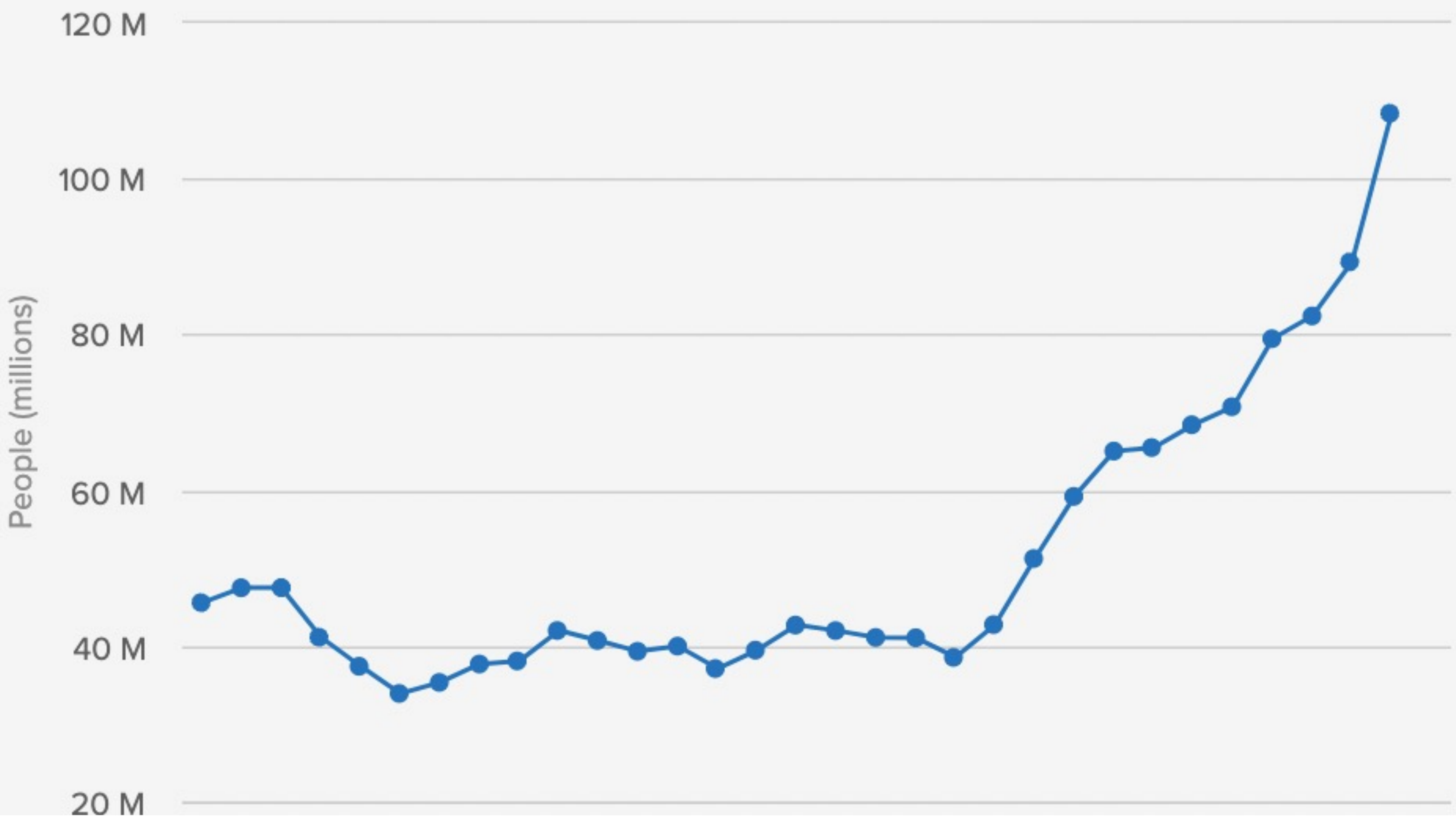
All Translation Tools Are Not Equal: Investigating the Quality of Language Translation for Forced Migration

Ameeta Agrawal[1], Lisa Singh[2], Elizabeth Jacobs[3], Yaguang Liu[2], Gwyneth Dunlevy[2], Rhitabrat Pokharel[1], Varun Uppala[2]

[1]Portland State University
[2]Georgetown University
[3]Max Planck Institute for Demography Research

108.4 million people worldwide were forcibly displaced

At the end of 2022 as a result of persecution, conflict, violence, human rights violations or events seriously disturbing public order.



