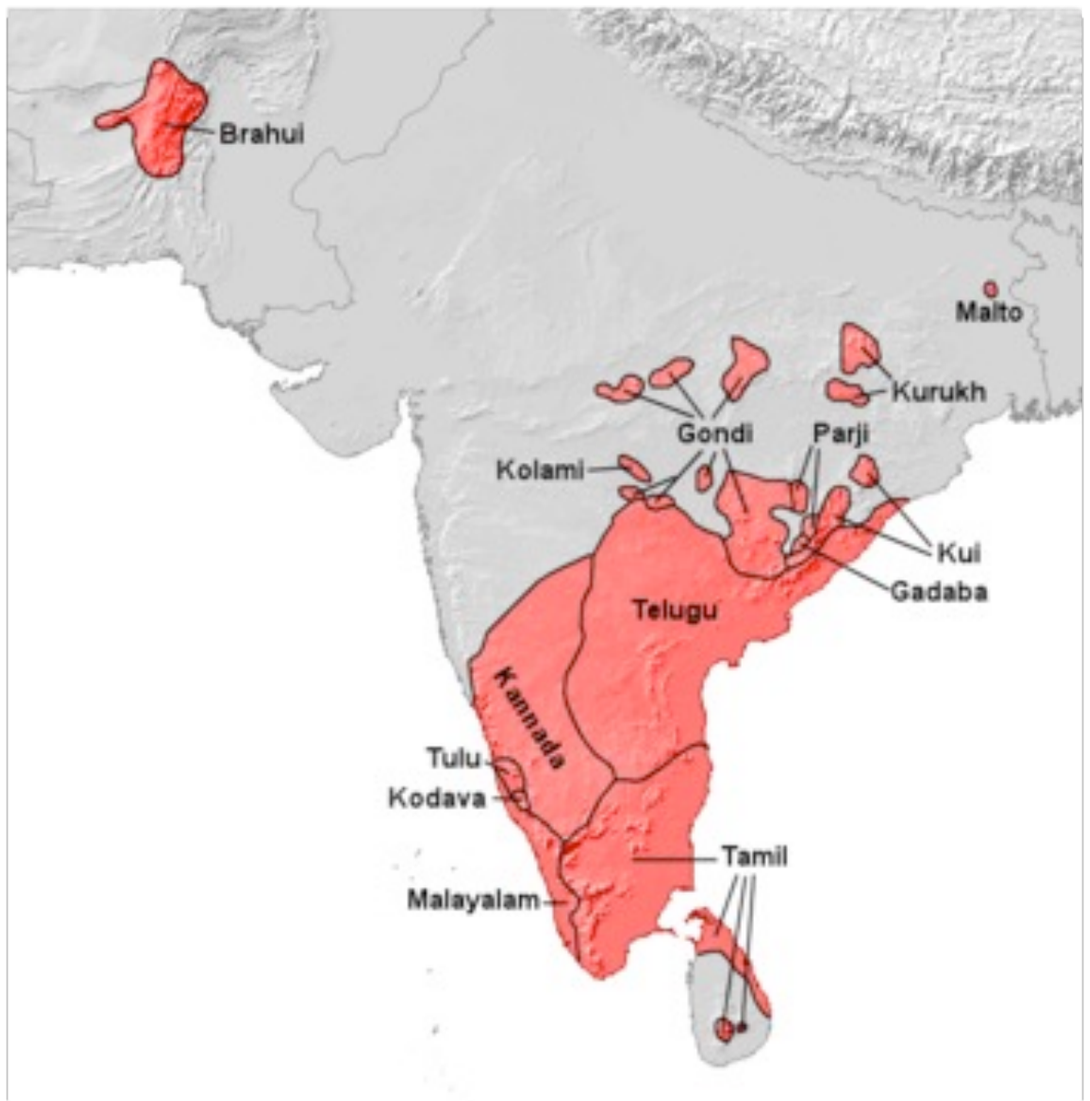
human language technology
center of excellence

