Low-Resource NLP

601.764

3/14/23

What is Low-Resource?

- NLU
- ASR
- QA
- MT
- IR

- Everything?
- < 50 Hours?
- ?
- < 100 K Sentences … ~ 2 M Words
- Industry thinks 1-2 orders of magnitude more
- ?

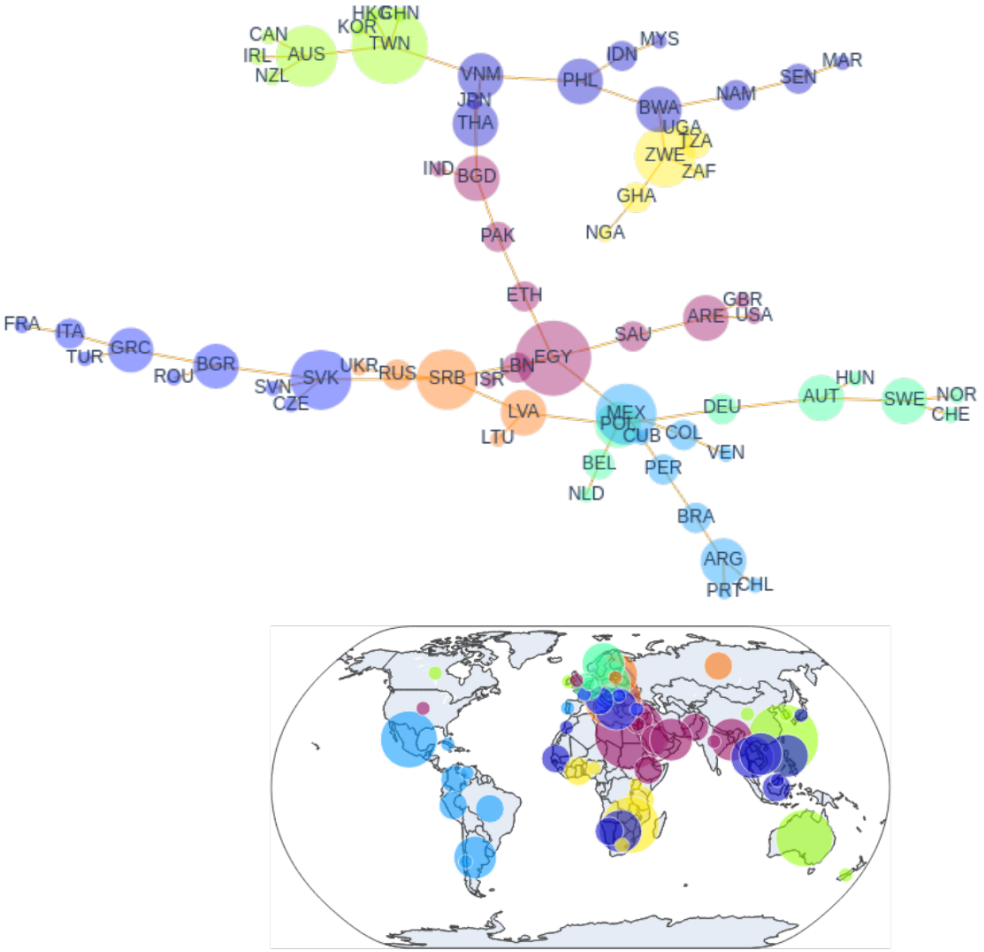


Figure 1: Example of a Geographic Representation network and it's corresponding location clusters (colored) recovered from the top-50 country-"expert" neurons of BLOOM. Notice that connected countries are either geographically or culturally close (e.g. south American cluster in light blue, African countries in yellow, South-East Asian countries in dark blue). *Note: node size is proportional to its degree in the graph.*

Geographic and Geopolitical Biases of Language Models

Fahim Faisal, Antonios Anastasopoulos
Department of Computer Science, George Mason University
{ffaisal,antonis}@gmu.edu

The Bible as a Parallel Corpus: Annotating the "Book of 2000 Tongues"

PHILIP RESNIK, MARI BROMAN OLSEN and MONA DIAB
Department of Linguistics and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA (E-mail: {resnik,molsen,mdiab}@umiacs.umd.edu)

13

A massively parallel corpus: the Bible in 100 languages

Christos Christodouloupoulos · Mark Steedman

Data

Datas

Partia

/erse.

615

aau, aaz, abx, aby, acf, acu, adz, aey, agd, agg, agm, agn, agr, agu, aia, ake, alp, alq, als, aly, ame, amk, amp, amr, amu, anh, anv, aoi, aoj, apb, apn, apu, apy, arb, arl, arn, arp, aso, ata, atb, atd, atg, auc, aui, auy, avt, awb, awk, awx, azg, azz, bao, bbb, bbr, bch, bco, bdd, bea, bel, bgs, bgt, bhg, bhl, big, bjr, bjv, bkd, bki, bkq, bkx, bla, blw, blz, bmh, bmk, bmr, bnp, boa, boj, bon, box, bqc, bre, bsn, bsp, bss, buk, bus, bvr, bxh, byx, bzd, bzj, cab, caf, cao, cap, car, cav, cax, cbc, cbi, cbk, cbr, cbs, cbt, cbu, cbv, cco, ces, cgc, cha, chd, chf, chk, chq, chz, cjo, cjv, cle, clu, cme, cmn, cni, cnl, cnt, cof, con, cop, cot, cpa, cpb, cpc, cpu, crn, crx, cso, cta, ctp, ctu, cub, cuc, cui, cut, cux, cwe, daa, dad, dah, ded, deu, dgr, dgz, dif, dik, dji, djk, dob, dwr, dww, dwy, eko, emi, emp, eng, epo, eri, ese, etr, faa, fai, far, for, fra, fuf, gai, gam, gaw, gdn, gdr, geb, gfk, ghs, gia, glk, gmv, gng, gnn, gnw, gof, grc, gub, guh, gui, gul, gum, guo, gvc, gvf, gwi, gym, gyr, hat, haw, hbo, hch, heb, heg, hix, hla, hlt, hns, hop, hrv, hub, hui, hus, huu, huv, hvn, ign, ikk, ikw, imo, inb, ind, ino, iou, ipi, ita, jac, jao, jic, jiv, jpn, jvn, kaq, kbc, kbh, kbm, kdc, kde, kdl, kek, ken, kew, kgk, kgp, khs, kje, kjs, kkc, kky, klt, klv, kms, kmu, kne, knf, knj, kos, kpf, kpg, kpj, kpw, kqa, kqc, kqf, kql, kqw, ksj, ksr, ktm, kto, kud, kue, kup, kvn, kwd, kwf, kwi, kwj, kyf, kyg, kyq, kyz, kze, lac, lat, lbb, leu, lex, lgl, lid, lif, lww, maa, maj, maq, mau, mav, maz, mbb, mbc, mbh, mbl, mbt, mca, mcb, mcd, mcf, mcp, mdy, med, mee, mek, meq, met, meu, mgh, mgw, mhl, mib, mic, mie, mig, mih, mil, mio, mir, mit, miz, mjc, mkn, mks, mlh, mlp, mmx, mna, mop, mox, mph, mpj, mpm, mpp, mps, mpx, mqb, mqj, msb, msc, msk, msm, msy, mti, muy, mva, mvn, mwc, mxb, mxp, mxq, mxt, myu, myw, myy, mzz, nab, naf, nak, nay, nbq, nca, nch, ncj, ncl, ncu, ndj, nfa, ngp, ngu, nhg, nhi, nho, nhr, nhu, nhw, nhy, nif, nin, nko, nld, nlg, nna, nnq, not, nou, npl, nsn, nss, ntj, ntp, nwi, nyu, obo, ong, ons, ood, opm, ote, otm, otn, otq, ots, pab, pad, pah, pao, pes, pib, pio, pir, pjt, plu, pma, poe, poi, pon, poy, ppo, prf, pri, ptp, ptu, pwg, quc, quf, quh, qul, qup, qvc, qve, qvh, qvm, qvn, qvs, qvw, qvz, qwh, qxh, qxn, qxo, rai, rkb, rmc, roo, rop, rro, ruf, rug, rus, sab, san, sbe, seh, sey, sgz, shj, shp, sim, sja, sll, smk, snc, snn, sny, som, soq, spa, spl, spm, sps, spy, sri, srm, srn, srp, srq, ssd, ssg, ssx, stp, sua, sue, sus, suz, swe, swh, swp, sxb, tac, tav, tbc, tbl, tbo, tbz, tca, tee, ter, tew, tfr, tgp, tif, tim, tiy, tke, tku, tna, tnc, tnn, tnp, toc, tod, toj, ton, too, top, tos, tpt, trc, tsw, ttc, tue, tuo, txu, ubr, udu, ukr, uli, ura, urb, usa, usp, uvl, vid, vie, viv, vmy, waj, wal, wap, wat, wbp, wed, wer, wim, wmt, wmw, wnc, wnu, wos, wrk, wro, wsk, wuv, xav, xed, xla, xnn, xon, xsi, xtd, xtm, yaa, yad, yal, yap, yaq, yby, ycn, yka, yml, yre, yuj, yut, yuw, yva, zaa, zab, zac, zad, zai, zaj, zam, zao, zar, zas, zat, zav, zaw, zca, zia, ziw, zos, zpc, zpl, zpo, zpq, zpu, zpv, zpz, zsr, ztq, zty, zyp

