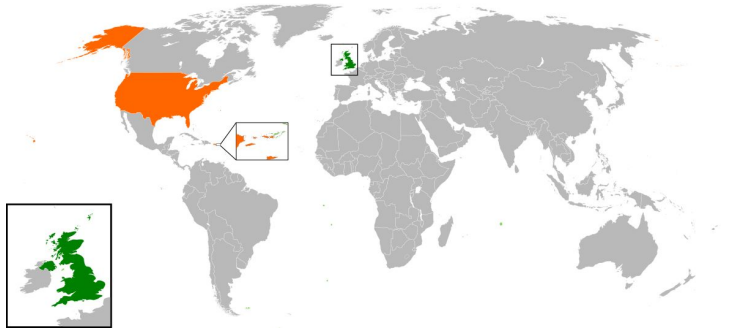
The Non-native Speaker Aspect : Indian English in Social Media

Rupak Sarkar, Sayantan Mahinder, Ashiqur R. Khudabukhsh

Indian English

- Pluricentric Language
- Variant Spoken in Indian Diaspora
- Largest second language variety of English



Data Set

- Social Media Comments from YouTube
  - India : 14 National News Channels
  - US : CNN, FOX, MSNBC
  - UK : BBC, Channel 4
- Articles
  - Highly circulated news outlets
  - India : Quint, Hindustan Times
  - US : Washington Post, New York Times

10k subset

$\mathcal{D}^{sm}_{en-in}$

$\mathcal{D}^{sm}_{en-us}$

$\mathcal{D}^{sm}_{en-gb}$

$\mathcal{D}^{na}_{en-in}$

$\mathcal{D}^{na}_{en-us}$

How is this different

- Wide range of linguistic proficiency
- Traditional Measures -
  - Frequent spelling errors
  - Limited variation in verb forms
  - Smaller Sentences
  - Article, preposition, pronoun usage

these people will be never change

please do not telicaste this idiot he will make this crisis severe as he did in delhi violance

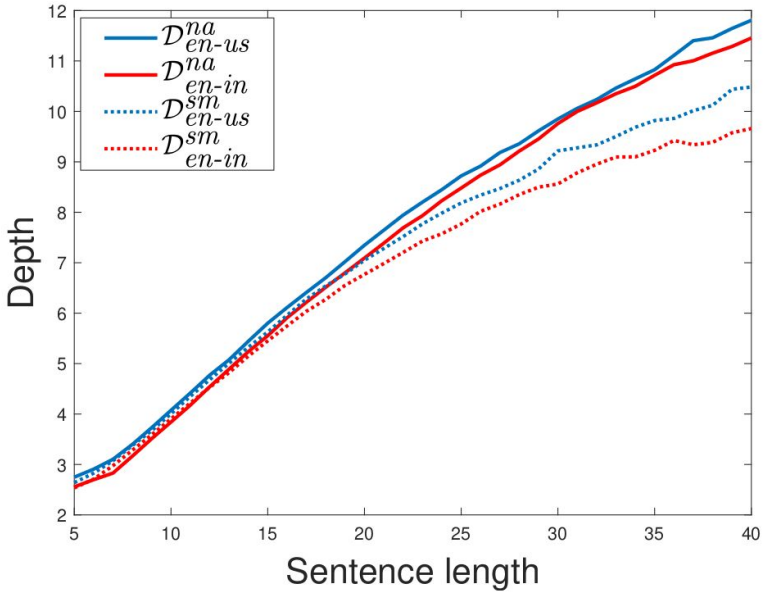
I am live in Assam.

forgotten to mention about bank employee whose who are working with high risk factors of coronavirus

Constituency Parsing

- Valid Sentence
- Parser Tree depth

Measure	$\mathcal{D}^{na}_{en\text{-}in}$	$\mathcal{D}^{na}_{en\text{-}us}$	$\mathcal{D}^{sm}_{en\text{-}in}$	$\mathcal{D}^{sm}_{en\text{-}us}$
Valid sentences	96.93	96.61	83.88	88.30



Cloze Test as a Linguistic Tool

- Cloze Test : Fill in the blank
- Have been used in -
  - Relation Extraction
  - Political Insight Mining
- We propose:
  - Linguistic quality assessment
  - Correct PoS Among top 10 predictions

	$\mathcal{D}^{na}_{en\text{-}in}$	$\mathcal{D}^{na}_{en\text{-}us}$	$\mathcal{D}^{sm}_{en\text{-}in}$	$\mathcal{D}^{sm}_{en\text{-}us}$
Overall	84.27	83.68	66.56	71.72
VERB	90.17	89.72	79.47	82.76
NOUN	86.20	85.95	62.03	67.96
ADP	89.78	89.24	75.96	75.88
ADJ	68.88	70.54	48.55	61.06
ADV	74.40	73.06	47.09	58.01

Measure	$\mathcal{D}^{na}_{en\text{-}in}$	$\mathcal{D}^{na}_{en\text{-}us}$	$\mathcal{D}^{sm}_{en\text{-}in}$	$\mathcal{D}^{sm}_{en\text{-}us}$
p@1	53.53	55.53	33.49	42.31
p@5	75.69	77.30	55.91	65.07
p@10	81.03	82.93	62.69	71.36

Conclusions

- College educated formal English
  - 29 % youth (18-24) go to college
  - USA: 88 %
  - UK: 61 %

[Source: UNESCO Institute for Statistics, 2018]

- As spoken in social media

Thank You !