JOHNS HOPKINS
UNIVERSITY

Languages

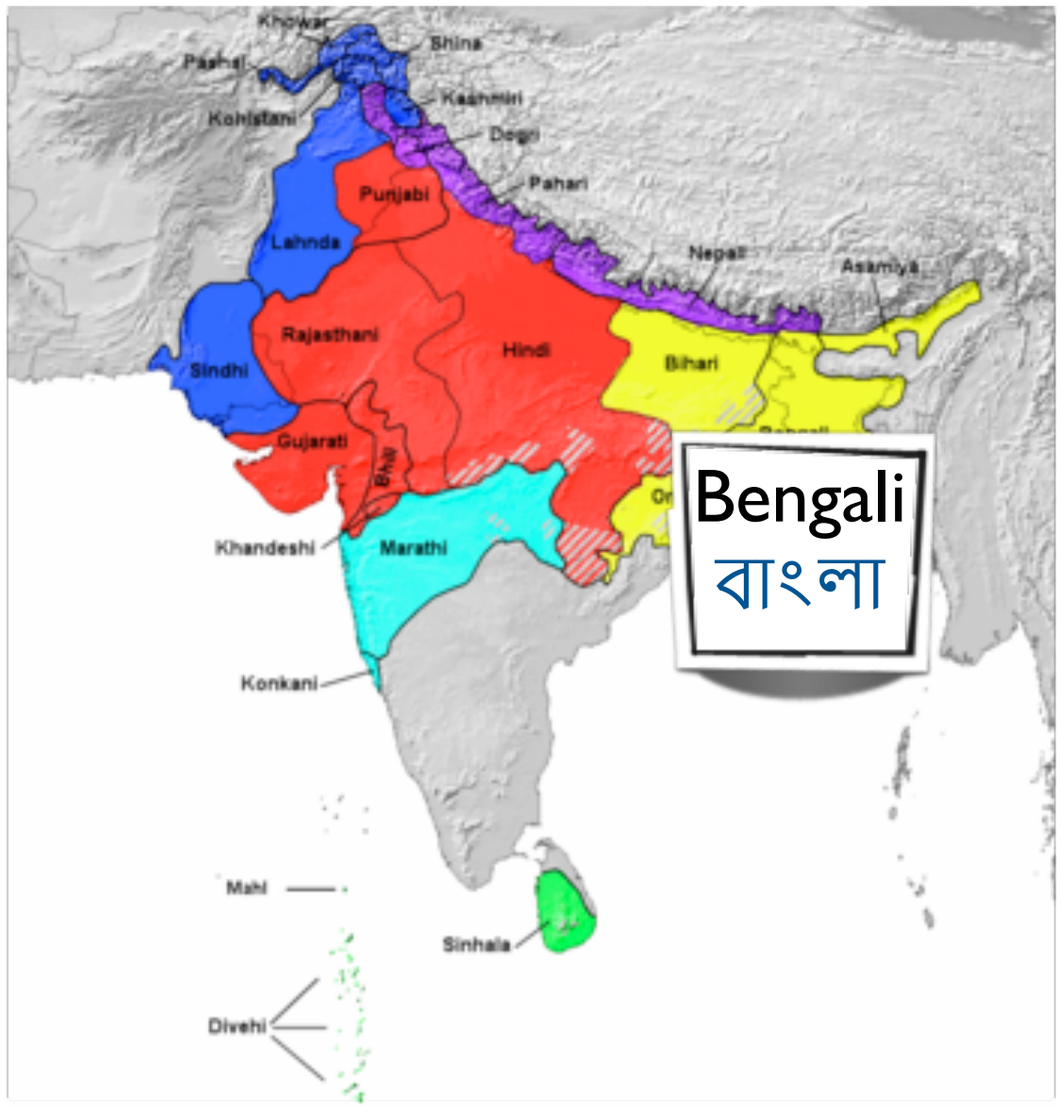


Indo-Aryan languages

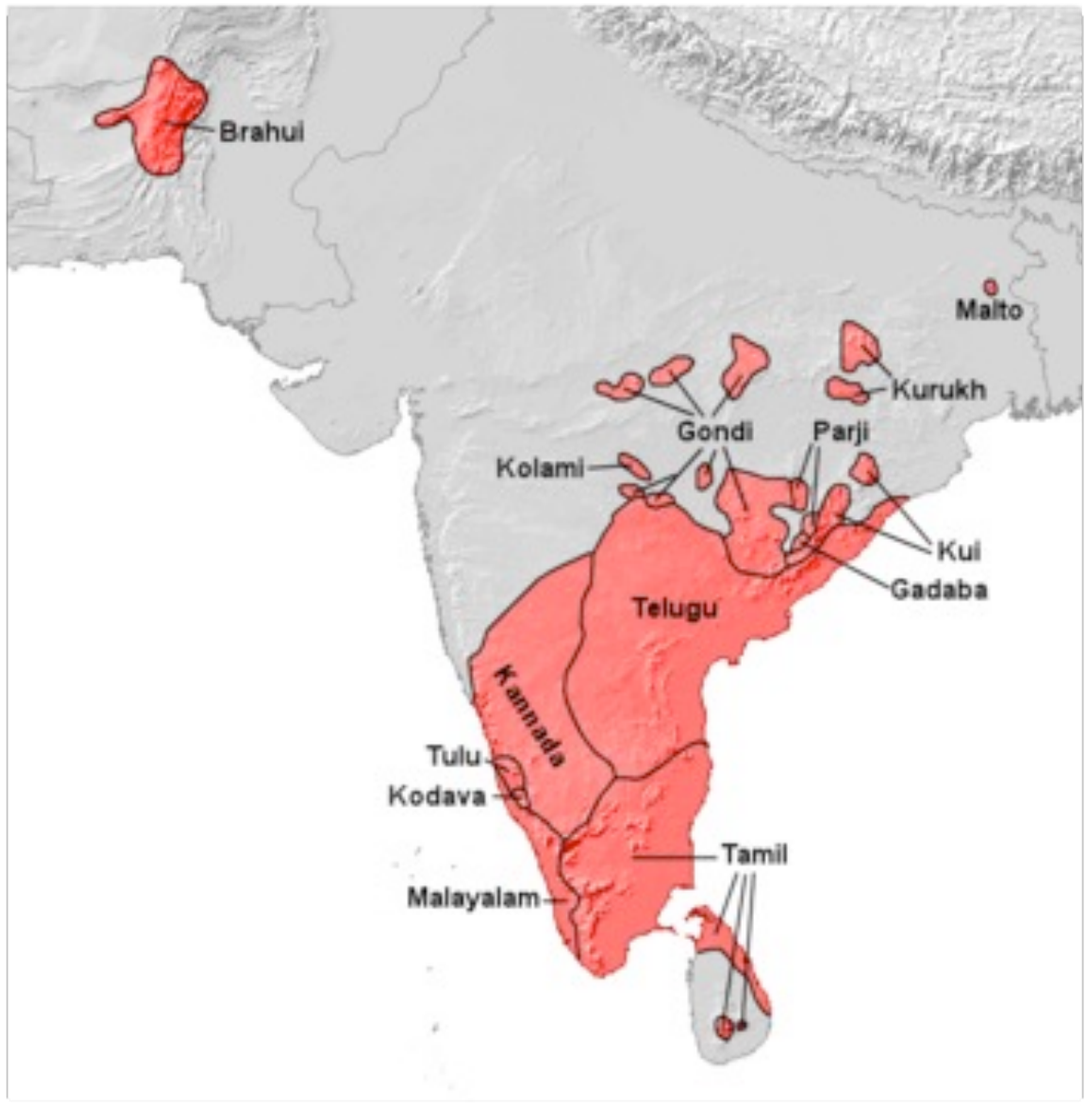


Dravidian languages

Languages

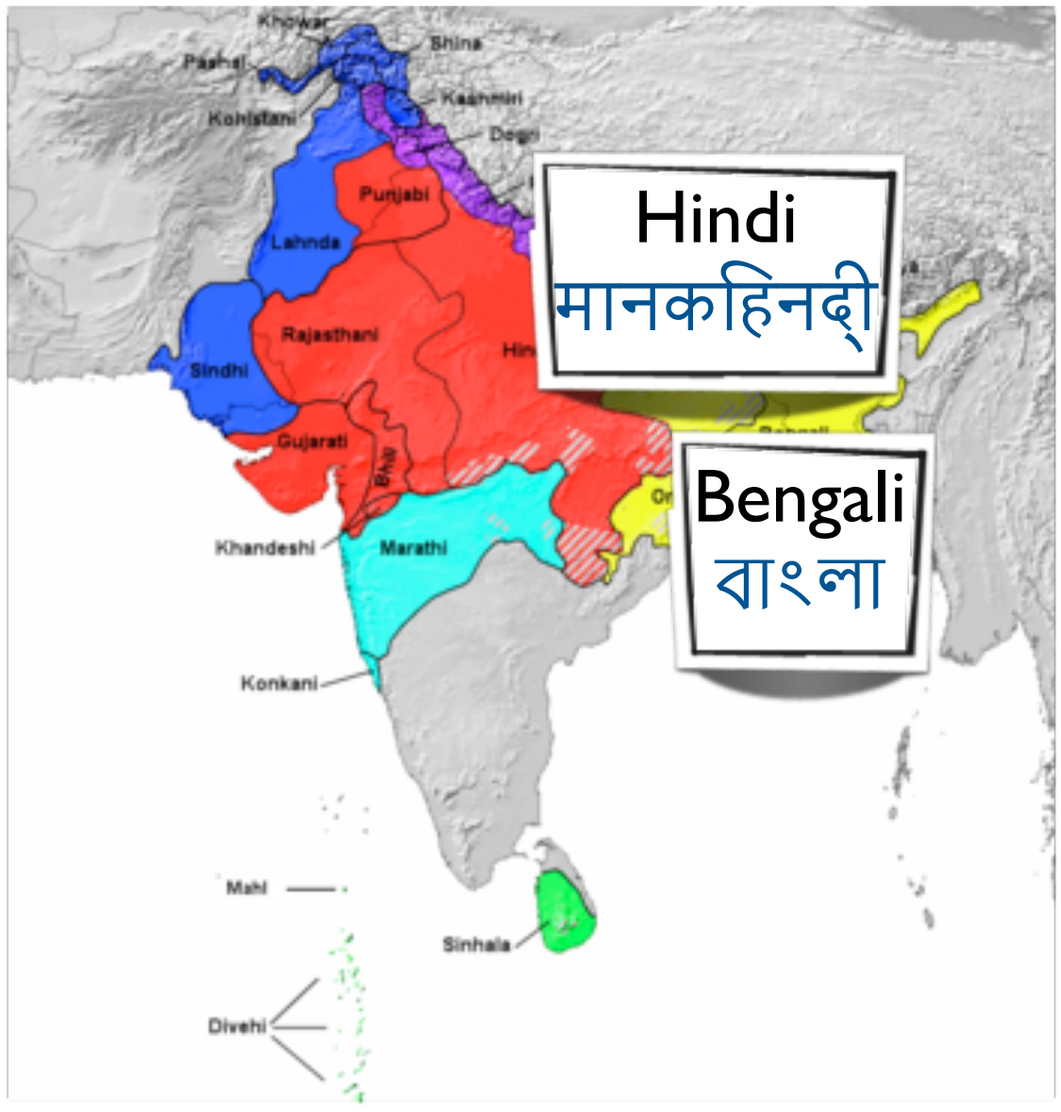


Indo-Aryan languages

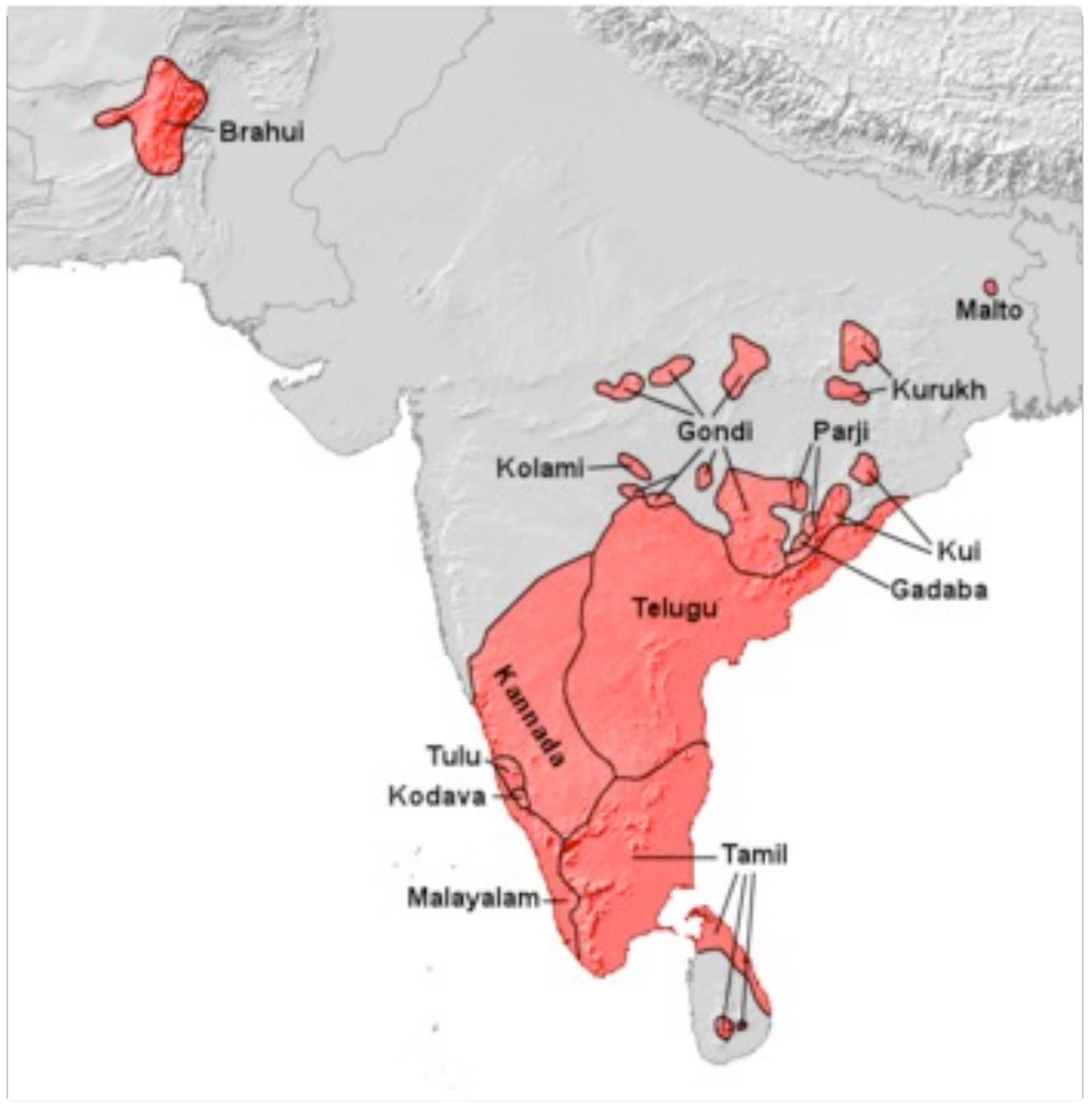


Dravidian languages

Languages

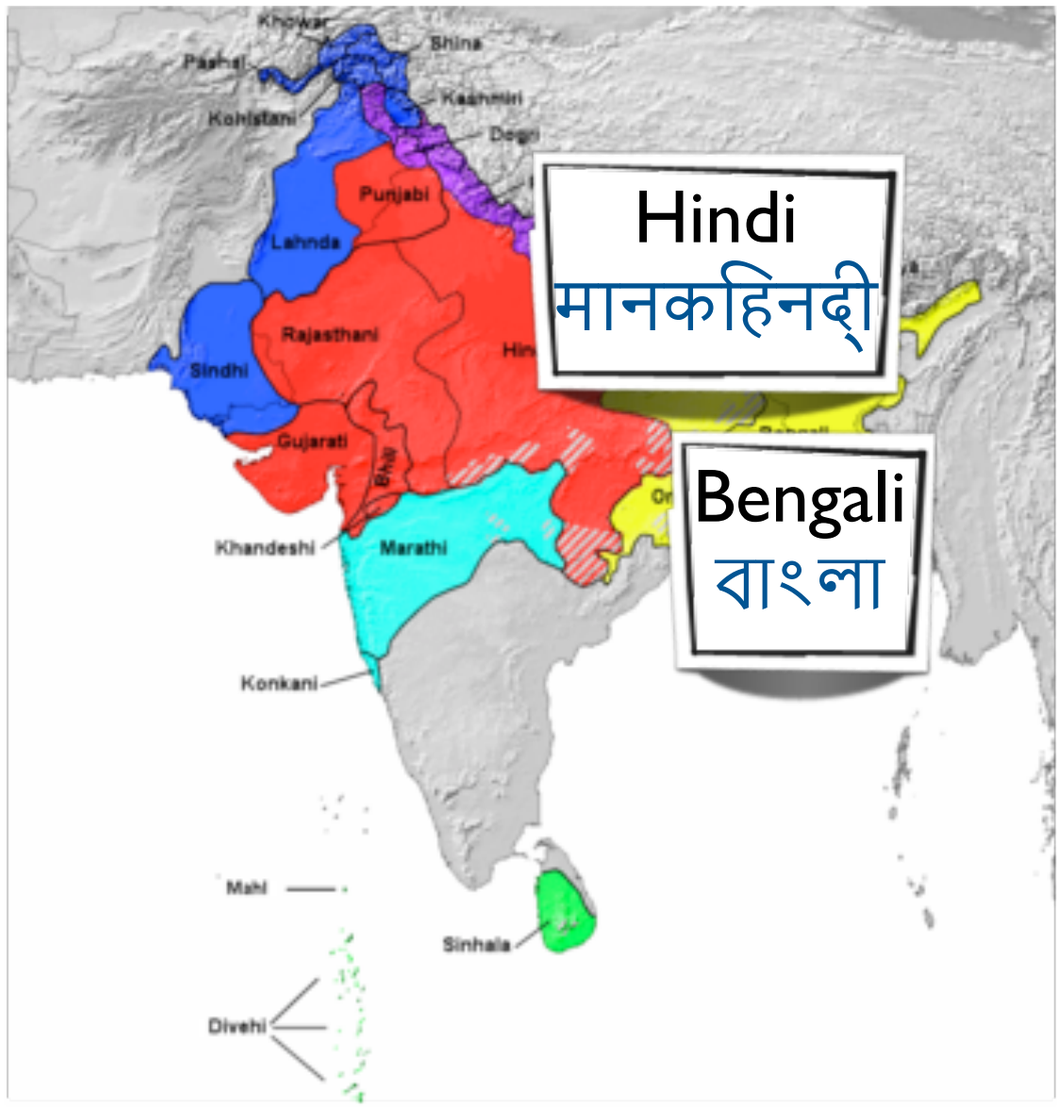


Indo-Aryan languages

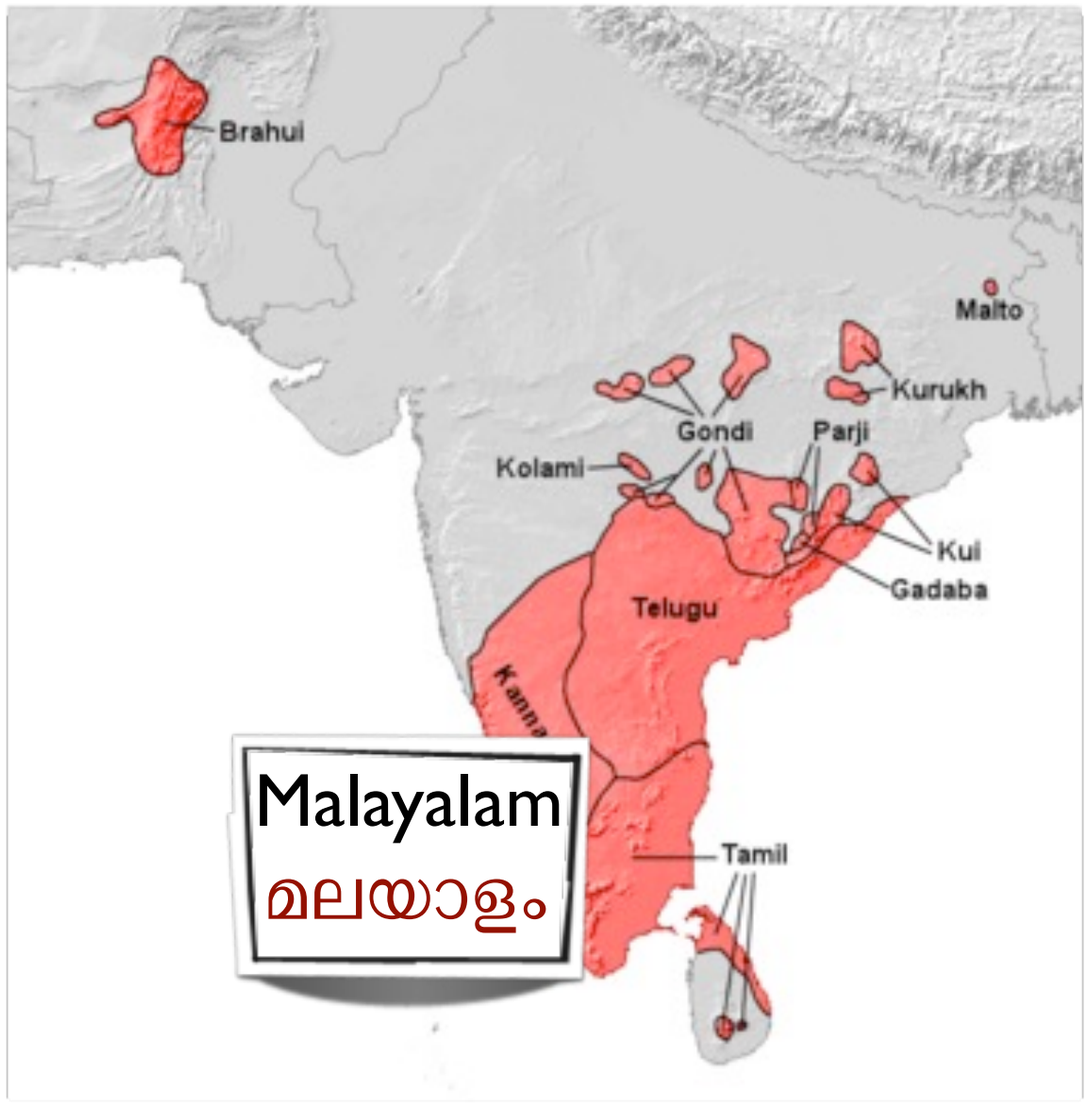


Dravidian languages

Languages

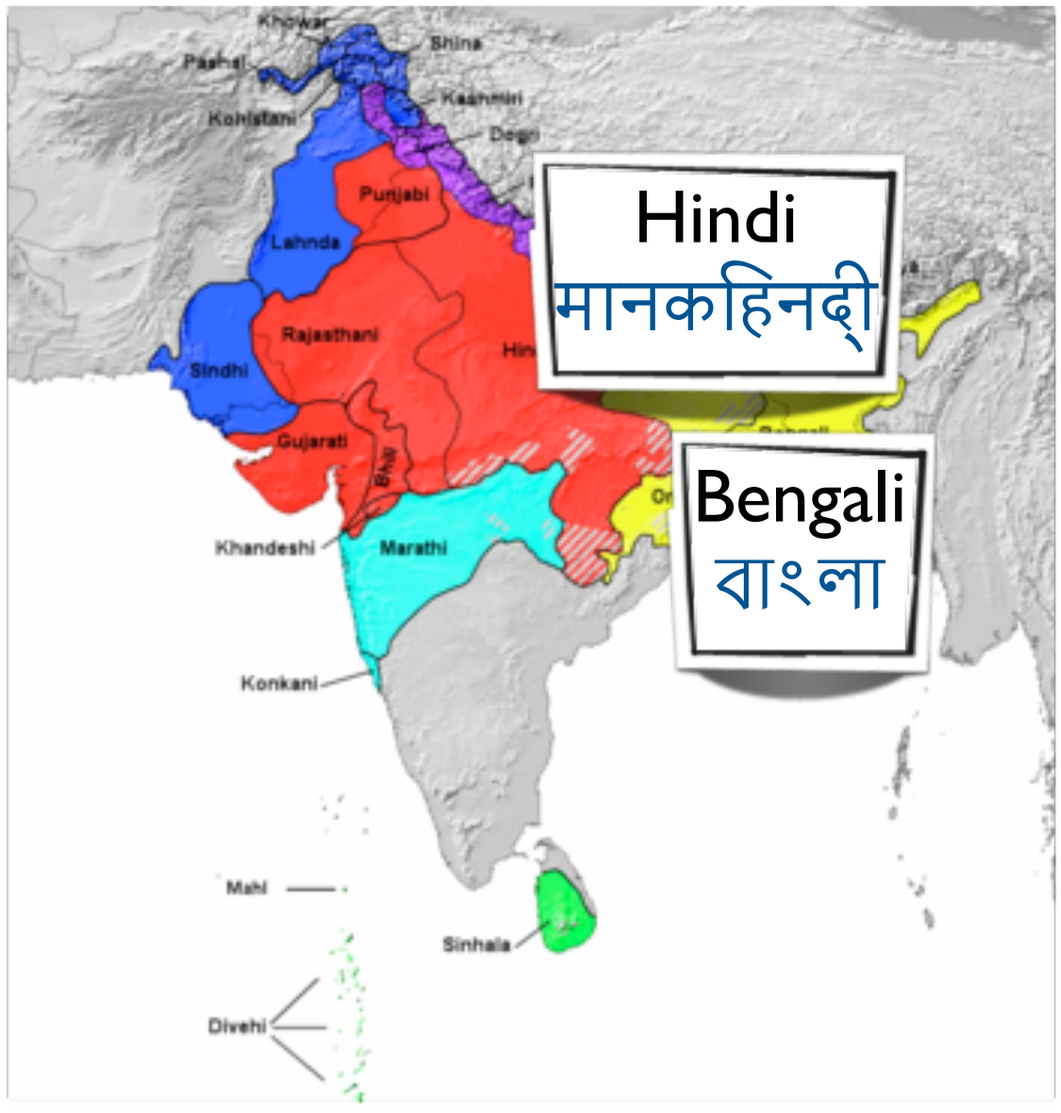


Indo-Aryan languages

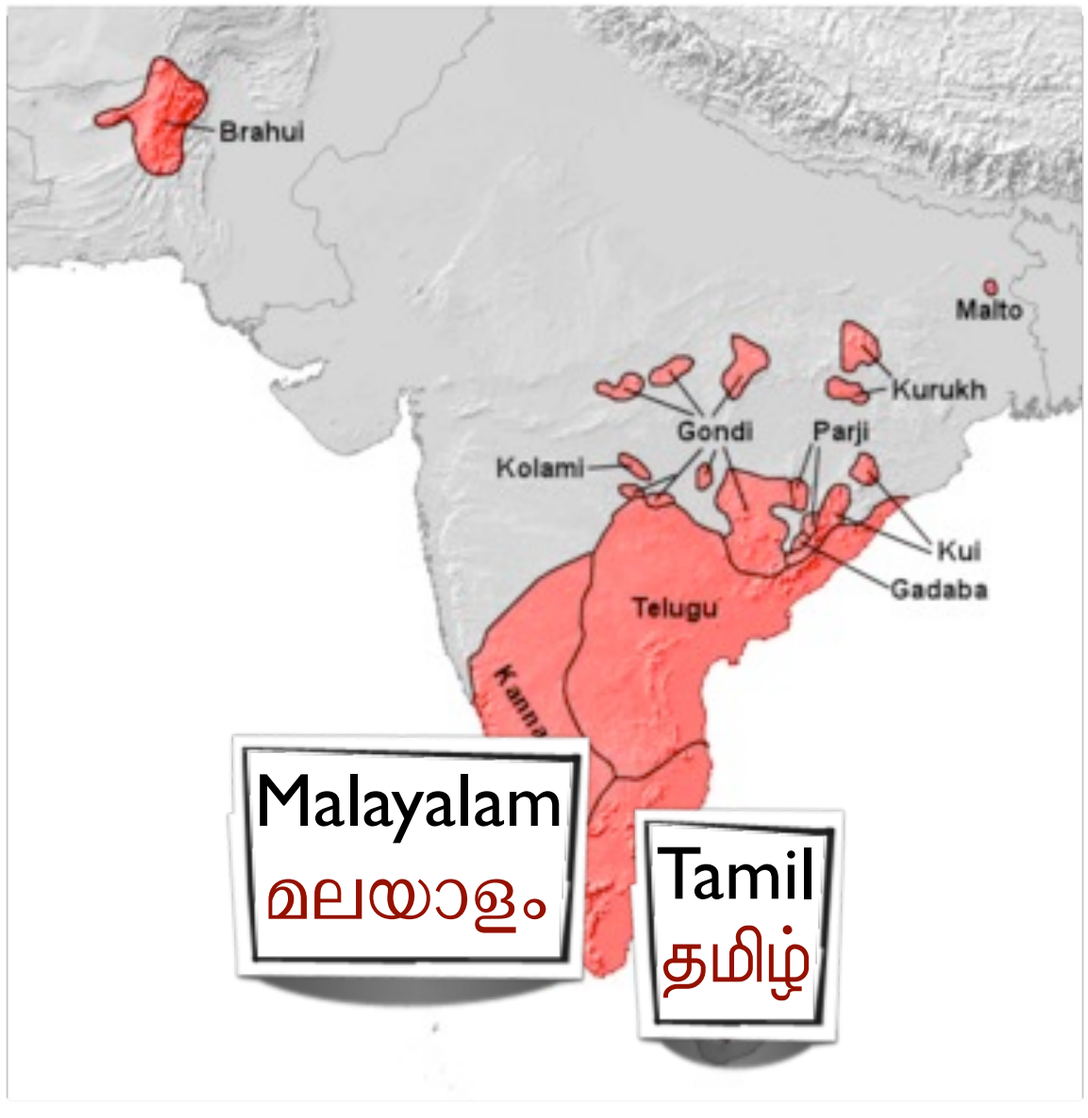


Dravidian languages

Languages

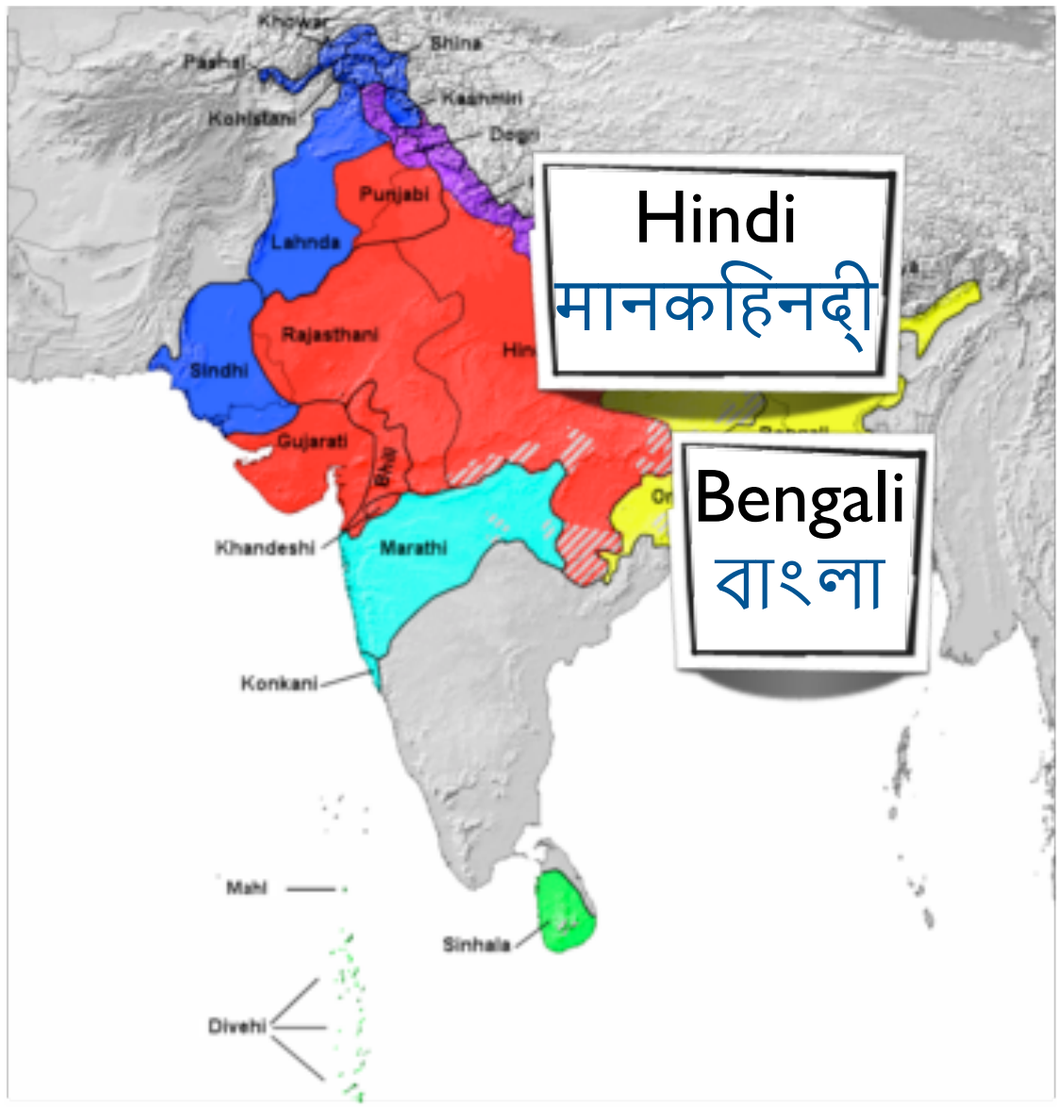


Indo-Aryan languages

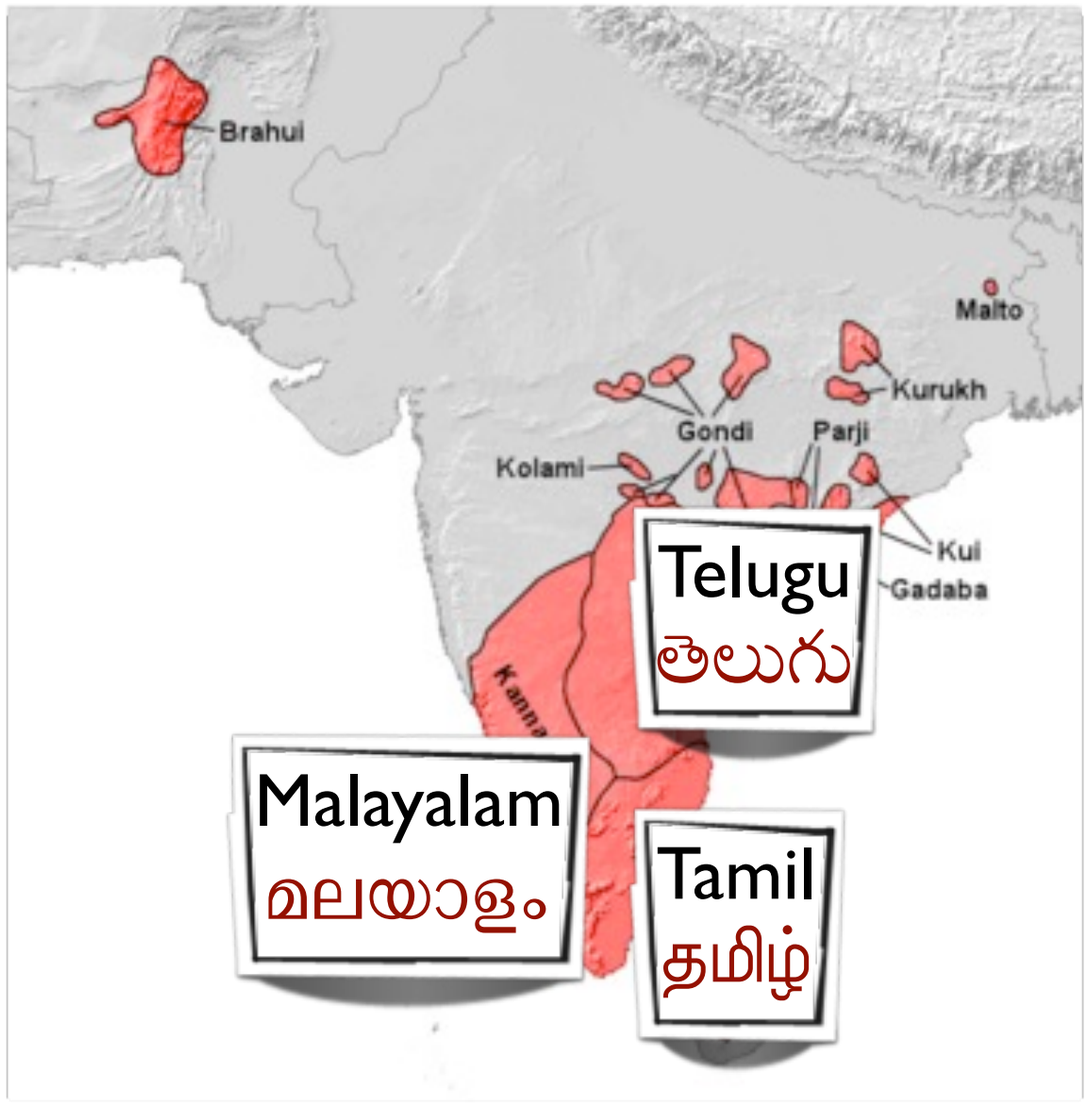


Dravidian languages

Languages

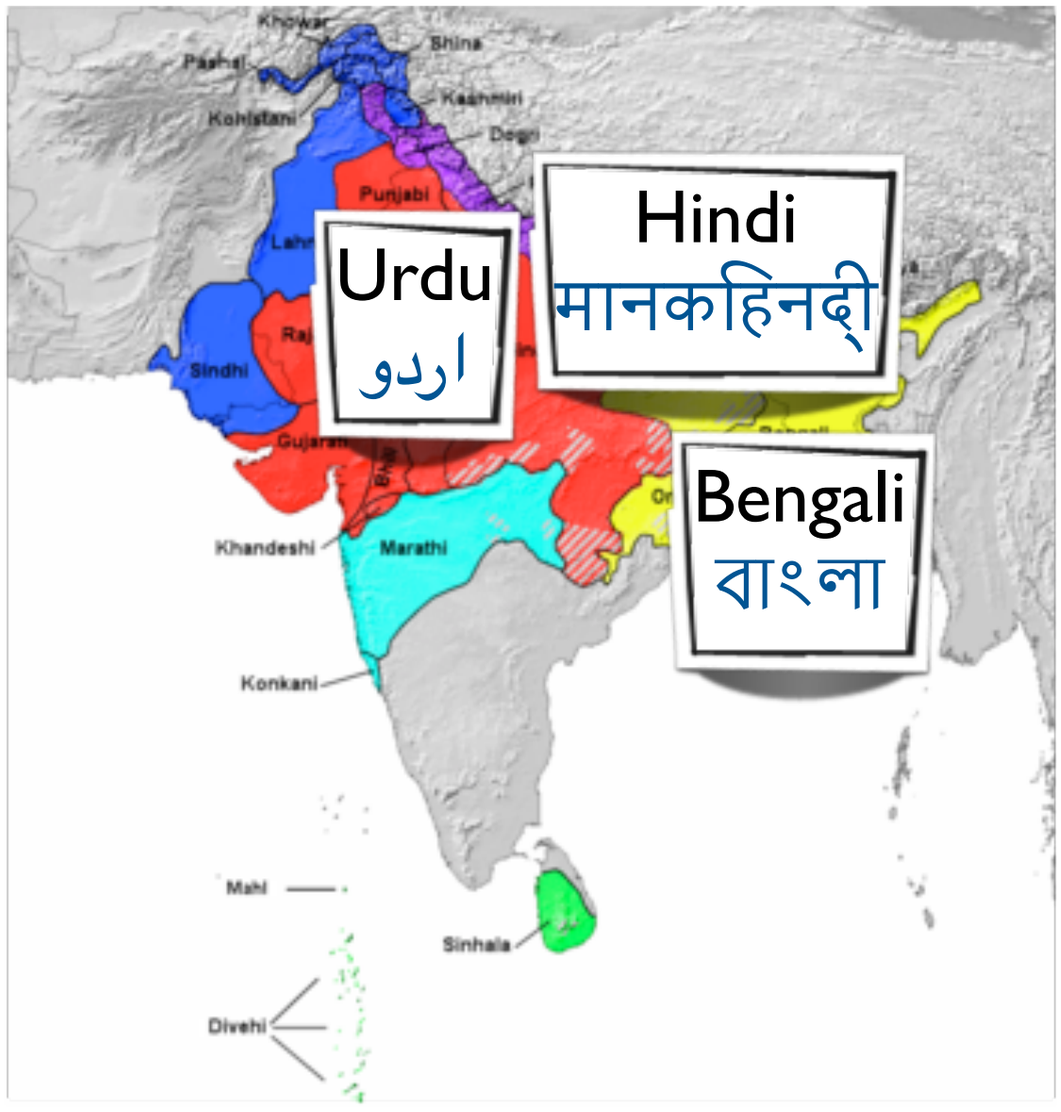


Indo-Aryan languages

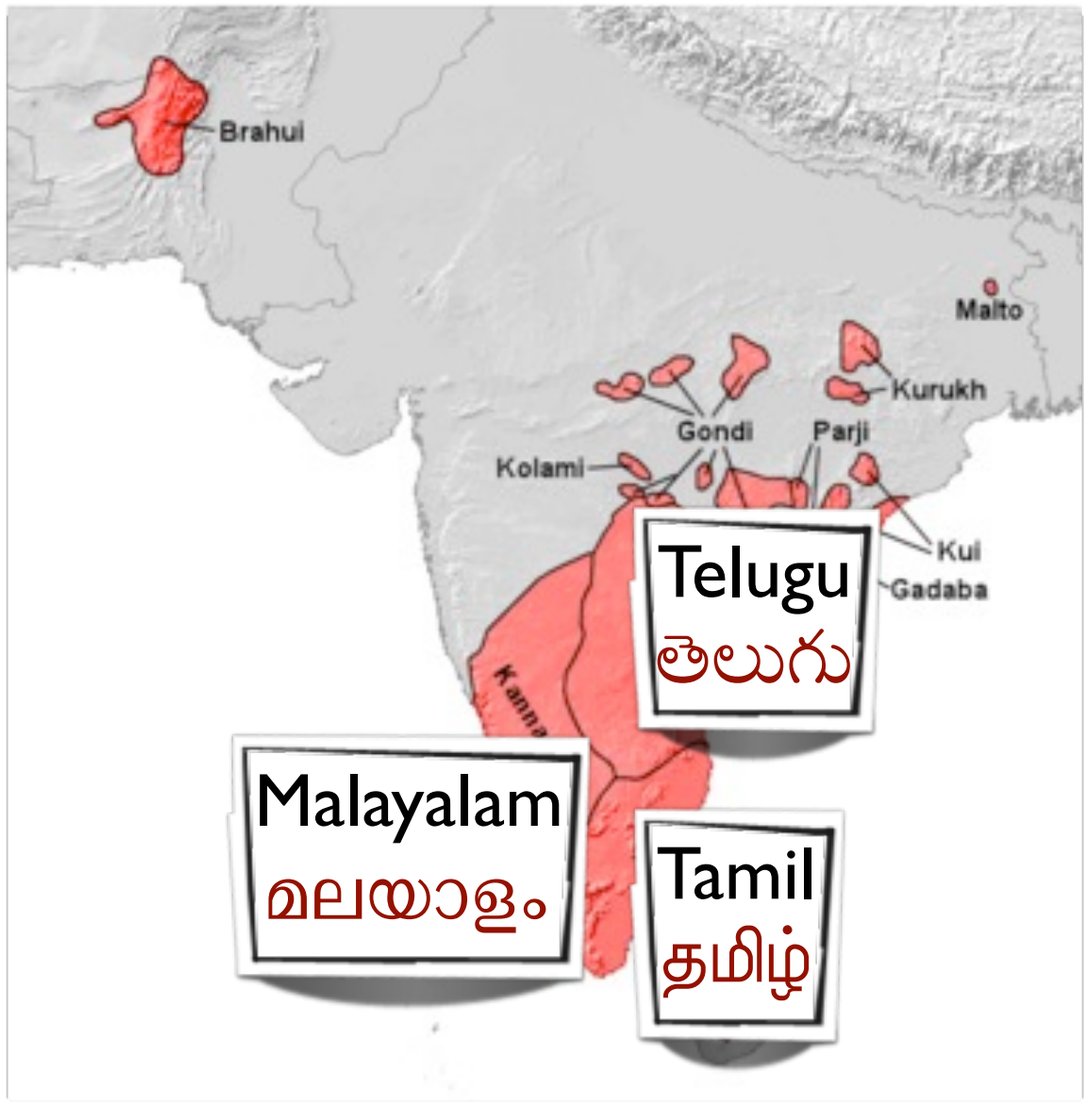


Dravidian languages

Languages