Emergencies | News and stories | Get involved

Ongoing emergencies >

Other countries

- Afghanistan emergency
- DR Congo emergency
- Rohingya emergency
- South Sudan emergency
- Syria emergency
- Ukraine emergency
- Venezuela situation

Official	Dari Persian, Pashto
Regional	Uzbek, Turkmen, Balochi, Pashayi, Nuristani
Minority	Arabic, Gujari, Urdu, Kyrgyz, Tajik, Sindhi

Official	French
National	Kituba, Lingala, Swahili and Tshiluba
Indigenous	More than 200

Official	English
National	Bari · Dinka · Luo · Murle · Nuer · Zande

Official	Modern Standard Arabic
Vernacular	Levantine Arabic and Mesopotamian Arabic
Minority	Kurdish, Turkish, Neo-Aramaic, Circassian, Chechen, Armenian, Greek

Official	Spanish
Indigenous	Languages of the Arawakan, Arutani-Sape, Cariban, Chibchan, Guahiban, Jirajaran, Timotean families

- Urgent need to process data at scale
- Organic data sources (social media, newspapers) offer insights in near real time
- But need to contextualize drivers of movement for different languages and locations
- Machine translation (MT) can help but less studied in the context of forced migration

We investigate off-the-shelf machine translation tools for forced migration

We translate words and phrases across high- and low-resource languages

We analyze trade-offs between cost and performance, and provide recommendations for data scientists and migration researchers

Translation Service	Cost
Microsoft	\$10/million characters
Google	\$20/million characters
DeepL	\$29/month
GPT-3	\$0.5/million characters
MUSE	Public Use
Opus-MT	Open Source
OpenNMT	Open Source
Human Translation	~ \$80/document

- Costs of MT services vary but all significantly lower than human translation costs
- Machine translation also generally faster than human translation

Factor	Example words/phrases
physical	ambush, armed forces, army, bomb, death, weapons, arms
environmental	earthquake, flood, hurricane
political	activist, protest, corruption, parliament
economic	currency, debt, unemployment, inflation, black market, sanction
health	COVID, dengue, malaria, AIDS, polio, Zika
food	crop, famine, drought, drinking water, food shortage, harvest, wheat

- Data consists of words/phrases representing macro-level drivers of forced migration
- Manual translations across four languages (English, Spanish, Arabic, Portuguese)

