Geographically-Informed Language Identification

Jonathan Dunn

jedunn@illinois.edu

www.jdunn.name

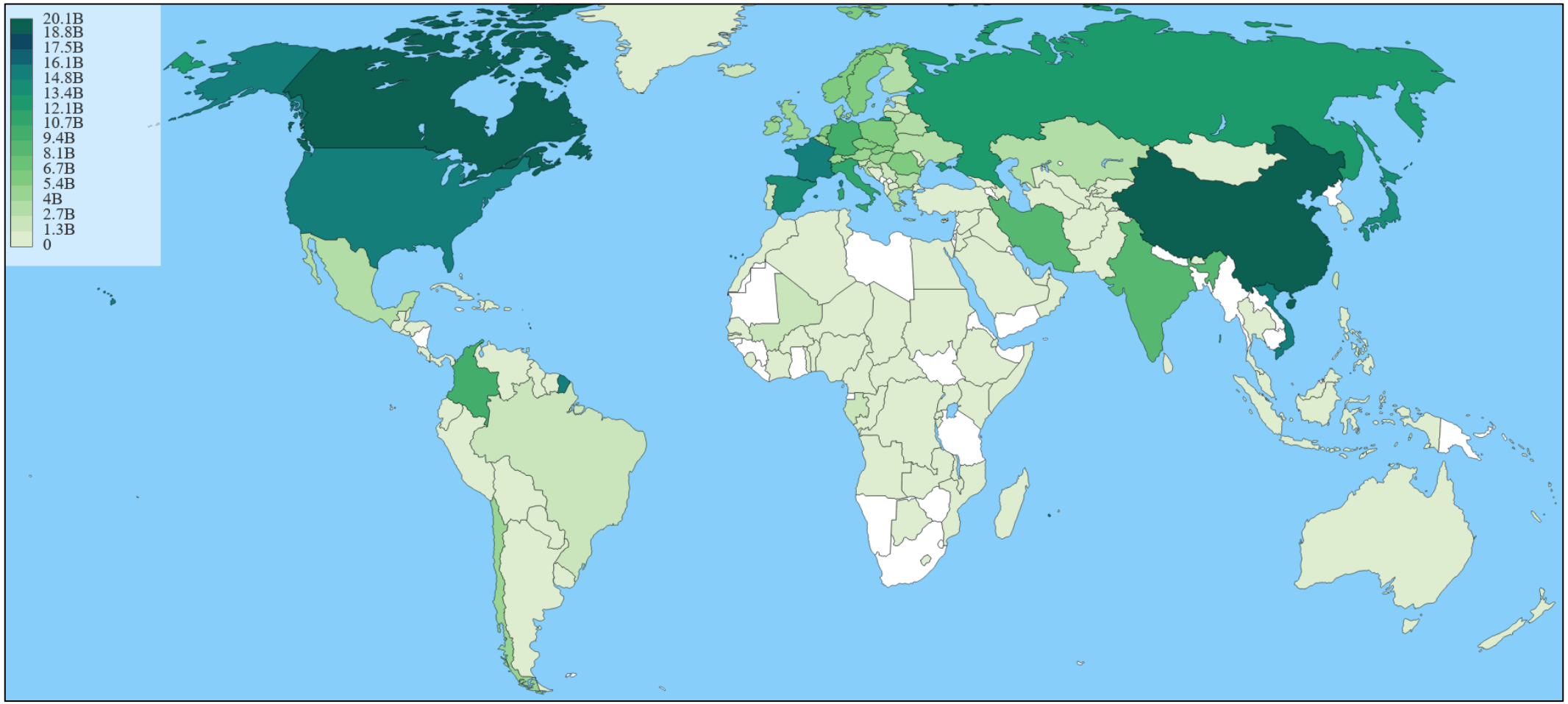
Lane Edwards-Brown

laneedwardsbrown@gmail.com

Do geographic priors improve LID for lower-resource languages?