Indo-Aryan languages



Dravidian languages

Languages

language	script	family	L1 (millions)
Bengali	বাংলা	Indo-Aryan	110
Hindi	मानकहिन्दी	Indo-Aryan	180
Malayalam	മലയാളം	Dravidian	35
Tamil	தமிழ்	Dravidian	65
Telugu	తెలుగు	Dravidian	69
Urdu	اردو	Indo-Aryan	60

Indo-Aryan languages

Dravidian languages

Characteristics

- Head-final
  - Subject-object-verb (SOV) word order

"The senator prepared her remarks"			
செனட்டர்	அவளை	கருத்துக்கள்	தயார் .
senator	*her*	*remarks*	*prepared .*

- Agglutinative morphology
  - inflectional: tense, person, number, gender, mood, voice
    - e.g., ஈரிநீன் / eeRineen ("climbed")

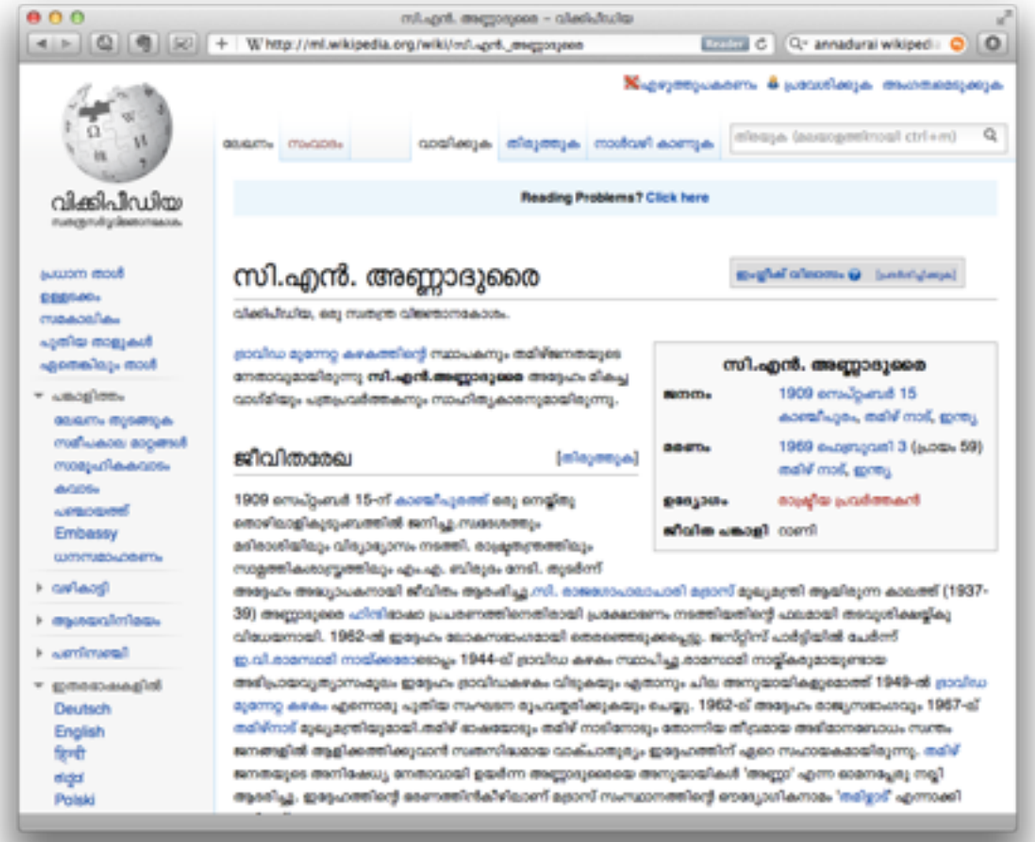
ஈறு	+	இன்	+	ஈன
eeRu		in		een
climb		*past*		*1p-sing-neuter*

www.google.com/transliterate/tamil

www.emille.lancs.ac.uk/lesal/tamil.pdf

Data collection

- We took the 100 most-popular Wikipedia articles in each language and translated them using Amazon's Mechanical Turk in a 3-step process