Creating a Massively Parallel Bible Corpus

Thomas Mayer, Michael Cysouw
Research Unit Quantitative Language Comparison
Philipps University of Marburg
thomas.mayer@uni-marburg.de, cysouw@uni-marburg.de

830+ Disappeared ….?

The Johns Hopkins University Bible Corpus: 1600+ Tongues for Typological Exploration

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and **David Yarowsky**

Center for Language and Speech Processing

Johns Hopkins University

(arya, rewicks, dlewis77, amueller, wswu, oadams, gnicola2, yarowsky)@jhu.edu,

post@cs.jhu.edu

Family	JHU	ETHN	JHU %	ETHN %
Niger–Congo	313	1542	19.43	20.63
Austronesian	277	1257	17.19	16.82
Trans-New Guinea	133	482	8.26	6.45
Sino-Tibetan	101	455	6.27	6.09
Indo-European	91	448	5.65	5.99
Otomanguean	83	178	5.15	2.38
Afro-Asiatic	67	377	4.16	5.04
Nilo-Saharan	52	206	3.23	2.76
Creole	27	93	1.68	1.24
Quechuan	27	44	1.68	0.59
Uto-Aztecan	26	61	1.61	0.82
Mayan	25	31	1.55	0.41
Maipurean	24	56	1.49	0.75
Turkic	20	41	1.24	0.55
Australian	19	381	1.18	5.10
Tucanoan	16	25	0.99	0.323
Tupian	15	76	0.93	1.02
Austro-Asiatic	14	167	0.87	2.23
Language isolate	14	88	0.87	1.18
Algic	12	42	0.74	0.56

Table 1: The top 20 largest language families in the JHUBC corpus. JHU percent denotes the percent of the languages in this corpus that are in each language family (normalized by 1611). Ethnologue percent denotes the percent of all Ethnologue languages that are a member of this family (normalized by 7474).

aai
acn
agm
alj
amn
apd
arz
avu
bam
beh
bible
blh
boj
bsn
bul
bzj
cbc
cfm
cjo
cnl
crh
ctu
czt
dig
dtp
emp
eus
for
gbo
gnd
gub
gux
hak
hil
hto
iba
ilo
ita
jiv
kat
kek
khk
kle
knk
kpx
kru
kus
kyz
ldi
lif
lsm
mak
mbf
mdy
mfq
mie
miz
mnb
mox
msy
mvp
myb
mzz
ncu
nhi
njm
nog
ntm
nyn
ong
pab
pbl
plw
prf
quc
qvi
qxo
roo
sbd
shp
sml
soy
srq
sus
tac
tca
tgl
tlh
tos
tsn
tuo
ubu
usp
wap
wnc
xho
xtd
yle
yva
zaw
zpu
Genesis
Exodus
Leviticus
Numbers
Deuteronomy
Joshua
Judges
Ruth
1 Samuel
2 Samuel
1 Kings
2 Kings
1 Chronicles
2 Chronicles
Ezra
Nehemiah
Esther
Job
Psalms
Proverbs
Ecclesiastes
Song of Solomon
Isaiah
Jeremiah
Lamentations
Ezekiel
Daniel
Hosea
Joel
Amos
Obadiah
Jonah
Micah
Nahum
Habakkuk
Zephaniah
Haggai
Zechariah
Malachi
Matthew
Mark
Luke
John
Acts
Romans
1 Corinthians
2 Corinthians
Galatians
Ephesians
Philippians
Colossians
1 Thessalonians
2 Thessalonians
1 Timothy
2 Timothy
Titus
Philemon
Hebrews
James
1 Peter
2 Peter
1 John
2 John
3 John
Jude
Revelation
1.0
0.8
0.6
0.4
0.2
0.0