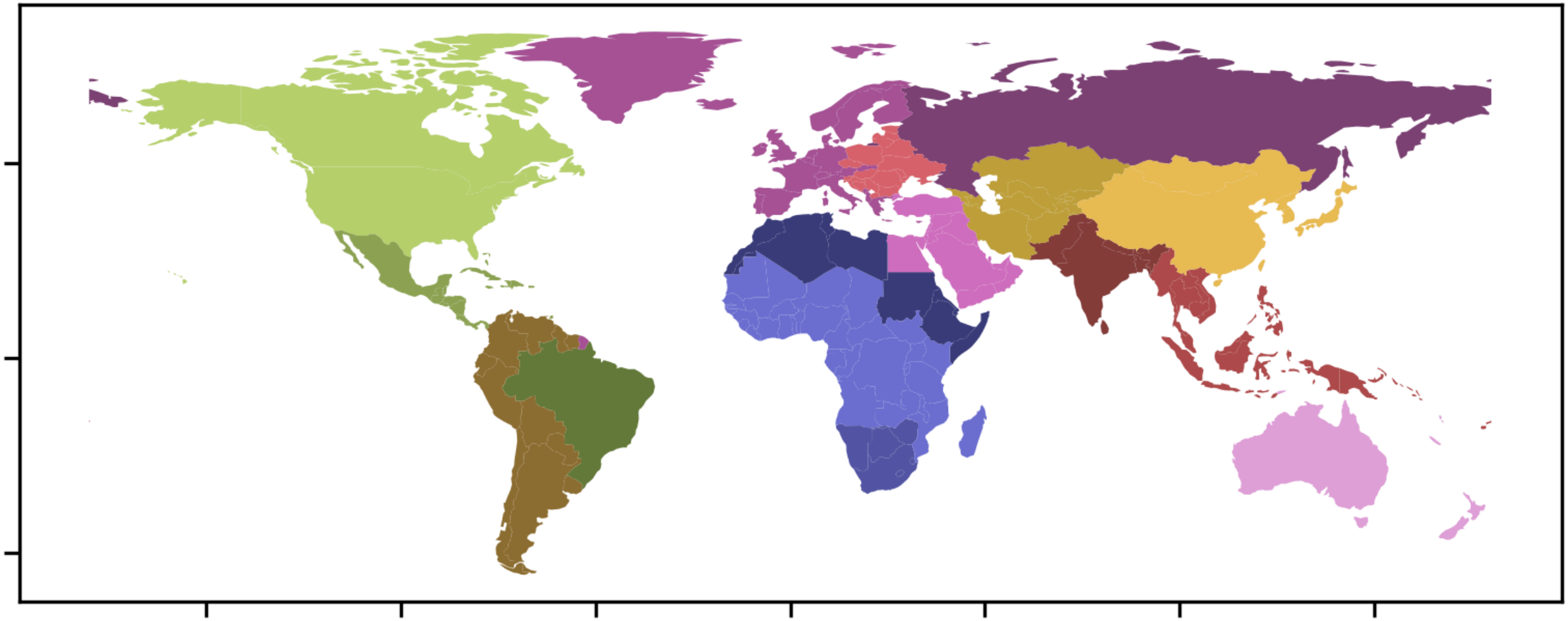
Goal: Enhanced LID for geo-referenced corpora

www.earthLings.io



Dunn, J. (2020). "Mapping Languages: The Corpus of Global Language Use." *Language Resources and Evaluation*.

16 Region-Specific LID Models



Data from LID Sources

Corpus	N. Langs
Bible Translations (Brown, 2014)	614
Global Voices News (Tiedemann, 2012)	41
JW 300 (Agić and Vulić, 2019)	380
Open Subtitles (Lison and Tiedemann, 2016)	62
QCRI Educational Domain (Tiedemann, 2012)	42
Tatoeba Sentences (Tiedemann, 2012)	309
Wikipedia Articles TensorFlow DataSets	280

Table 1: Primary Sources of Training Data

Language Data from Glottolog

International Languages

Language	Abbrv.	Language	Abbrv.
Amharic	amh	Korean	kor
Arabic	ara	Mandarin	zho
Bengali	ben	Marathi	mar
English	eng	Polish	pol
Farsi	fas	Portuguese	por
French	fra	Punjabi	pan
German	deu	Russian	rus
Gujarati	guj	Spanish	spa
Hausa	hau	Swahili	swa
Hindi	hin	Tagalog	tgl
Indonesian	ind	Tamil	tam
Italian	ita	Telugu	tel
Japanese	jpn	Thai	tha
Javanese	jav	Turkish	tur
Kannada	kan	Urdu	urd
		Vietnamese	vie

Table 2: International languages which are included in each regional model.