  1. Dictionary construction
  2. Page translation
  3. Vote gathering



amazon mechanical turk
beta
Artificial Artificial Intelligence

1. Dictionary construction

- We built a source-language vocabulary, and solicited translations for each word from Turkers
- Each word was presented along with four sentences it occurred in
- As controls, we used titles, which link pages in Wikipedia across languages

2. Translation

- We took the 100 most-popular Wikipedia articles in each language and translated them using Amazon's Mechanical Turk

அண்ணாதுரை மிகச் சிறந்த தமிழ் சொற்பொழிவாளரும் மேடைப் பேச்சாளரும் ஆவார் .

Annadurai best Tamil lecturer stage spokesman is .

amazon mechanical turk beta
Artificial Artificial Intelligence

four non-expert translations

annadurai was an excellent orator and a public speaker .
annadurai is a very good speaker
annadurai is one of the best tamil speeches and also stage speeches .
annathurai was the great reader and also the stage speaker .

3. Votes

- In a final task, we collected five votes on which of the four translations was the best

அண்ணாதுரை மிகச் சிறந்த தமிழ் சொற்பொழிவாளரும் மேடைப் பேச்சாளரும் ஆவார் .

Annadurai best Tamil lecturer stage spokesman is .

training sentences 8504-8507

amazon mechanical turk
Artificial Artificial Intelligence

four non-expert translations

annadurai was an excellent orator and a public speaker .
annadurai is a very good speaker
annadurai is one of the best tamil speeches and also stage speeches .
annathurai was the great reader and also the stage speaker .

3. Votes

- In a final task, we collected five votes on which of the four translations was the best

அண்ணாதுரை மிகச் சிறந்த தமிழ் சொற்பொழிவாளரும் மேடைப் பேச்சாளரும் ஆவார் .