JW300: A Wid… Corpus for Low Resource Languages

Departme…
IT Universit…
z…

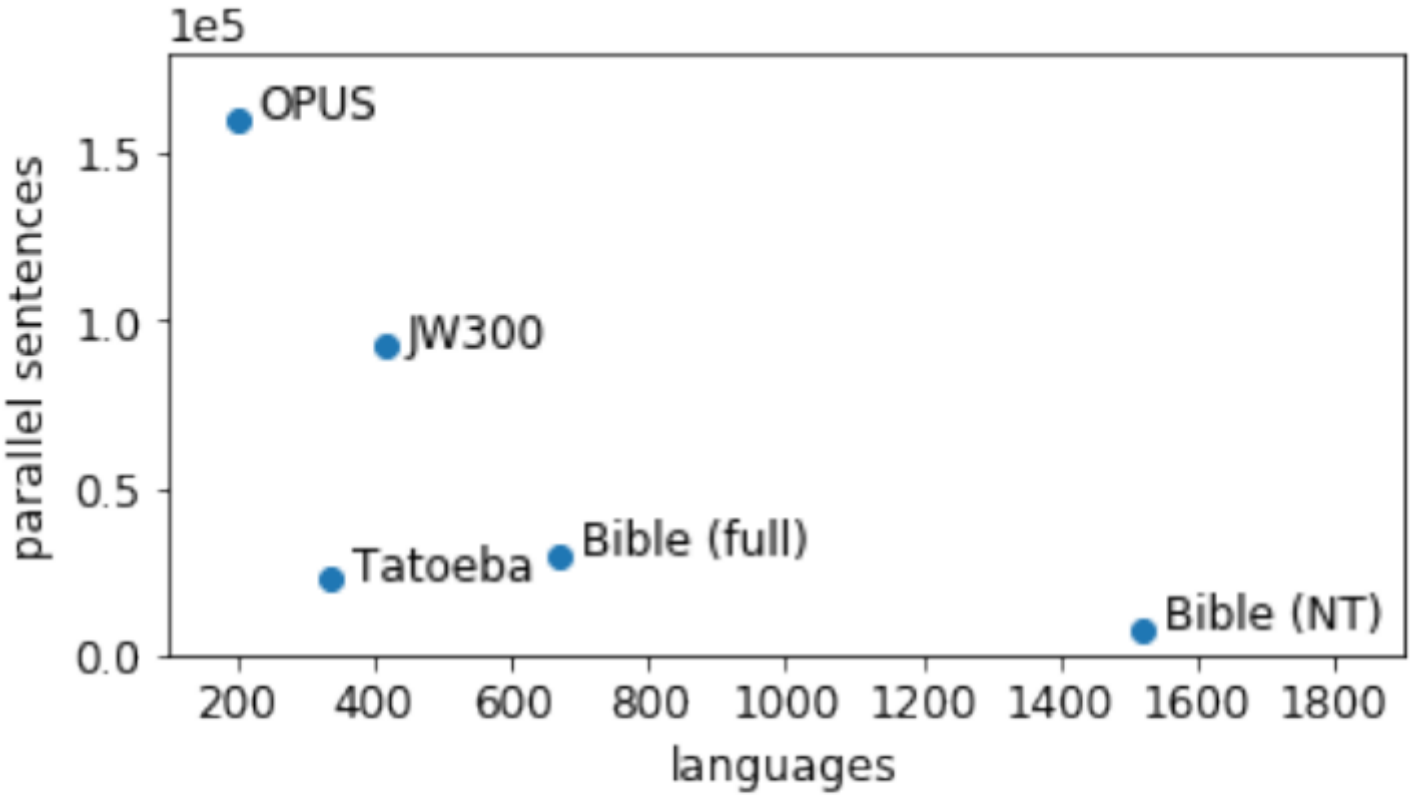


Figure 1: Our dataset JW300 in comparison to other massive parallel text collections with respect to multilingual breadth and volume of parallel sentences. The y-axis depicts the mean number of parallel sentences per language pair.

LORELEI

- Low Resource Languages for Emergent Incidents (LORELEI)

Overview of the DARPA LORELEI Program

Caitlin Christianson[1] · Jason Duncan[2] · Boyan Onyshkevych[1]

UN marks anniversary of devastating 2010 Haiti earthquake

UN Photo/Logan Abassi | UN peacekeepers take a break while working through the rubble of the UN mission in Haiti's headquarters in Port au Prince, in the aftermath of the devastating January 2010 earthquake.

https://news.un.org/en/story/2022/01/1109632

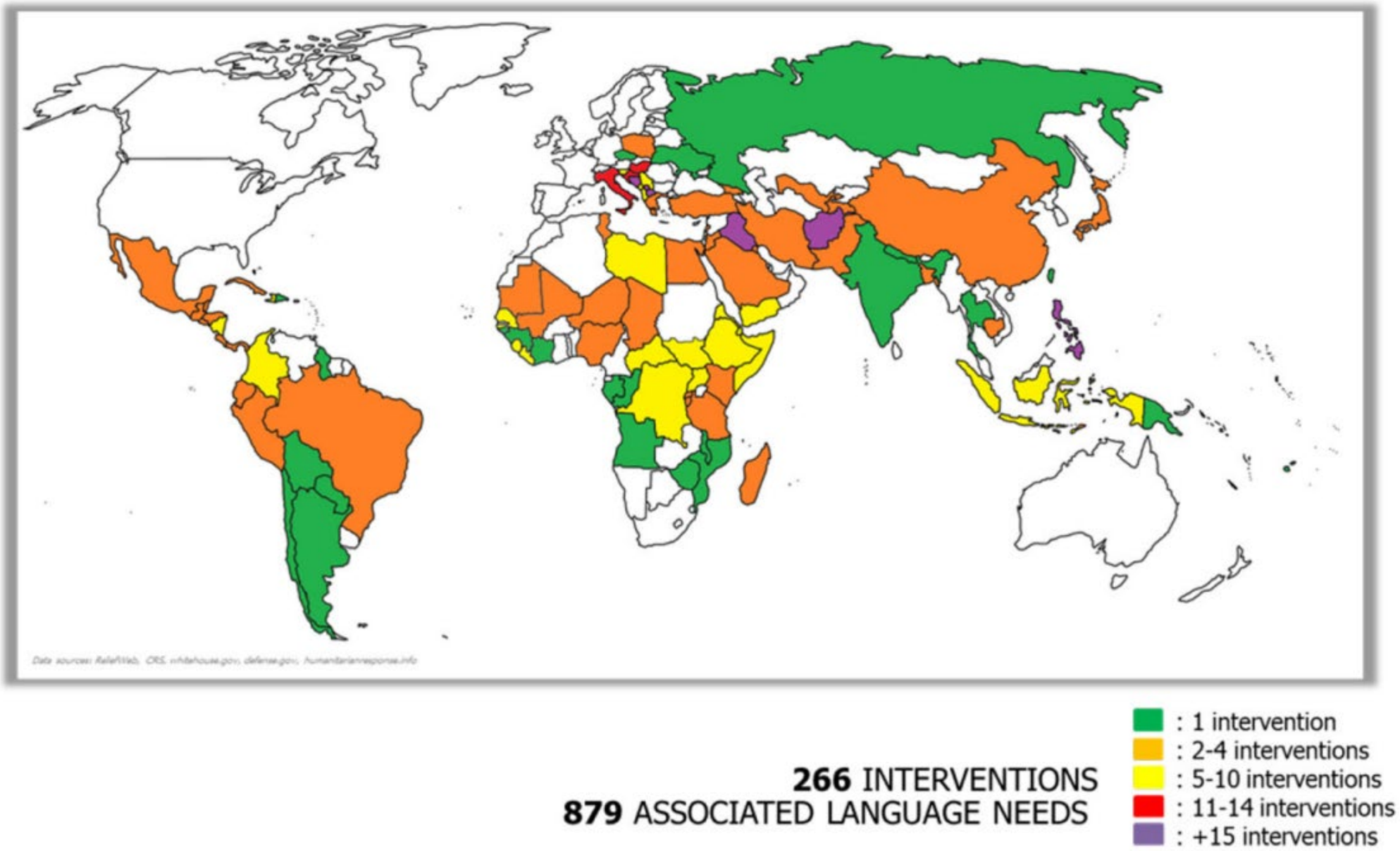


Fig. 1 Interventions involving deployment of U.S. personnel 1990–2014

Kinyarwanda	Uyghur
Oromo	Uzbek
Sinhala	IL11 (undisclosed)
Tigrinya	IL12 (undisclosed)
Ukrainian	