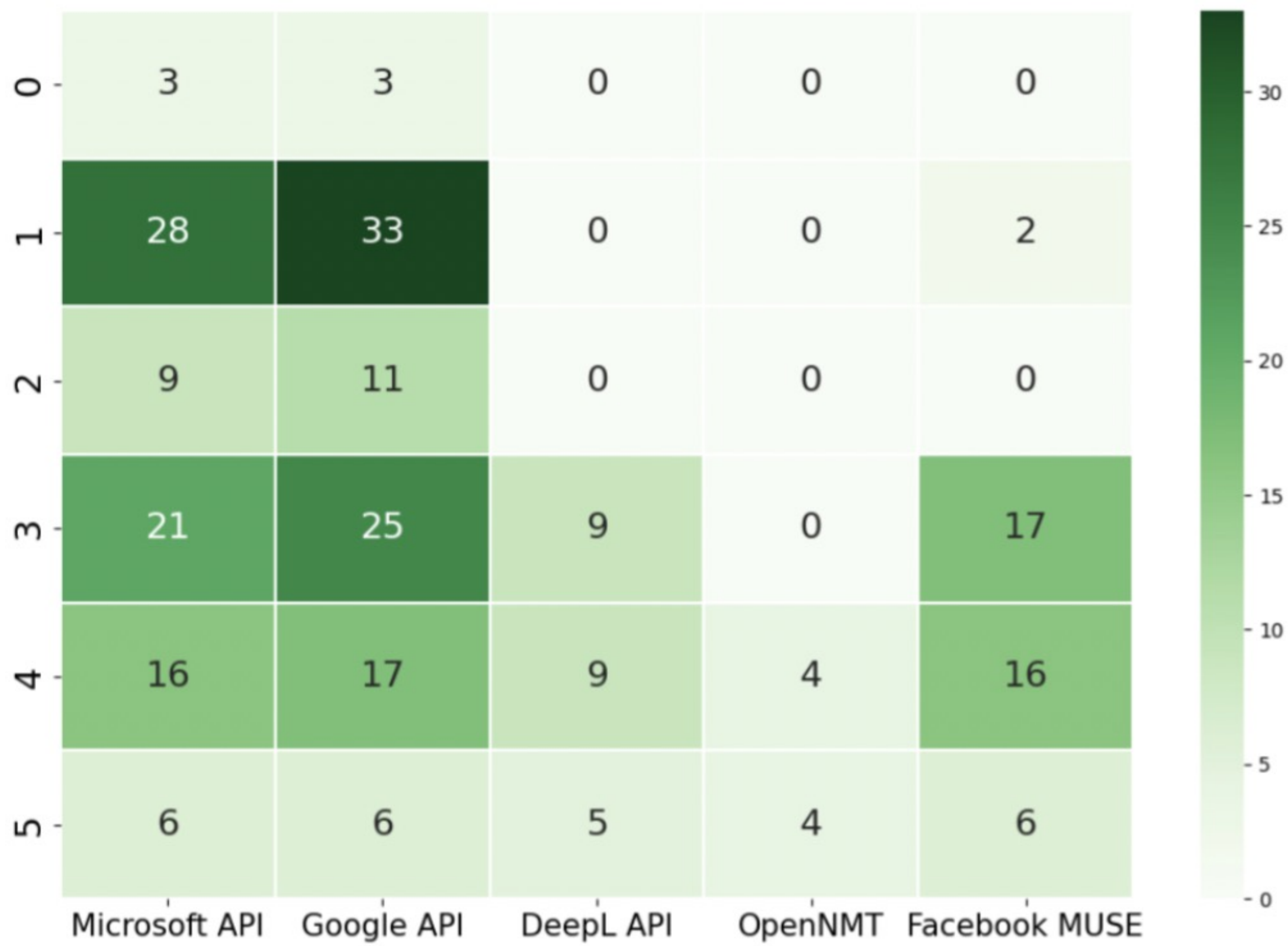
English	Spanish	Arabic	Portuguese
activist	activista	ناشط	ativisto
crime	crimen	جريمة	crime
dialogue	diálogo	حوار	discussão
government	gobierno	حكومة	governo
law	ley	قانون	lei
oppression	opresión	قمع	oppressão
political	político	سياسي	político
voter	votante	ناخب	eleitor

Feature	Microsoft	Google	DeepL	OpenNMT	MUSE	GPT-3	Opus-MT
Has PoS Capabilities?	Yes, but not for all	No	No	Yes	unknown	No	No
Returns Multiple Words with Confidence?	Yes	No	Yes	-	-	Yes	No
Usability:	Small amount of customization	Easy to use API	Easy to use API	Small amount of customization	Small amount of customization	Easy to use API	Easy to use API
Number of Languages Supported:	103	109	24	8 (train others)	100	unknown	unknown

- Seven off-the-shelf machine translation services

“high”-resource	**“medium”-resource**	**“low”-resource**
English (5)	Ukrainian (3)	Somali (1)
Arabic (5)		Haitian Creole (0)
Spanish (5)		Dari (0)
Portuguese (4)		

- Eight languages from different regions of recent insecurity



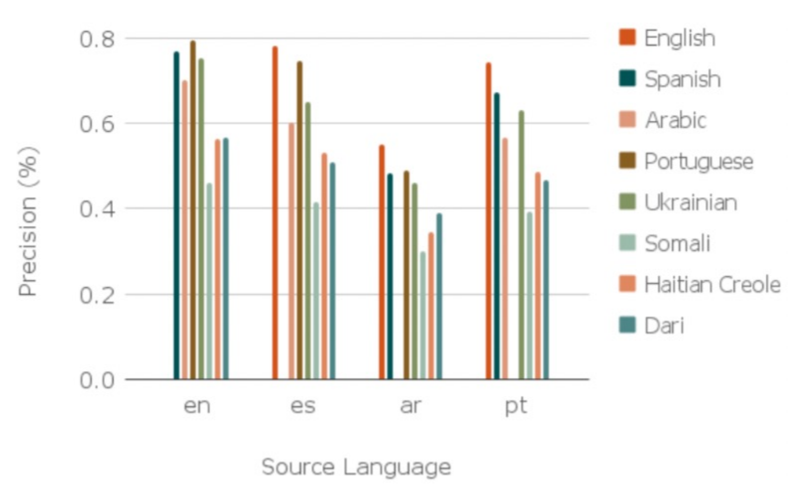
- Very few languages are available through OpenNMT
- DeepL and Facebook MUSE support more high-resource languages than low-resource ones
- Microsoft and Google provide reasonable coverage across the spectrum, only services currently available for the lowest-resource language category ('0')