Results

Region	N. Langs	F-Score *Geo*	F-Score *Baseline*	Test Samples
Africa, North	44	**0.990**	0.886	621k
Africa, Southern	58	**0.982**	0.903	1,053k
Africa, Sub	166	**0.980**	0.947	1,931k
America, Central	188	**0.991**	0.961	2,965k
America, North	68	**0.993**	0.902	1,017k
America, South	129	**0.995**	0.960	4,612k
America, Brazil	88	**0.996**	0.945	818k
Asia, Central	54	**0.988**	0.906	777k
Asia, East	46	**0.990**	0.892	1,131k
Asia, South	60	**0.986**	0.914	979k
Asia, Southeast	325	**0.990**	0.972	3,992k
Europe, East	65	**0.978**	0.908	3,132k
Europe, West	108	**0.967**	0.921	5,473k
Europe, Russia	65	**0.984**	0.914	1,098k
Middle East	53	**0.988**	0.904	801k
Oceania	49	**0.984**	0.890	745k

Geo = GeoLID in current region

Baseline = Same architecture, no geo

Models = fastText-based classifier

Results for only local languages

Region	N. Langs	F-Score *Geo*	F-Score *Baseline*	Test Samples
Africa, North	13	**0.986**	0.972	75k
Africa, Southern	27	**0.976**	0.964	442k
Africa, Sub	135	**0.979**	0.970	1,313k
America, Central	157	**0.991**	0.983	2,339k
American, North	37	**0.994**	0.945	363k
America, South	98	**0.997**	0.994	3,747k
America, Brazil	57	**0.999**	0.996	287k
Asia, Central	23	0.984	0.983	237k
Asia, East	15	0.993	0.980	410k
Asia, South	29	0.983	0.983	340k
Asia, Southeast	294	**0.991**	0.985	3,291k
Europe, East	34	**0.975**	0.961	2,363k
Europe, West	77	**0.963**	0.950	4,565k
Europe, Russia	34	**0.979**	0.972	525k
Middle East	22	**0.986**	0.981	255k
Oceania	18	**0.978**	0.959	168k

Geo = GeoLID in current region

Baseline = Same architecture, no geo

Models = fastText-based classifier

OpenLID as extra test set

Region	N.	Geo	Baseline
Africa, North	35	**0.990**	0.969
Africa, Southern	44	**0.988**	0.968
Africa, Sub	65	**0.983**	0.968
America, Central	36	**0.987**	0.970
America, North	31	**0.986**	0.966
America, South	35	**0.990**	0.969
America, Brazil	33	**0.993**	0.967
Asia, Central	43	**0.991**	0.974
Asia, East	38	0.964	0.945
Asia, South	42	0.947	0.925
Asia, Southeast	48	0.936	0.926
Europe, East	53	**0.992**	0.978
Europe, West	67	**0.991**	0.982
Europe, Russia	42	**0.992**	0.974
Middle East	41	**0.994**	0.973
Oceania	35	0.976	0.960