Table 1: LORELEI Incident Languages

Akan (Twi)	Swahili
Amharic	Tagalog
Arabic	Tamil
Bengali	Thai
Farsi	Vietnamese
Hindi	Wolof
Hungarian	Yoruba
Indonesian	Zulu
Mandarin	English (partial)
Russian	Hausa (partial)
Somali	Turkish (partial)
Spanish	

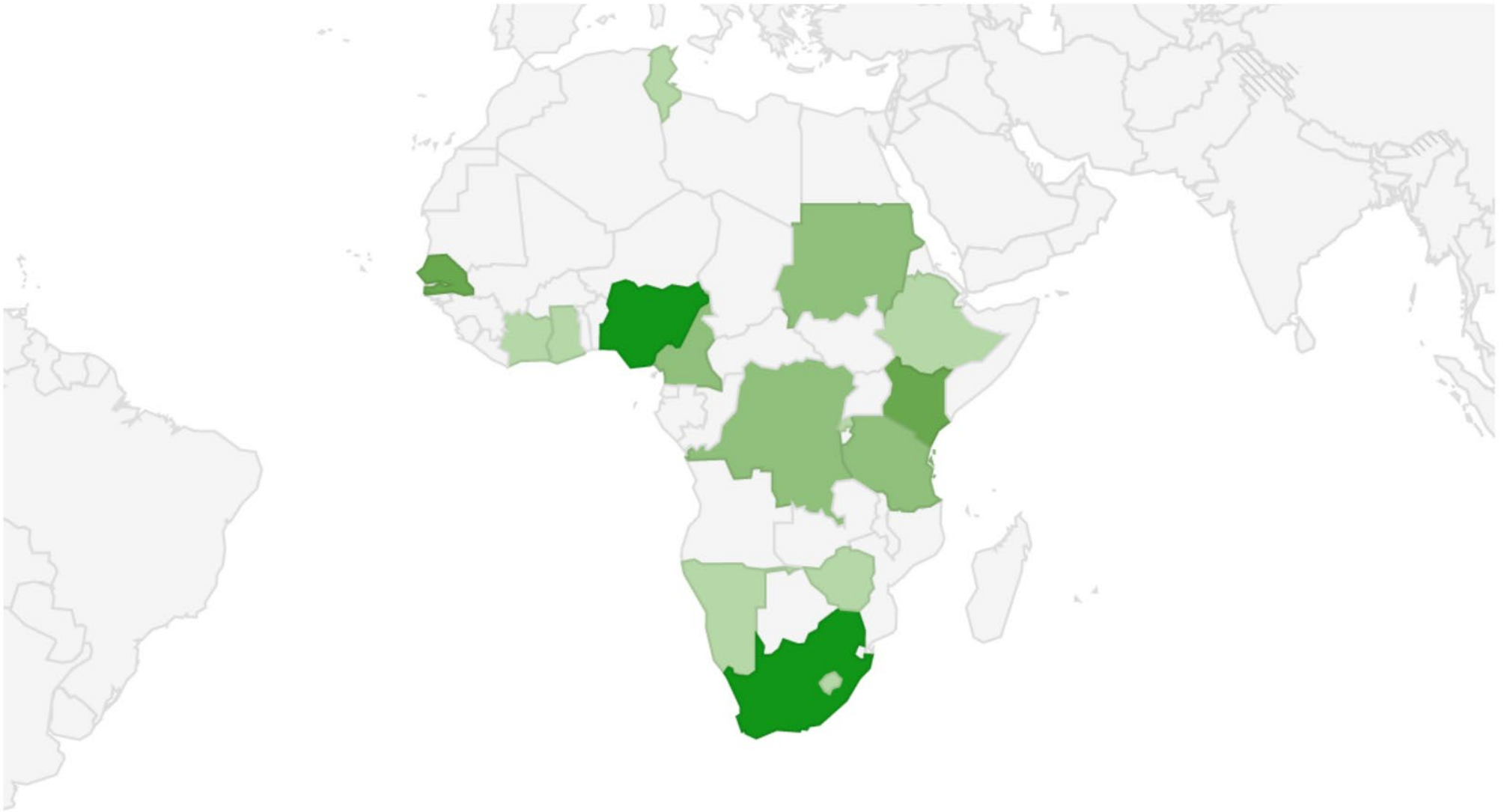
Table 2: LORELEI Representative Languages

Masakhane

- Roughly translates to “We build together” in isiZulu

Masakhane

A grassroots NLP community for Africa, by Africans



Above: Countries where Masakhane is active since launch

https://venturebeat.com/ai/the-masakhane-project-wants-machine-translation-and-ai-to-transform-africa/

Towards Neural Machine Translation for African Languages

Jade Z. Abbott
Retro Rabbit
ja@retrorabbit.co.za

Laura Martinus
Human Language Technologies, CSIR
lmartinus@csir.co.za

2018

Benchmarking Neural Machine Translation for Southern African Languages

Laura Martinus
Explore Data Science Academy, South Africa
laura@explore-ai.net

Jade Z. Abbott
Retro Rabbit, South Africa
jabbott@retrorabbit.co.za

2019

MasakhaNER: Named Entity Recognition for African Languages

David Ifeoluwa Adelani[1*], Jade Abbott[2*], Graham Neubig[3], Daniel D'souza[4*], Julia Kreutzer[5*], Constantine Lignos[6*], Chester Palen-Michel[6*], Happy Buzaaba[7*], Shruti Rijhwani[3], Sebastian Ruder[8], Stephen Mayhew[9], Israel Abebe Azime[10*], Shamsuddeen H. Muhammad[11,12*], Chris Chinenye Emezue[13*], Joyce Nakatumba-Nabende[14*], Perez Ogayo[15*], Aremu Anuoluwapo[16*], Catherine Gitau[*], Derguene Mbaye[*], Jesujoba Alabi[17*], Seid Muhie Yimam[18], Tajuddeen Rabiu Gwadabe[19*], Ignatius Ezeani[20*], Rubungo Andre Niyongabo[21*], Jonathan Mukiibi[14], Verrah Otiende[22*], Iroro Orife[23*], Davis David[*], Samba Ngom[*], Tosin Adewumi[24*], Paul Rayson[20], Mofetoluwa Adeyemi[*], Gerald Muriuki[14], Emmanuel Anebi[*], Chiamaka Chukwuneke[20], Nkiruka Odu[25], Eric Peter Wairagala[14], Samuel Oyerinde[*], Clemencia Siro[*], Tobius Saul Bateesa[14], Temilola Oloyede[*], Yvonne Wambui[*], Victor Akinode[*], Deborah Nabagereka[14], Maurice Katusiime[14], Ayodele Awokoya[26*], Mouhamadane MBOUP[*], Dibora Gebreyohannes[*], Henok Tilaye[*], Kelechi Nwaike[*], Degaga Wolde[*], Abdoulaye Faye[*], Blessing Sibanda[27*], Orevaoghene Ahia[28*], Bonaventure F. P. Dossou[29*], Kelechi Ogueji[30*], Thierno Ibrahima DIOP[*], Abdoulaye Diallo[*], Adewale Akinfaderin[*], Tendai Marengereke[*], and Salomey Osei[10*]