Evaluation: one-way vs. round-trip

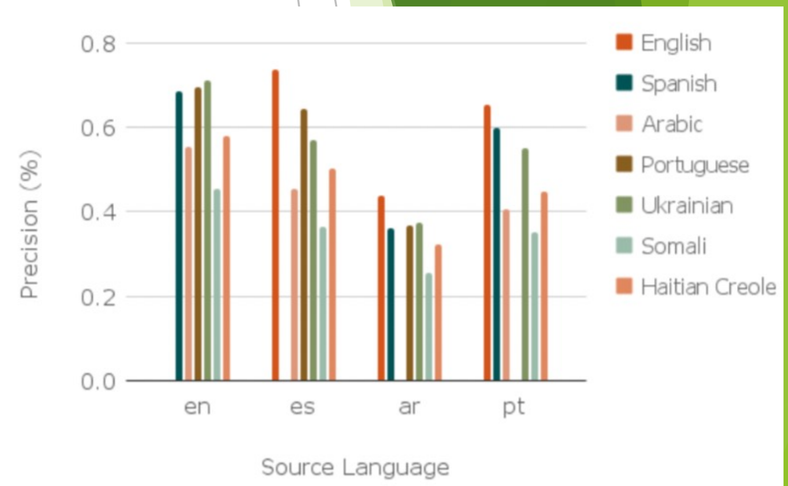
- When parallel translated data available, **one-way evaluation** (English → Spanish)
- But when parallel translations not available? **Round-trip evaluation** (English → Spanish → English)
- Both show similar *trends*

Source-Target	Microsoft		Google		DeepL		GPT-3		Opus-MT	
	One-way	Round-trip	One-way	Round-trip	One-way	Round-trip	One-way	Round-trip	One-way	Round-trip
en-sp	0.78	0.76	**0.80**	0.68	0.75	0.64	0.72	0.58	0.74	0.75
en-ar	0.54	0.70	**0.57**	0.55	-	-	0.26	0.21	0.22	0.18
en-pt	**0.75**	0.79	0.71	0.69	0.69	0.64	0.69	0.64	0.70	-

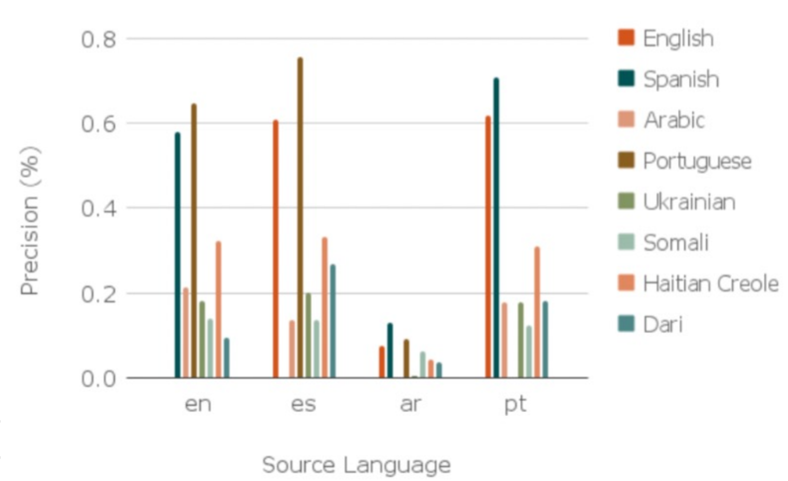
- Microsoft MT and Google MT tools show similar trends, with Microsoft yielding slightly better performance (~80%) than Google (~70%)
- GPT-3 shows distinct gap between high- and low-resource languages (~70% to ~30%)
- Takeaway: slightly costlier options such as Microsoft/Google do bring additional advantages over the less expensive GPT-3