Annadurai best Tamil lecturer stage spokesman is .

training sentences 8504-8507

amazon mechanical turk
Artificial Artificial Intelligence

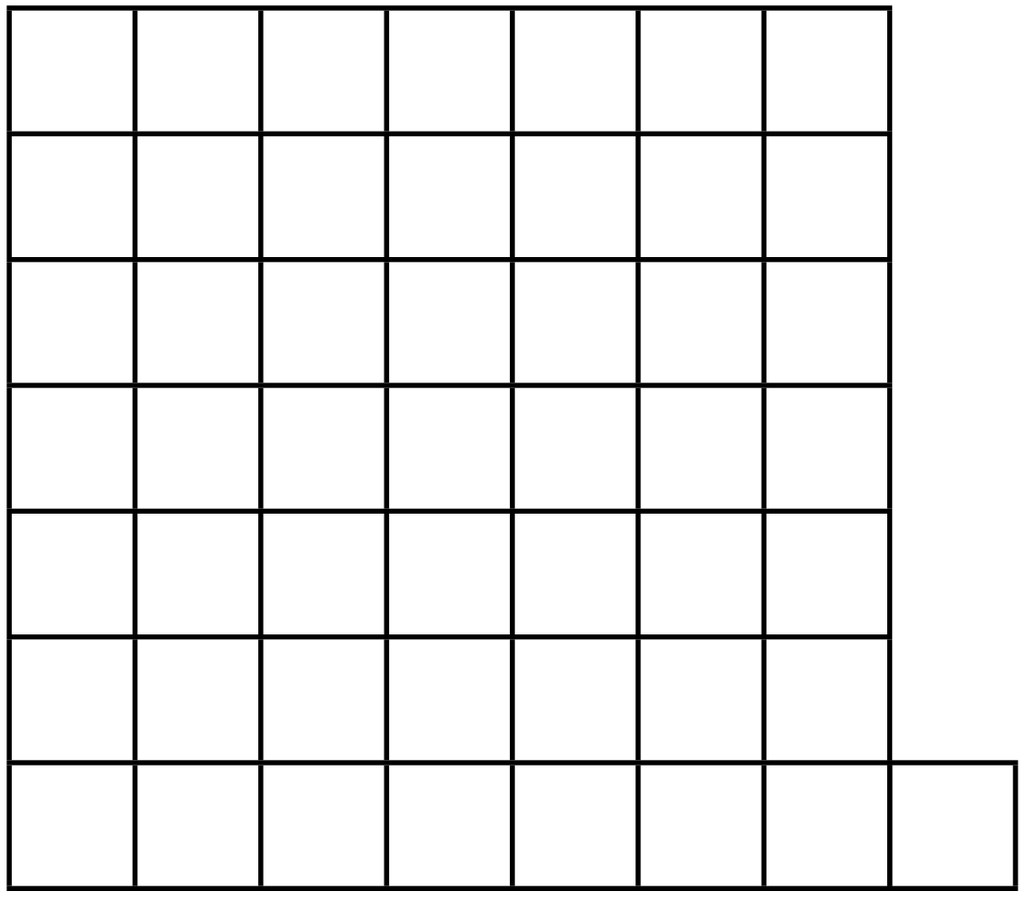
four non-expert translations

5 nnadurai was an excellent orator and a public speaker .
annadurai is a very good speaker
annadurai is one of the best tamil speeches and also stage speeches .
annathurai was the great reader and also the stage speaker .

Data collection

- Obtained about 500K English words for training, another 35K for tuning and testing

Indic languages
0.5m English words



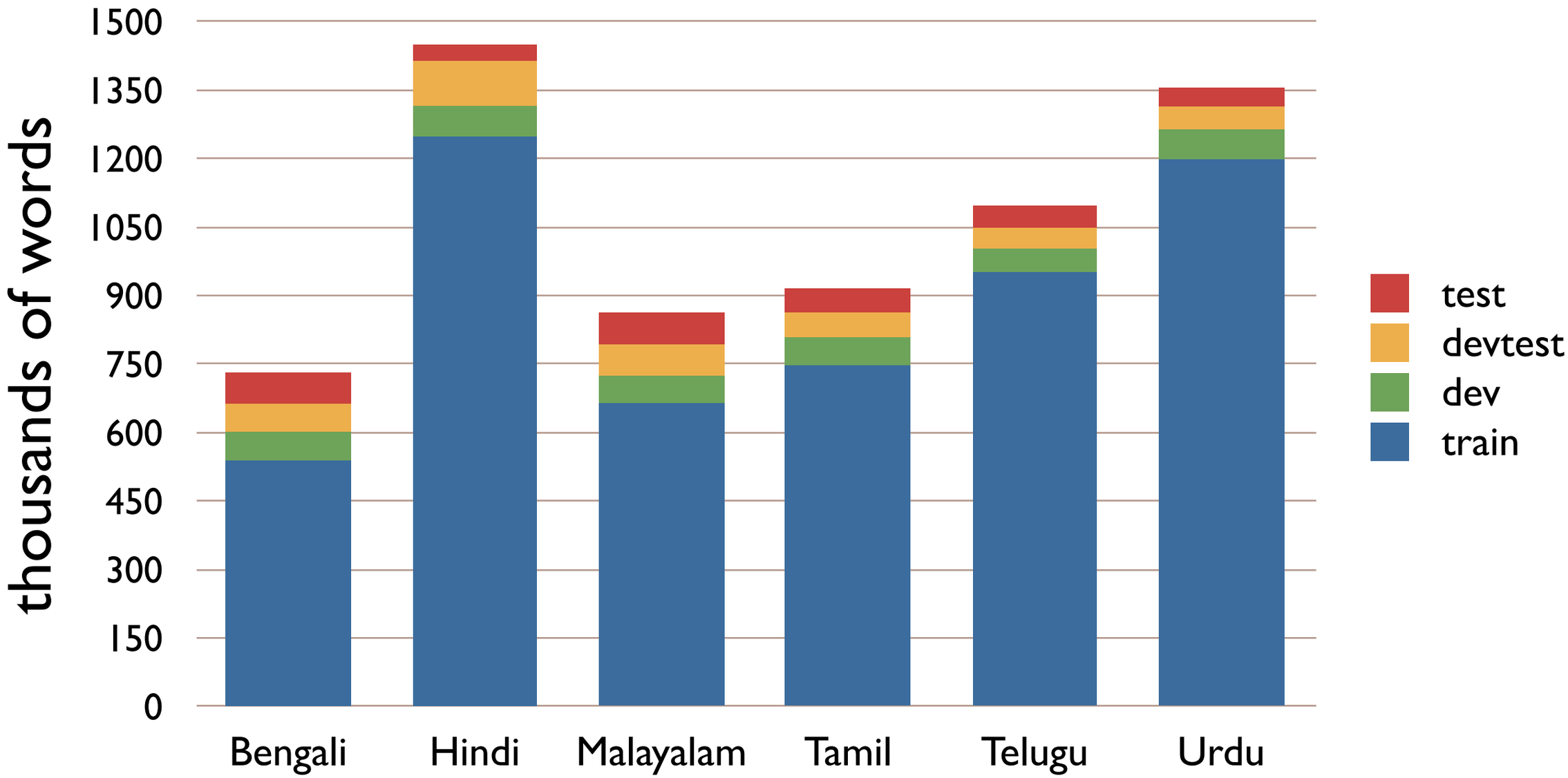
Europarl ES-EN
50m English words

Data splits

- We produce four datasets: TRAIN, DEV, DEVTEST, TEST
- Steps:
  - We manually assigned documents to one of seven categories
  - We assigned categories to datasets in round-robin fashion

PLACES	PEOPLE	THINGS	SEX	RELIGION
Agra	Gautama Buddha	*Air pollution*	Anal sex	Bhagavad Gita
Bihar	Harivansh Rai Bachchan	**Earth**	**Kama Sutra**	Diwali
China	Indira Gandhi	*Essay*	**Masturbation**	Hanuman
Delhi	Jaishankar Prasad	*Ganges*	**Penis**	Hinduism
Himalayas	Jawaharlal Nehru	*General knowledge*	Sex positions	Holi
India	Kabir	**Global warming**	**Sexual intercourse**	**Islam**
Mumbai	Kalpana Chawla	*Pollution*	**Vagina**	Mahabharata
Nepal	Mahadevi Varma	*Solar energy*	**LANGUAGE & CULTURE**	Puranas
Pakistan	Meera	*Terrorism*		Quran
Rajasthan	Mohammed Rafi	**TECH**	Ayurveda	Ramayana
Red Fort	**Mahatma Gandhi**	Blog	Constitution of India	Shiva
Taj Mahal	Mother Teresa	**Google**	Cricket	Taj Majal: Shiva Temple?
United States	Navbharat Times	Hindi Web Reso	**English language**	Vedas
Uttar Pradesh	Premchand	**Internet**	Hindi Cable News	**Vishnu**
PEOPLE	Rabindranath Tagor	*Mobile phone*	Hindi literature	**PEOPLE**
A. P. J. Abdul Kalam	Rani Lakshmibai	News aggregator	Hindi-Urdu grammar	Subhas Chandra Bose
Aishwarya Rai	Sachin Tendulkar	RSS	*Horoscope*	Surdas
Akbar	Sarojini Naidu	**Wikipedia**	Indian cuisine	Swami Vivekananda
Amitabh Bachchan	**EVENTS**	YouTube	Sanskrit	Tulsidas
Barack Obama	History of India		Standard Hindi	
Bhagat Singh	**World War II**			
Dainik Jagran				

Data splits


thousands of words
1500
1350
1200
1050
900
750
600
450
300
150
0
Bengali
Hindi
Malayalam
Tamil
Telugu
Urdu
test
devtest
dev
train

Data quality issues

- Translation quality

அஜித்	குமாரின்	முதல்	வெற்றிப்	படம்	ஆசை	.
ajith	*kumar*	*first*	*successful*	*movie*	*assai*	.

training sentence 17

- Translations

Data quality issues

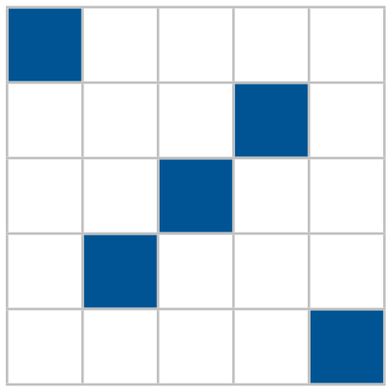
- Translation quality

அஜித்	குமாரின்	முதல்	வெற்றிப்	படம்	ஆசை	.
ajith	*kumar*	*first*	*successful*	*movie*	*assai*	.

training sentence 17

- Translations

- aasai was the first successfull movie for ajith kumar .