2021

Language	Sentence
English	The Emir of Kano turbaned Zhang who has spent 18 years in Nigeria
Amharic	የካኖ ኢምር በናይጀርያ ፩፰ ዓመት ያሳለፈውን ዛንግን ዋና መሪ አደረጉት
Hausa	Sarkin Kano yayi wa Zhang wanda yayi shekara 18 a Najeriya sarauta
Igbo	Onye Emir nke Kano kpubere Zhang okpu onye nke nọgoro afọ iri na asatọ na Naijiria
Kinyarwanda	Emir w'i Kano yimitse Zhang wari umaze imyaka 18 muri Nijeriya
Luganda	Emir w'e Kano yatikkidde Zhang amaze emyaka 18 mu Nigeria
Luo	Emir mar Kano ne orwakone turban Zhang ma osedak Nigeria kwuom higni 18
Nigerian-Pidgin	Emir of Kano turban Zhang wey don spend 18 years for Nigeria
Swahili	Emir wa Kano alimvisha kilemba Zhang ambaye alikaa miaka 18 nchini Nigeria
Wolof	Emiiru Kanó dafa kaala kii di Zhang mii def Nigeria fukki at ak juróom ñett
Yorùbá	Ẹ́míà ìlú Kánò wé láwàní lé orí Zhang ẹni tí ó ti lo ọdún méjìdínlógún ní orílè-èdè Nàìjíríà

Table 2: Example of named entities in different languages. PER, LOC, and DATE are in colours purple, orange, and green, respectively.

Language	Data Source	Train/ dev/ test	# Anno.	PER	ORG	LOC	DATE	% of Entities in Tokens	# Tokens
Amharic	DW & BBC	1750/ 250/ 500	4	730	403	1,420	580	15.13	37,032
Hausa	VOA Hausa	1903/ 272/ 545	3	1,490	766	2,779	922	12.17	80,152
Igbo	BBC Igbo	2233/ 319/ 638	6	1,603	1,292	1,677	690	13.15	61,668
Kinyarwanda	IGIHE news	2110/ 301/ 604	2	1,366	1,038	2,096	792	12.85	68,819
Luganda	BUKEDDE news	2003/ 200/ 401	3	1,868	838	943	574	14.81	46,615
Luo	Ramogi FM news	644/ 92/ 185	2	557	286	666	343	14.95	26,303
Nigerian-Pidgin	BBC Pidgin	2100/ 300/ 600	5	2,602	1,042	1,317	1,242	13.25	76,063
Swahili	VOA Swahili	2104/ 300/ 602	6	1,702	960	2,842	940	12.48	79,272
Wolof	Lu Defu Waxu & Saabal	1,871/ 267/ 536	2	731	245	836	206	6.02	52,872
Yorùbá	GV & VON news	2124/ 303/ 608	5	1,039	835	1,627	853	11.57	83,285

Table 3: Statistics of our datasets including their source, number of sentences in each split, number of annotators, number of entities of each label type, percentage of tokens that are named entities, and total number of tokens.

MasakhaNER 2.0: Africa-centric Transfer Learning for Named Entity Recognition

David Ifeoluwa Adelani[1,2,*], Graham Neubig[3], Sebastian Ruder[4], Shruti Rijhwani[3], Michael Beukman[5*], Chester Palen-Michel[6*], Constantine Lignos[6*], Jesujoba O. Alabi[1*], Shamsuddeen H. Muhammad[7*], Peter Nabende[8*], Cheikh M. Bamba Dione[9*], Andiswa Bukula[10], Rooweither Mabuya[10], Bonaventure F. P. Dossou[11*], Blessing Sibanda[*], Happy Buzaaba[12*], Jonathan Mukiibi[8*], Godson Kalipe[*], Derguene Mbaye[13*], Amelia Taylor[14*], Fatoumata Kabore[15*], Chris Chinenye Emezue[16*], Anuoluwapo Aremu[*], Perez Ogayo[3*], Catherine Gitau[*], Edwin Munkoh-Buabeng[17*], Victoire M. Koagne[*], Allahsera Auguste Tapo[18*], Tebogo Macucwa[19*], Vukosi Marivate[19*], Elvis Mboning[*], Tajuddeen Gwadabe[*], Tosin Adewumi[20*], Orevaoghene Ahia[21*], Joyce Nakatumba-Nabende[8*], Neo L. Mokono[19*], Ignatius Ezeani[22*], Chiamaka Chukwuneke[22*], Mofetoluwa Adeyemi[23*], Gilles Q. Hacheme[24*], Idris Abdulmumin[25*], Odunayo Ogundepo[23*], Oreen Yousuf[15*], Tatiana Moteu Ngoli[*], Dietrich Klakow[1]