(a) Microsoft MT system

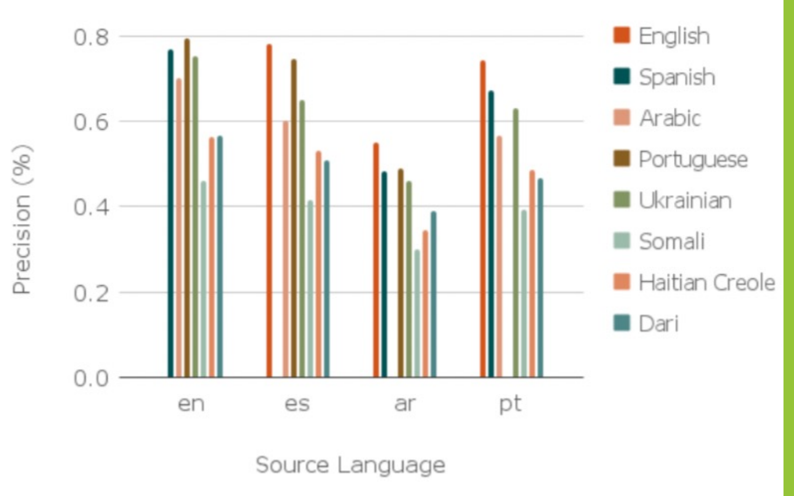


(b) Google MT system

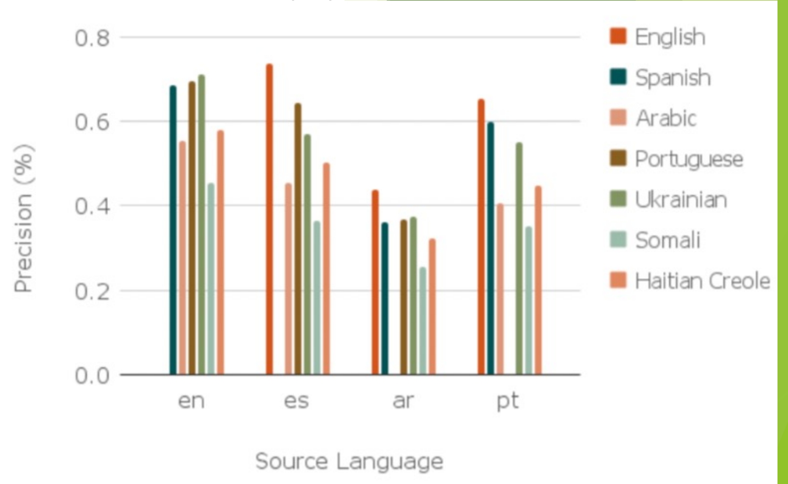


(c) GPT-3 model

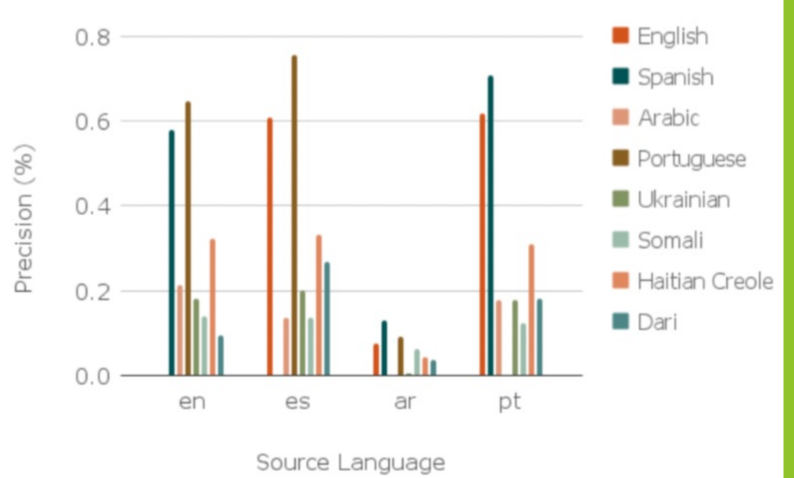
- English seems to be the best 'source' language
- This is encouraging because having English data can help scale up multilingual translations of forced migration organic data
- In comparing four higher-resource languages (English (5), Spanish (5), Arabic (5), Portuguese (4)), Arabic does poorly even though higher than Portuguese
- Similarly, in comparing three lower-resource languages (Somali (1), Haitian Creole (0), Dari (0)), it is Somali that does the poorest