Geo = GeoLID in current region

Baseline = Same architecture, no geo

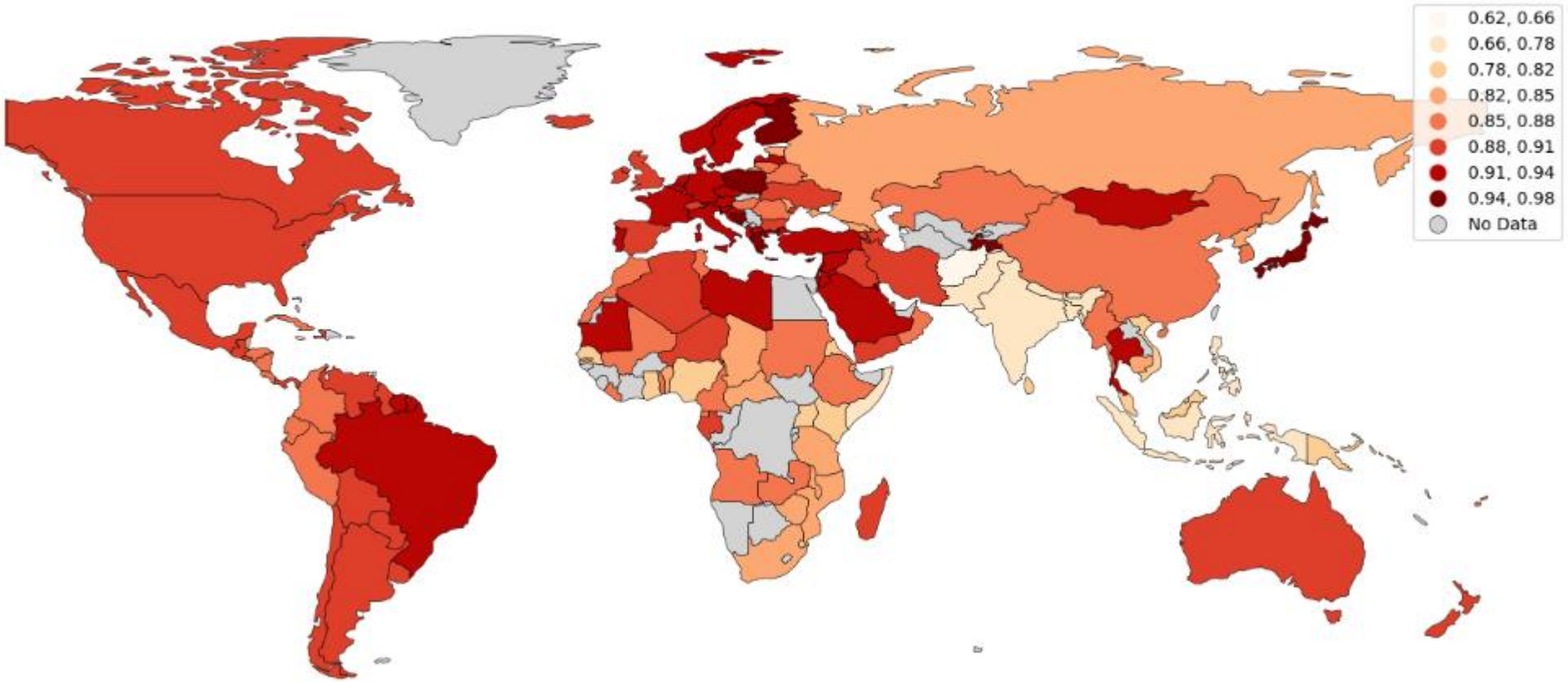
Models = fastText-based classifier

Are some languages poorly represented?

Model	Lang.	Geo	Non-Geo
Africa, Southern	sot	0.88	0.86
Africa, Sub	bam	0.69	0.53
Africa, Sub	fuh	0.69	0.58
Africa, Sub	ffm	0.70	0.59
Africa, Sub	plt	0.86	0.85
Africa, Sub	tum	0.86	0.83
Africa, Sub	eng	0.87	0.34
Africa, Sub	run	0.89	0.89
America, Central	kek	0.87	0.85
Asia, South	dty	0.79	0.80
Asia, Southeast	pam	0.81	0.76
Asia, Southeast	cbk	0.85	0.81
Asia, Southeast	spa	0.85	0.41
Asia, Southeast	tet	0.85	0.78
Asia, Southeast	bjn	0.86	0.82
Asia, Southeast	gor	0.87	0.84
Europe, East	eng	0.87	0.34
Europe, East	rmy	0.87	0.75
Europe, Russia	mdf	0.89	0.87
Europe, West	eng	0.86	0.34
Europe, West	ile	0.87	0.88
Oceania	cha	0.86	0.78
Oceania	bjn	0.88	0.82

Table 6: Complete list of all languages with an f-score below 0.90 in the geographic models.