Data quality issues

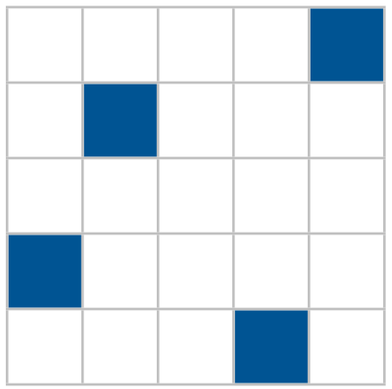
- Translation quality

அஜித்	குமாரின்	முதல்	வெற்றிப்	படம்	ஆசை	.
ajith	*kumar*	*first*	*successful*	*movie*	*assai*	.

training sentence 17

- Translations

- aasai was the first successfull movie for ajith kumar .
- first film by ajith kumar was ' asai '



Data quality issues

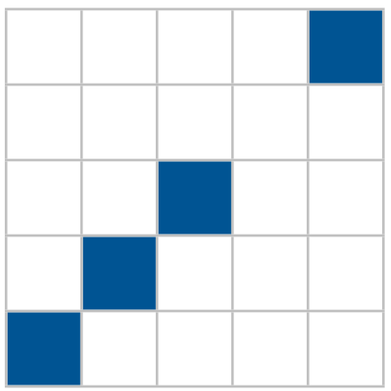
- Translation quality

அஜித்	குமாரின்	முதல்	வெற்றிப்	படம்	ஆசை	.
ajith	*kumar*	*first*	*successful*	*movie*	*assai*	.

training sentence 17

- Translations

- aasai was the first successfull movie for ajith kumar .
- first film by ajith kumar was ' asai '
- ajith kumar first victory is aasai



Data quality issues

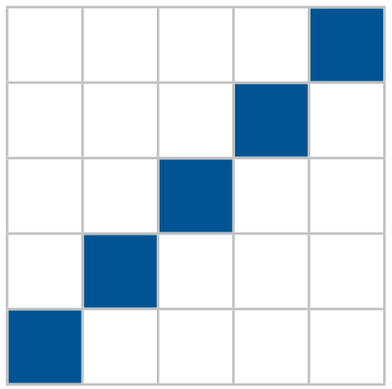
- Translation quality

அஜித்	குமாரின்	முதல்	வெற்றிப்	படம்	ஆசை	.
ajith	*kumar*	*first*	*successful*	*movie*	*assai*	.

training sentence 17

- Translations

- aasai was the first successfull movie for ajith kumar .
- first film by ajith kumar was ' asai '
- ajith kumar first victory is aasai
- ajithkumar first success movie is aasai



Data quality issues

- Inconsistent orthography

இலங்கையில்	சோழர்	ஆட்சி
In Sri Lanka	*Chola*	*ruled*

- Translations
  - in srilanka **solar government**
  - **chola** rule in sri lanka .
  - in srilanka **chozhas** ruled
  - **chola** reign in sri lanka

Data quality issues

- Poor alignments

Legend

- false positive
- true positive
- false negative

Research Questions

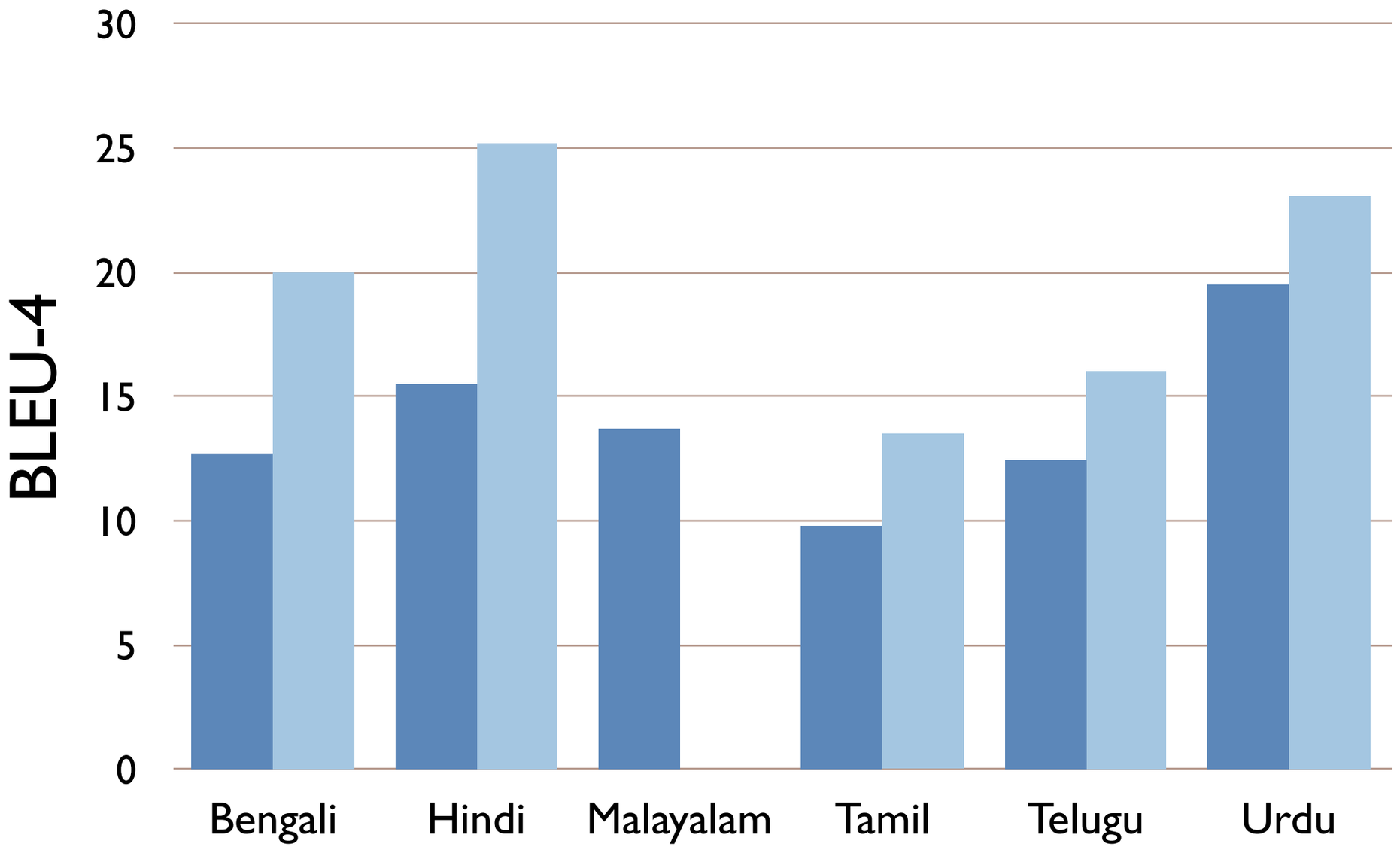
1. How well does SMT work on these languages?
2. Do linguistic annotations help?
3. How important is translation quality?

Q1: How well can we do?

- Hiero
  - Linguistically un-informed grammars that define lexicalized (re)orderings, extracted from aligned text

X → $X^{(1)}$ உறுதி செய்கிறது $X^{(2)}$, $X^{(1)}$ confirmed $X^{(2)}$

Q1: How well can we do?



scores are the mean of three MERT runs

Q2: Do linguistic annotations help?

- Syntax-augmented machine translation (SAMT)
  - Linguistically informed grammars extracted with the aid of a target-side parse tree

S+. → PRP+VBZ(1) உறுதி செய்கிறது .(2), PRP+VBZ(1) confirmed .(2)

- SAMT grammars are particularly well-motivated
  - Syntax should help describe high-level SOV → SVO reordering
  - Previously well-attested for Urdu (Baker et al., SCALE 2009)

Q2: Do linguistic annotations help?

— BLEU-4 scores —

Language
Bengali
Hindi
Malayalam
Tamil
Telugu
Urdu

scores are the mean of three MERT runs

Q2: Do linguistic annotations help?

— BLEU-4 scores —

Language	Hiero
Bengali	12.72
Hindi	15.53
Malayalam	13.72
Tamil	9.81
Telugu	12.46
Urdu	19.53

scores are the mean of three MERT runs

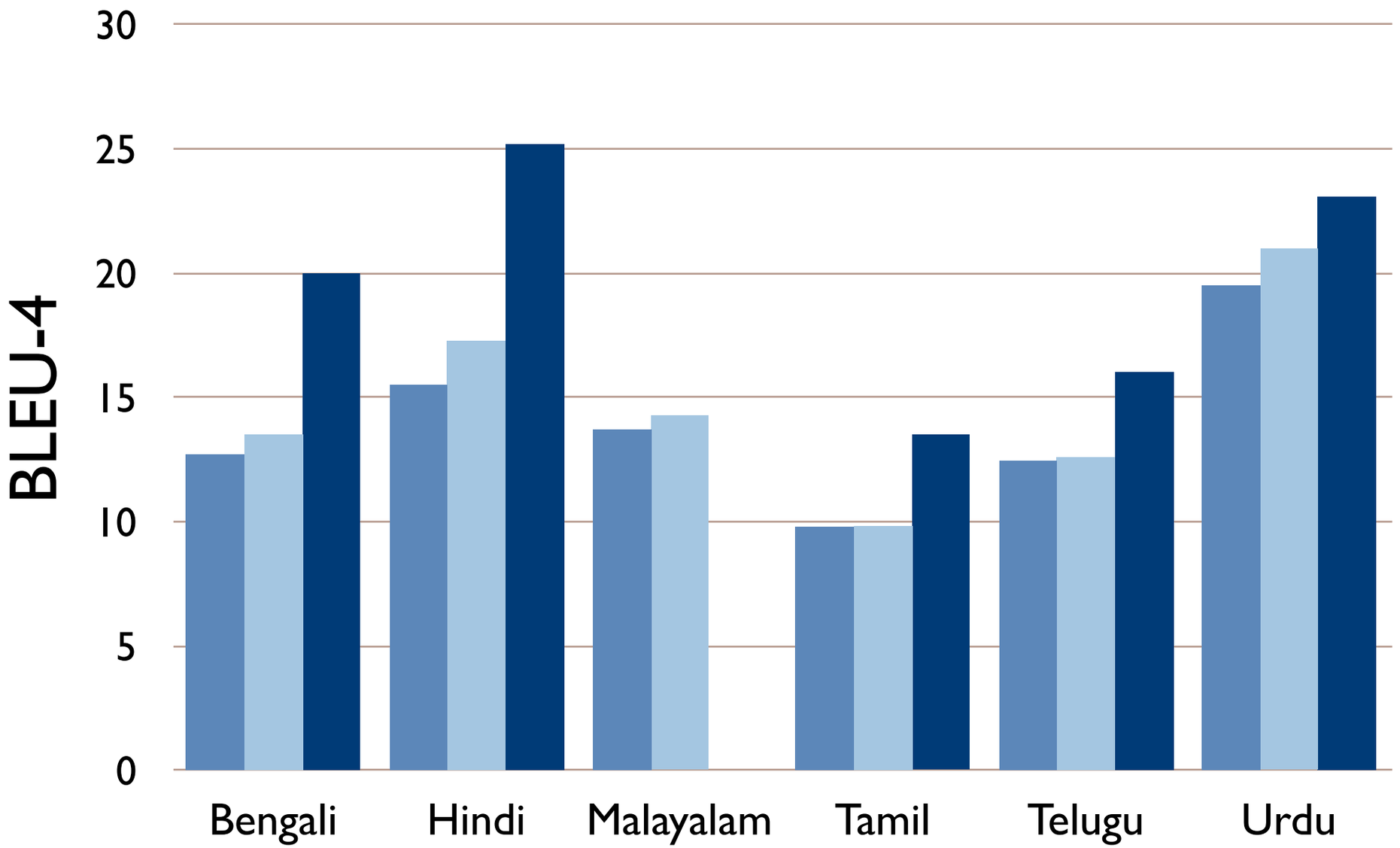
Q2: Do linguistic annotations help?