(a) Microsoft MT system



(b) Google MT system

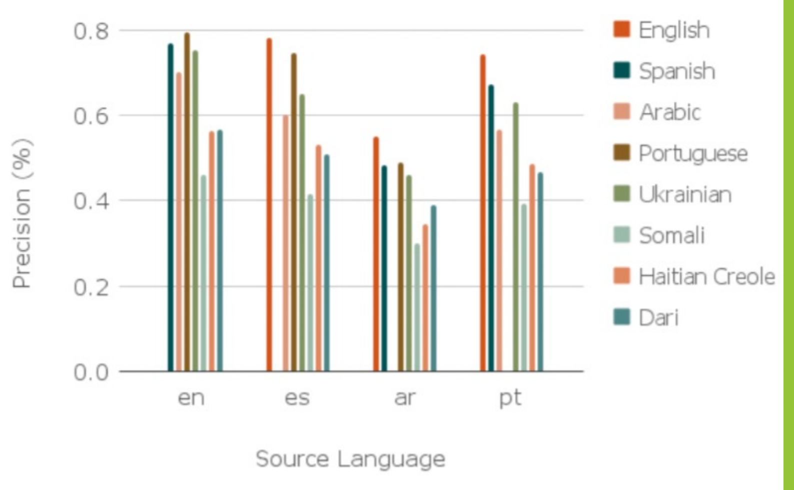


(c) GPT-3 model

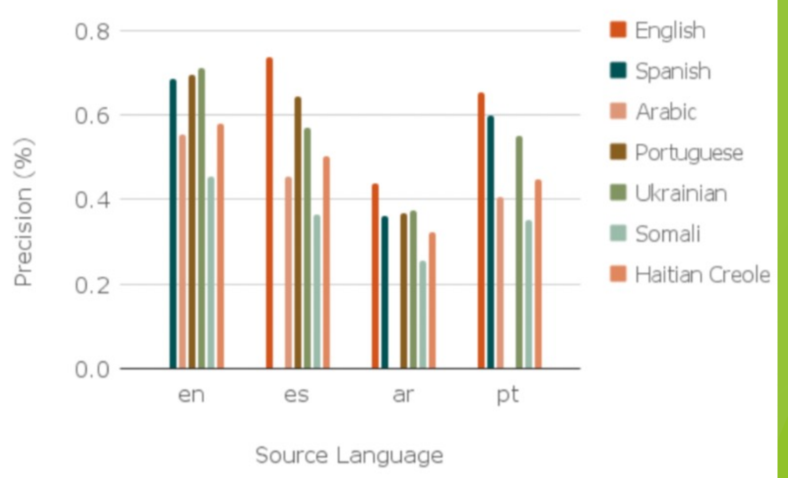
▶ One possible explanation:

- Both Arabic and Somali belong to the Afro-Asiatic language family
- The rest (except Haitian Creole) all belong to the Indo-European language family

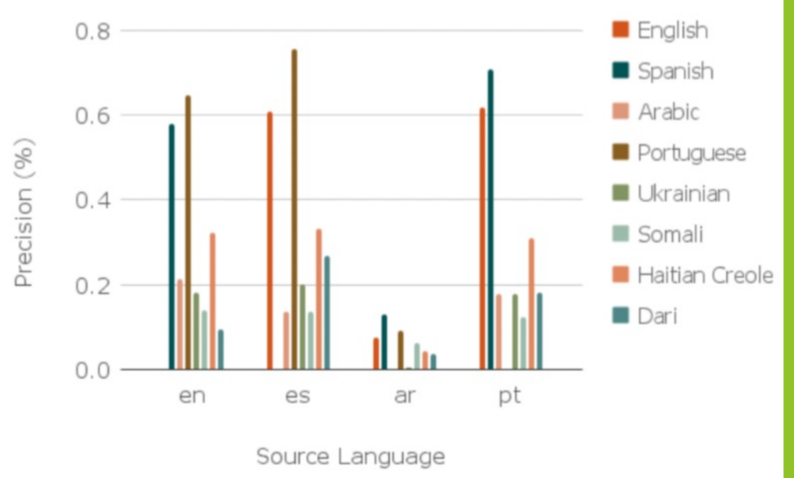
▶ In short, MT tools work better not only for higher-resource languages, but also for languages from certain language families