2022

Language	Family	African Region	No. of Speakers	Source	Train / dev / test	% Entities in Tokens	# Tokens
Bambara (bam)	NC / Mande	West	14M	MAFAND-MT (Adelani et al., 2022)	4462/ 638/ 1274	6.5	155,552
Ghomálá' (bbj)	NC / Grassfields	Central	1M	MAFAND-MT (Adelani et al., 2022)	3384/ 483/ 966	11.3	69,474
Éwé (ewe)	NC / Kwa	West	7M	MAFAND-MT (Adelani et al., 2022)	3505/ 501/ 1001	15.3	90420
Fon (fon)	NC / Volta-Niger	West	2M	MAFAND-MT (Adelani et al., 2022)	4343/ 621/ 1240	8.3	173,099
Hausa (hau)	Afro-Asiatic / Chadic	West	63M	Kano Focus and Freedom Radio	5716/ 816/ 1633	14.0	221,086
Igbo (ibo)	NC / Volta-Niger	West	27M	IgboRadio and Ka Ọdị Taa	7634/ 1090/ 2181	7.5	344,095
Kinyarwanda (kin)	NC / Bantu	East	10M	IGIHE, Rwanda	7825/ 1118/ 2235	12.6	245,933
Luganda (lug)	NC / Bantu	East	7M	MAFAND-MT (Adelani et al., 2022)	4942/ 706/ 1412	15.6	120,119
Luo (luo)	Nilo-Saharan	East	4M	MAFAND-MT (Adelani et al., 2022)	5161/ 737/ 1474	11.7	229,927
Mossi (mos)	NC / Gur	West	8M	MAFAND-MT (Adelani et al., 2022)	4532/ 648/ 1294	9.2	168,141
Naija (pcm)	English-Creole	West	75M	MAFAND-MT (Adelani et al., 2022)	5646/ 806/ 1613	9.4	206,404
Chichewa (nya)	NC / Bantu	South-East	14M	Nation Online Malawi	6250/ 893/ 1785	9.3	263,622
chiShona (sna)	NC / Bantu	South	12M	VOA Shona	6207/ 887/ 1773	16.2	195,834
Kiswahili (swa)	NC / Bantu	East & Central	98M	VOA Swahili	6593/ 942/ 1883	12.7	251,678
Setswana (tsn)	NC / Bantu	South	14M	MAFAND-MT (Adelani et al., 2022)	3489/ 499/ 996	8.8	141,069
Akan/Twi (twi)	NC / Kwa	West	9M	MAFAND-MT (Adelani et al., 2022)	4240/ 605/ 1211	6.3	155,985
Wolof (wol)	NC / Senegambia	West	5M	MAFAND-MT (Adelani et al., 2022)	4593/ 656/ 1312	7.4	181,048
isiXhosa (xho)	NC / Bantu	South	9M	Isolezwe Newspaper	5718/ 817/ 1633	15.1	127,222
Yorùbá (yor)	NC / Volta-Niger	West	42M	Voice of Nigeria and Asejere	6877/ 983/ 1964	11.4	244,144
isiZulu (zul)	NC / Bantu	South	27M	Isolezwe Newspaper	5848/ 836/ 1670	11.0	128,658

Table 1: **Languages and Data Splits for MasakhaNER 2.0 Corpus**. Language, family (NC: Niger-Congo), number of speakers, news source, and data split in number of sentences

AfroLM: A Self-Active Learning-based Multilingual Pretrained Language Model for 23 African Languages

Bonaventure F. P. Dossou[1,2,*], **Atnafu Lambebo Tonja**[3,*], **Oreen Yousuf**[4,*], **Salomey Osei**[5,*], **Abigail Oppong**[6,*], **Iyanuoluwa Shode**[7,*], **Oluwabusayo Olufunke Awoyomi**[8,*], **Chris Chinenye Emezue**[1,9,*]

[*]Masakhane NLP, [1]Mila Quebec AI Institute, Canada, [2] McGill University, Canada, [3]Instituto Politécnico Nacional, Mexico, [4]Uppsala University, Sweden, [5]Universidad de Deusto, Spain [6]Ashesi University, Ghana, [7]Montclair State University, USA, [8]The College of Saint Rose, USA, [9]Technical University of Munich, Germany

2022

Languages	Family	Writing System	African Region	No of Speakers	Initial # of Sentences	Source	Size (MB)
Amharic (amh)	Afro-Asiatic/Semitic	Ge'ez script	East	57M	655,079	✣,✝,★	279
Afan Oromo (orm)	Afro-Asiatic/Cushitic	Latin script	East	37.4M	50,105	✝	9.87
Bambara (bam)	NC/Manding	Latin, Arabic(Ajami), N'ko	West	14M	6,618	✣	1.00
Ghomálá' (bbj)	NC/Grassfields	Latin script	Central	1M	4,841	✣	0.50
Éwé (ewe)	NC/Kwa	Latin (Ewe alphabet)	West	7M	5,615	✣	0.50
Fon (fon)	NC/Volta-Niger	Latin script	West	1.7M	5,448	✣	1.00
Hausa (hau)	Afro-Asiatic/Chadic	Latin (Boko alphabet)	West	63M	1,626,330	✣,✝,★	208
Igbo (ibo)	NC/Volta-Niger	Latin (Önwu alphabet)	West	27M	437,737	✣,✝,★	63
Kinyarwanda (kin)	NC/Rwanda-Rundi	Latin script	Central	9.8M	84,994	➳,✝,✣	37.70
Lingala (lin)	NC/Bang	Latin script	Central & East	45M	398,440	✣	45.90
Luganda (lug)	NC/Bantu	Latin script (Ganda alphabet)	East	7M	74,754	✝,✣	8.34
Luo (luo)	Nilo-Saharan	Latin script	East	4M	8,684	✝	1.29
Mooré (mos)	NC/Gur	Latin script	West	8M	27,908	✣,✝	5.05
Chewa (nya)	NC/Nyasa	Latin script	South & East	12M	8,000	✣	1.66
Naija (pcm)	English-Creole	Latin script	West	75M	345,694	✣,✝,★	101
Shona (sna)	NC/Bantu	Latin script (Shona alphabet)	Southeast	12M	187,810	✣,✝	32.80
Swahili (swa)	NC / Bantu	Latin script (Roman Swahili alphabet)	East & Central	98M	1,935,485	✣,✝,★	276
Setswana (tsn)	NC / Bantu	Latin (Tswana alphabet)	South	14M	13,958	✣,✝	2.21
Akan/Twi (twi)	NC / Kwa	Latin script	West	9M	14,701	✣	1.61
Wolof (wol)	NC / Senegambia	Latin (Wolof alphabet)	West	5M	13,868	✝	2.20
Xhosa (xho)	NC/Zunda	Latin (Xhosa alphabet)	South	20M	93,288	✣,✝	17.40
Yorůbá (yor)	NC / Volta-Niger	Latin (Yorùbá alphabet)	West	42M	290,999	✣,✝,★	45.9
isiZulu (zul)	NC / Bantu	Latin (Zulu alphabet)	South	27M	194,562	✣,✝	33.70

Table 1: **Languages Corpora Details**. **Legends**: (Adelani et al., 2022a) → ✣, (Alabi et al., 2022a) → ✝, (Kelechi et al., 2021) → ★, (Niyongabo et al., 2020) → ➳.