Where do Geo and Non-Geo Models Disagree?

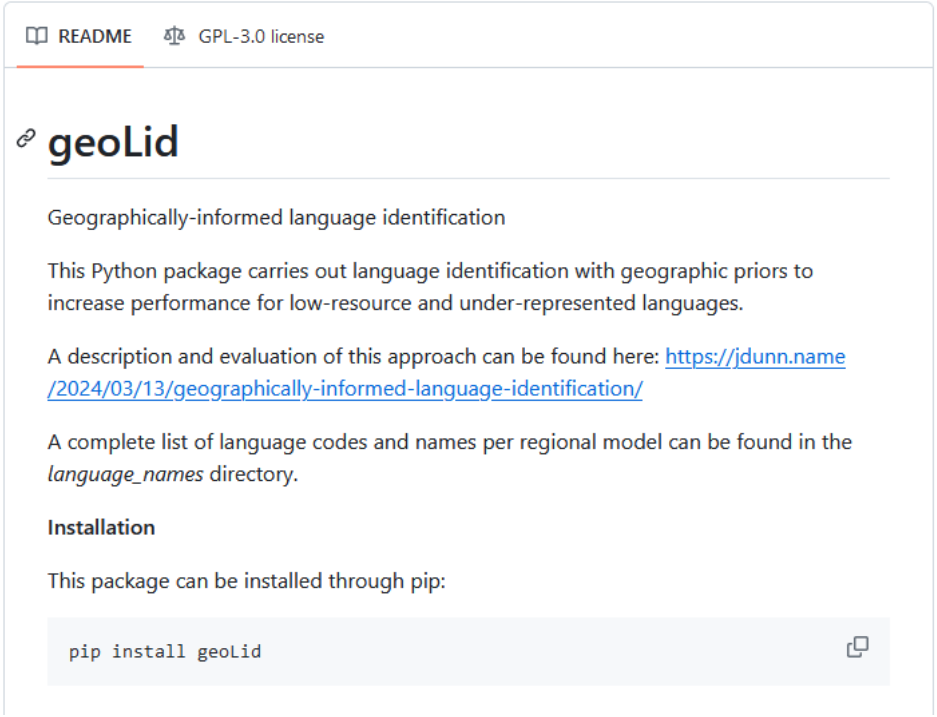


Where do Geo and Non-Geo Models Disagree?

Region	N.	Agree	Samples
Africa, North	9	84.93%	10.21 mil
Africa, Southern	3	83.08%	3.77 mil
Africa, Sub	26	84.94%	26.71mil
America, Central	14	87.28%	16.71 mil
America, North	2	89.62%	2.97 mil
America, South	11	88.85%	14.40 mil
America, Brazil	1	92.41%	1.53 mil
Asia, Central	7	85.52%	9.28 mil
Asia, East	5	89.31%	5.96 mil
Asia, South	7	74.20%	8.42 mil
Asia, Southeast	14	83.86%	13.39 mil
Europe, East	15	89.88%	21.20 mil
Europe, West	23	91.03%	32.19 mil
Europe, Russia,	1	83.03%	1.50 mil
Middle East	12	91.80%	16.33 mil
Oceania	7	85.56%	4.75 mil
Total	**157**	**87%**	**189 mil**

Conclusions

Geographic priors allow better language ID, when available



README GPL-3.0 license

geoLid

Geographically-informed language identification

This Python package carries out language identification with geographic priors to increase performance for low-resource and under-represented languages.

A description and evaluation of this approach can be found here: https://jdunn.name/2024/03/13/geographically-informed-language-identification/

A complete list of language codes and names per regional model can be found in the *language_names* directory.

Installation

This package can be installed through pip:

```
pip install geoLid
```

github.com/jonathandunn/geoLid

Thanks!

- Jonathan Dunn
- jedunn@illinois.edu
- www.jdunn.name

Lane Edwards-Brown

laneedwardsbrown@gmail.com