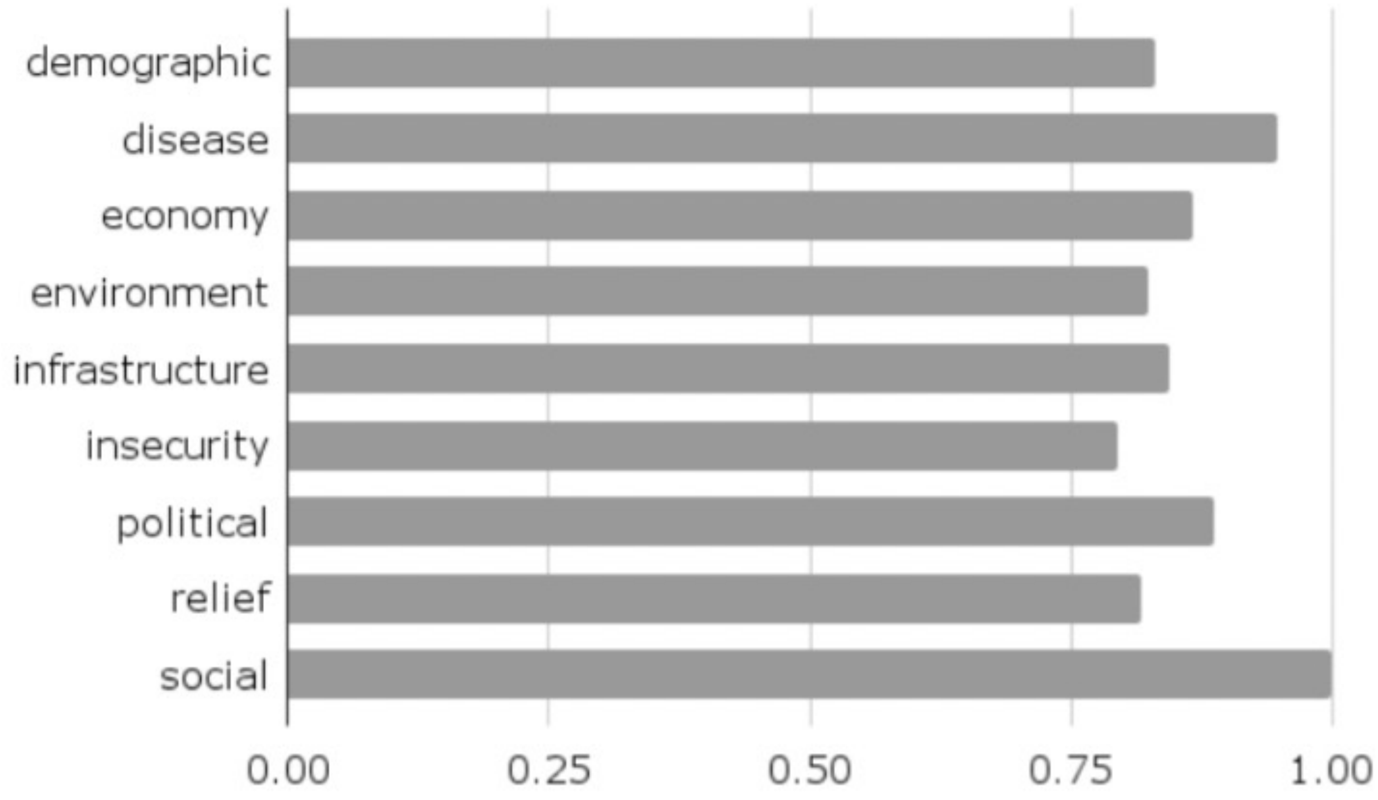
(a) Microsoft MT system



(b) Google MT system



(c) GPT-3 model



- For drivers of forced migration, precision ranges from 0.8 to 1.0
  - “insecurity” is the lowest, perhaps because it has a larger set of words than the other topics, and the words may have more synonyms
  - “social” is the highest
- Different translation performance across topics

Factor	Microsoft	Google
demographic	English Arabic English	English Ukrainian English
disease	English Somali English	Spanish English Spanish
economy	Portuguese English Portuguese	English Ukrainian English
environment	English Spanish English	English Ukrainian English
infrastructure	English Spanish English	English Ukrainian English
insecurity	English Portuguese English	English Portuguese English
political	Spanish English Spanish	Spanish English Spanish
relief	English Portuguese English	English Spanish English
social	Spanish Portuguese Spanish	English Portuguese English

- While English, Spanish, Portuguese, Arabic generally do well, two lower resource languages emerged as surprising target languages
  - For the factor "disease", Somali as a target language obtains highest accuracy
  - Ukrainian also obtains good translations, possibly due to current situation
- We hypothesize that languages that use factor words in a similar context are likely to have a higher match rate
- Overall, Microsoft Translator emerged as a stronger MT tool than Google Translate