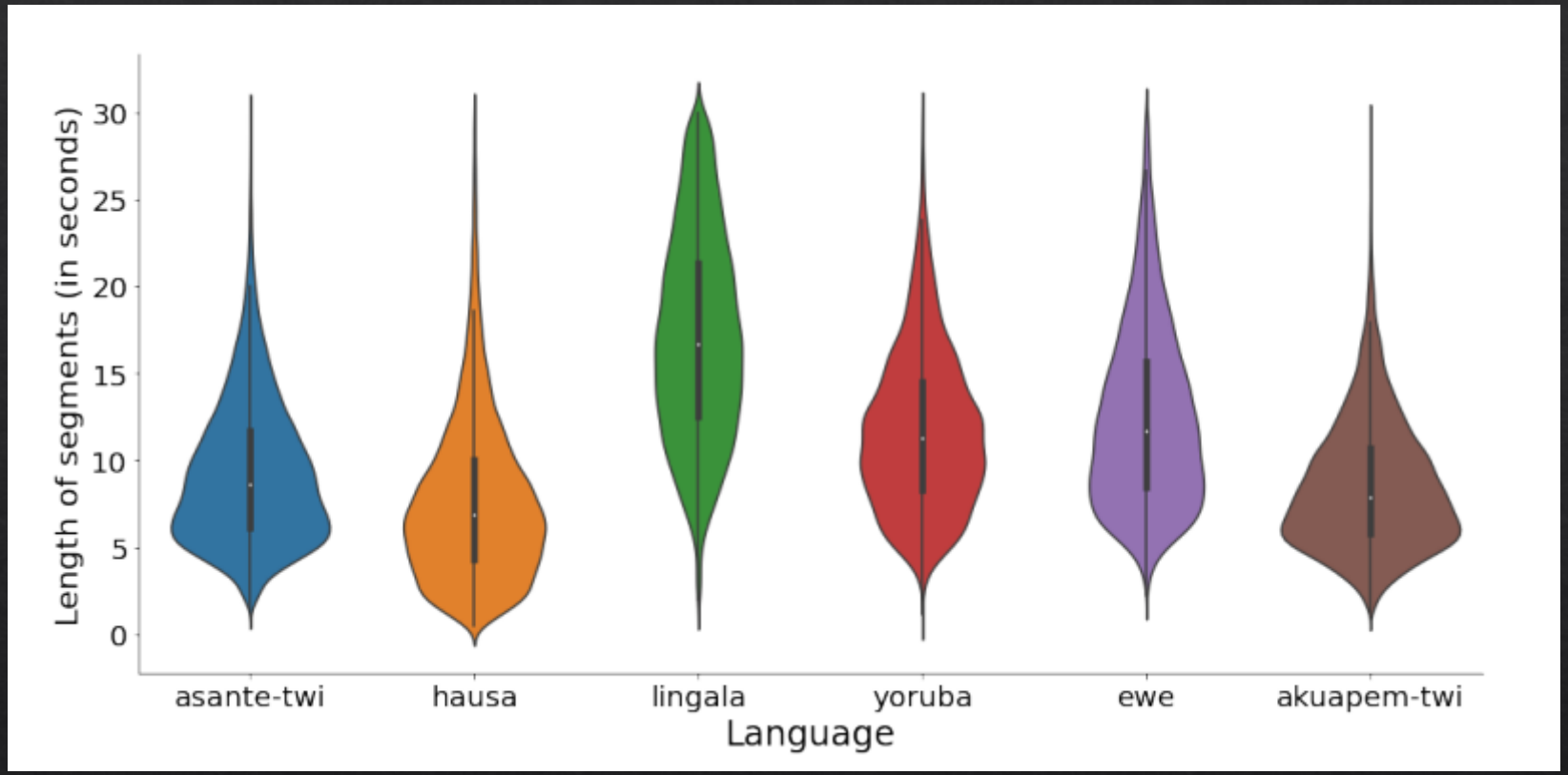
…. And returning

BibleTTS: a large, high-fidelity, multilingual, and uniquely African speech corpus

Josh Meyer[1], *David Ifeoluwa Adelani*[2], *Edresson Casanova*[1,3], *Alp Öktem*[4,5], *Daniel Whitenack*[6] *Julian Weber*[1], *Salomon Kabongo*[7], *Elizabeth Salesky*[8], *Iroro Orife*[9], *Colin Leong, Perez Ogayo*[10], *Chris Emezue*[11], *Jonathan Mukiibi* [12], *Salomey Osei, Apelete Agbolo*[13], *Victor Akinode, Bernard Opoku*[14], *Samuel Olanrewaju, Jesujoba Alabi*[2], *Shamsuddeen Muhammad*

[∀] Masakhane NLP, Africa [1] Coqui, USA [2] Saarland University, Germany [3] University of São Paulo, Brazil [4] CLEAR Global, USA [5] Col·lectivaT, Spain [6] SIL International [7] Leibniz Universität Hannover, Germany [8] Johns Hopkins University, USA [9] Niger-Volta LTI, USA, [10] Carnegie Mellon University, USA [11] Technical University of Munich, Germany [12] Makerere University, Uganda [13] Ewegbe Akademi, Togo, [14] Kwame Nkrumah University of Science and Technology, Ghana

josh@coqui.ai, didelani@lsv.uni-saarland.de, edresson@coqui.ai, alp.oktem@clearglobal.org, dan_whitenack@sil.org


Length of segments (in seconds)
0
5
10
15
20
25
30
asante-twi
hausa
lingala
yoruba
ewe
akuapem-twi
Language

Language	Classification	African Region	No. of speakers
Éwé [ewe]	Niger-Congo / Kwa	West	5.5M
Hausa [hau]	Afro-Asiatic / Chadic	West	77M
Kikuyu [kik]	Niger-Congo / Bantu	East	8.2M
Lingala [lin]	Niger-Congo / Bantu	Central	40M
Luganda [lug]	Niger-Congo / Bantu	East	11M
Luo [luo]	Nilo-Saharan / Luo–Acholi	East	5.3M
Chichewa [nya]	Niger-Congo / Bantu	South-East	14M
Akuapem Twi [aka]	Niger-Congo / Akan	West	626k
Asante Twi [aka]	Niger-Congo / Akan	West	3.8M
Yorùbá [yor]	Niger-Congo / Volta-Niger	West	46M

Building African Voices

Perez Ogayo[†], *Graham Neubig*[†‡], *Alan W Black*[†]

[†]Language Technologies Institute, Carnegie Mellon University
[‡]Inspired Cognition
Pittsburgh, PA, USA