— BLEU-4 scores —

Language	Hiero	SAMT	Difference
Bengali	12.72	**13.53**	+0.81
Hindi	15.53	**17.29**	+1.76
Malayalam	13.72	**14.28**	+0.56
Tamil	9.81	9.85	+0.04
Telugu	12.46	12.61	+0.15
Urdu	19.53	**20.99**	+1.46

scores are the mean of three MERT runs

Q2: Do linguistic annotations help?



scores are the mean of three MERT runs

Q3: Does translation quality matter?

- Recall that we have four redundant translations for each Indian language sentence, along with independently-obtained votes about which is best
- We trained models on a quarter of the data
  1. selected randomly
  2. selected by plurality (breaking ties randomly)
- And tested on the same test sets

Q3: Does text quality matter?

	Hiero	
Language	random	best
Bengali	9.43	9.29
Hindi	11.74	12.18
Malayalam	-	-
Tamil	7.73	7.48
Telugu	10.49	10.61
Urdu	13.51	**14.26**

scores are the mean of three MERT runs

Q3: Does text quality matter?

	Hiero		SAMT
Language	random	best	random
Bengali	9.43	9.29	9.65
Hindi	11.74	12.18	12.61
Malayalam	-	-	-
Tamil	7.73	7.48	7.88
Telugu	10.49	10.61	10.75
Urdu	13.51	**14.26**	14.63

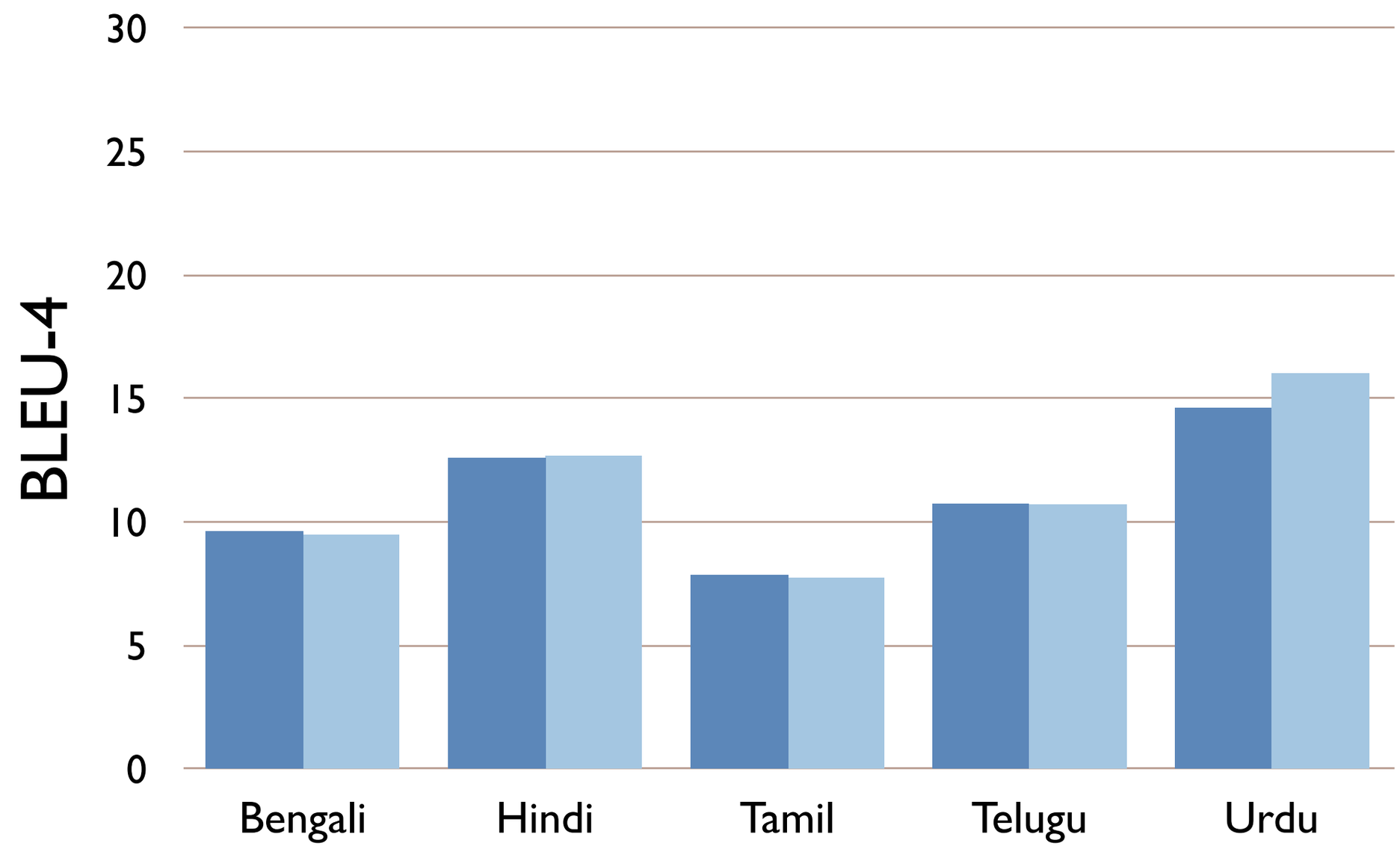
scores are the mean of three MERT runs

Q3: Does text quality matter?

	Hiero		SAMT	
Language	random	best	random	best
Bengali	9.43	9.29	9.65	9.50
Hindi	11.74	12.18	12.61	12.69
Malayalam	-	-	-	-
Tamil	7.73	7.48	7.88	7.76
Telugu	10.49	10.61	10.75	10.72
Urdu	13.51	**14.26**	14.63	**16.03**

scores are the mean of three MERT runs

Q3: Does text quality matter?



random
highest-voted

scores are the mean of three MERT runs

Future directions

- Morphology
  - We took the word segmentations as given, yet we know these languages to be highly agglutinative
  - Better segmentation should help at all stages, from alignment to decoding

Future directions

Tamil-English dataset

Urdu-English dataset

misspelling	count
japenese	91
japans	40
japenes	9
japenies	3
japaeneses	3
japeneese	1
japense	1

Future directions

- Text normalization: standardizing orthography would help immensely

Tamil-English dataset

Urdu-English dataset

misspelling	count
japenese	91
japans	40
japenes	9
japenies	3
japaeneses	3
japeneese	1
japense	1

Future directions

- Text normalization: standardizing orthography would help immensely

Tamil-English dataset

இலங்கையில் சோழர் ஆட்சி
in srilanka solar government
chola rule in sri lanka .
in srilanka chozhas ruled
chola reign in sri lanka

Urdu-English dataset

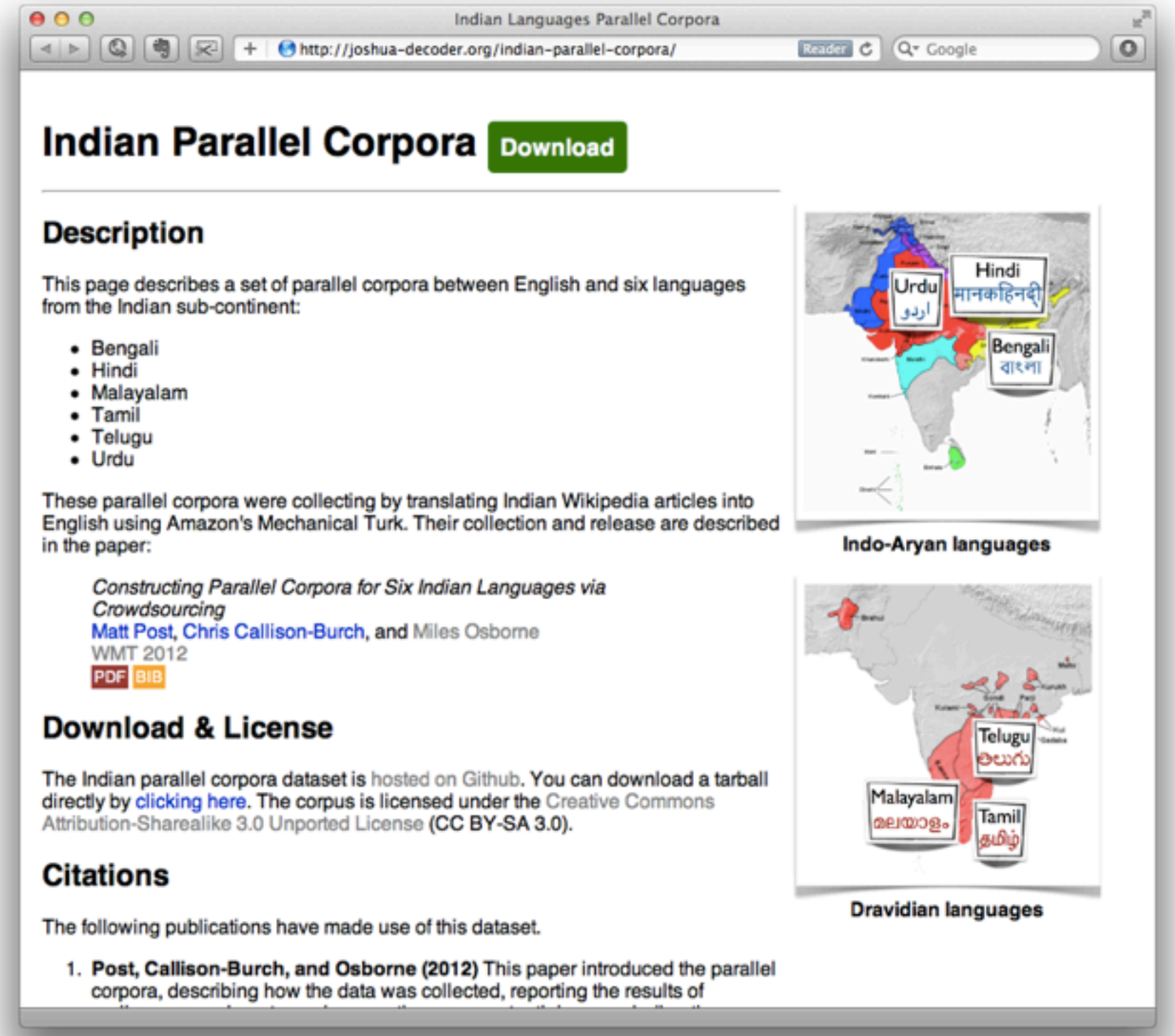
misspelling	count
japenese	91
japans	40
japenes	9
japenies	3
japaeneses	3
japeneese	1
japense	1

Summary

joshua-decoder.org/indian-parallel-corpora

- A suite of six low-resource, head-final, morphologically rich languages from the Indian subcontinent
- Provided data splits for comparisons
- Ideas can be tested in an afternoon on a variety of languages
- We suggest future work in the areas of morphology, normalization, and domain adaptation
- The website will track uses of the data as well as the best test-set scores

joshua-decoder.org/indian-parallel-corpora



Thanks

- Support: Google, Microsoft, EuroMatrixPlus, DARPA
- Lexi Birch
- Ghouse Ismail