{aogayo, gneubig, awb}@cs.cmu.edu

2022

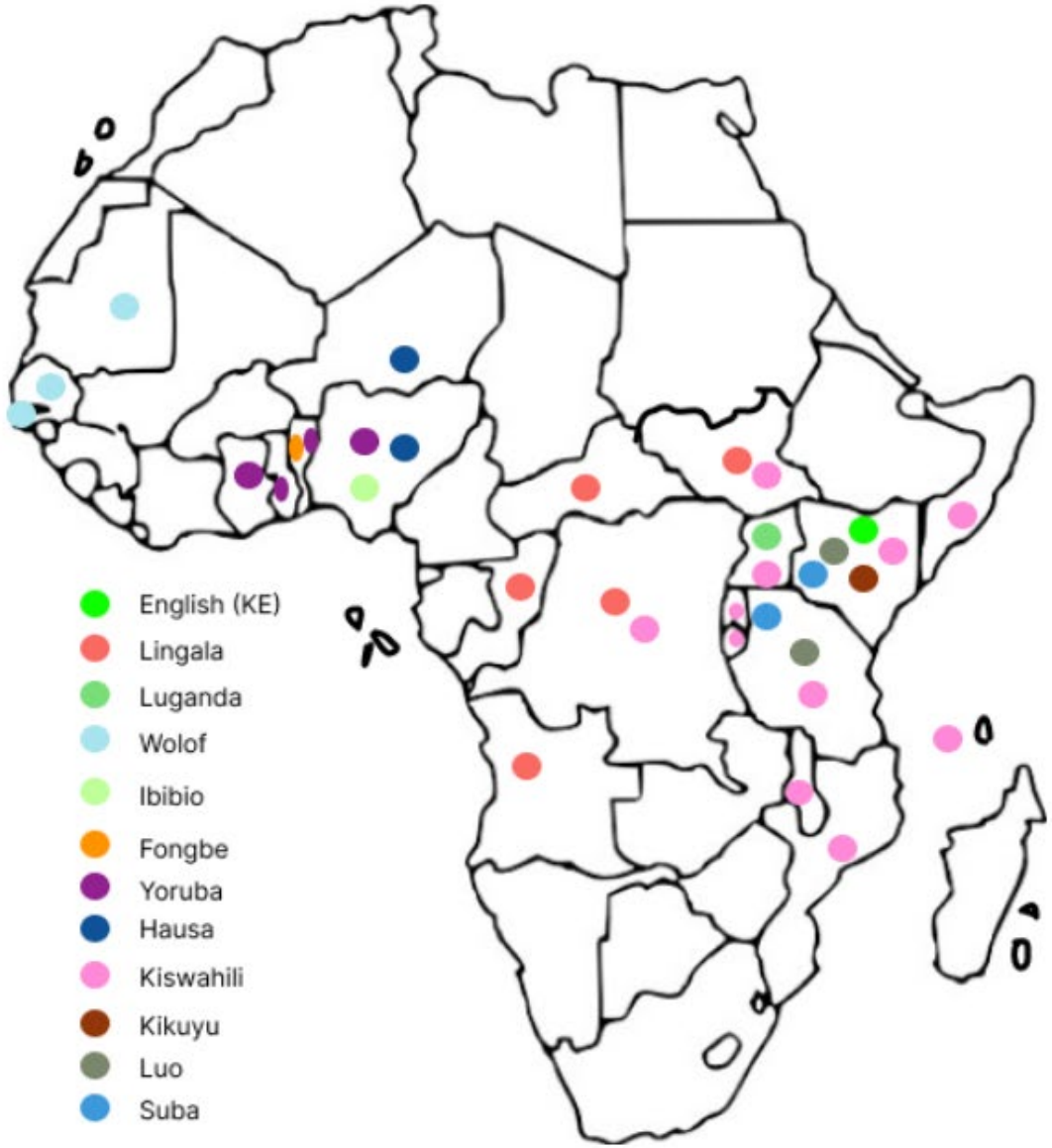


Figure 1: *Current language coverage of African Voices*

3rd Workshop on African Natural Language Processing

AfricaNLP 2022

Virtual | Apr 25 2021 | https://africanlp.masakhane.io/ | maxmed@microsoft.com

Please see the venue website for more information.
Submission Start: Feb 03 2021 11:59PM UTC-0, End: Mar 03 2021 11:59PM UTC-0

Accept | Reject

Search by paper title and metadata

IGBOSUM1500 - INTRODUCING THE IGBO TEXT SUMMARIZATION DATASET

CHINEDU EMMANUEL MBONU, Chiamaka Ijeoma Chukwuneke, Roseline Uzoamaka Paul, Ignatius Ezeani, Ikechukwu Onyenwe

09 Mar 2022 (modified: 02 May 2022) AfricaNLP 2022 Readers: Everyone 1 Reply

Show details

A Comparison of Topic Modeling and Classification Machine Learning Algorithms on Luganda Data

Bateesa Saul Tobius, Claire Babirye, Joyce Nakatumba-Nabende, Andrew Katumba

09 Mar 2022 (modified: 21 Apr 2022) AfricaNLP 2022 Readers: Everyone 1 Reply

Show details

Building Text and Speech Datasets for Low Resourced Languages: A Case of Languages in East Africa

Claire Babirye, Joyce Nakatumba-Nabende, Andrew Katumba, Ronald Ogwang, Jeremy Tusubira Francis, Jonathan Mukiibi, Medadi Ssentanda, Lilian D Wanzare, Davis David

09 Mar 2022 (modified: 21 Apr 2022) AfricaNLP 2022 Readers: Everyone 1 Reply

Show details

Goud.ma: a News Article Dataset for Summarization in Moroccan Darija

FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech

Alexis Conneau[†‡1], *Min Ma*[†2], *Simran Khanuja*[†2], *Yu Zhang*[2], *Vera Axelrod*[2], *Siddharth Dalmia*[3], *Jason Riesa*[2], *Clara Rivera*[2], *Ankur Bapna*[‡2]

[1]Meta AI Research, [2]Google Research, [3]Carnegie Mellon University

aconneau@fb.com, {minm,simrankh,ngyuzh,vaxelrod,ankurbpn}@google.com, sdalmia@cs.cmu.edu

Dataset Paper

- n-way parallel
- 102 Languages
- FLoRes-101

FLoRes-101

The FLORES-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation

Naman Goyal[1], **Cynthia Gao**[1], **Vishrav Chaudhary**[1], **Peng-Jen Chen**[1], **Guillaume Wenzek**[2], **Da Ju**[1], **Sanjana Krishnan**[1], **Marc'Aurelio Ranzato**[1], **Francisco Guzmán**[1], **Angela Fan**[2,3]

[1]Facebook AI Research, USA, [2]Facebook AI Research, France, [3]LORIA

flores@fb.com

FLoRes-101

- 3001 English Sentences
- 101 Languages + English → 10,100 Language Pairs

FLoRes v1.0

- Best Dataset Paper at EMNLP 2019 … last conference before the pandemic

The FLORES Evaluation Datasets for Low-Resource Machine Translation: Nepali–English and Sinhala–English

Francisco Guzmán[♡♦] **Peng-Jen Chen**[♡★] **Myle Ott**[★] **Juan Pino**[♦]
Guillaume Lample[★‡] **Philipp Koehn**[■] **Vishrav Chaudhary**[♦] **Marc'Aurelio Ranzato**[★]
[♦]Facebook Applied Machine Learning [★]Facebook AI Research
[‡]Sorbonne Universités [■]Johns Hopkins University
{fguzman,pipibjc,myleott,juancarabina,guismay,vishrav,ranzato}@fb.com
phi@jhu.edu

FLoRes-101

- High Quality
- Good Coverage
- Diverse topics
- Meta-Data
  - Document Level Translation
  - Multimodal Translation
  - Text Classification

FLoRes-101 BLEU

- spBLEU (Kenton does not like)
- Points out problems with many languages & BLEU… but
  - Requires Training
  - spBLEU does not represent how humans look at translations
  - Many of the challenges still there

FLoRes-101

- English Wikimedia
- 1/3 Wikinews
- 1/3 Wikijunior (non-fiction books ages 0-12)
- 1/3 WikiVoyage

FLoRes-101 Creation Workflow

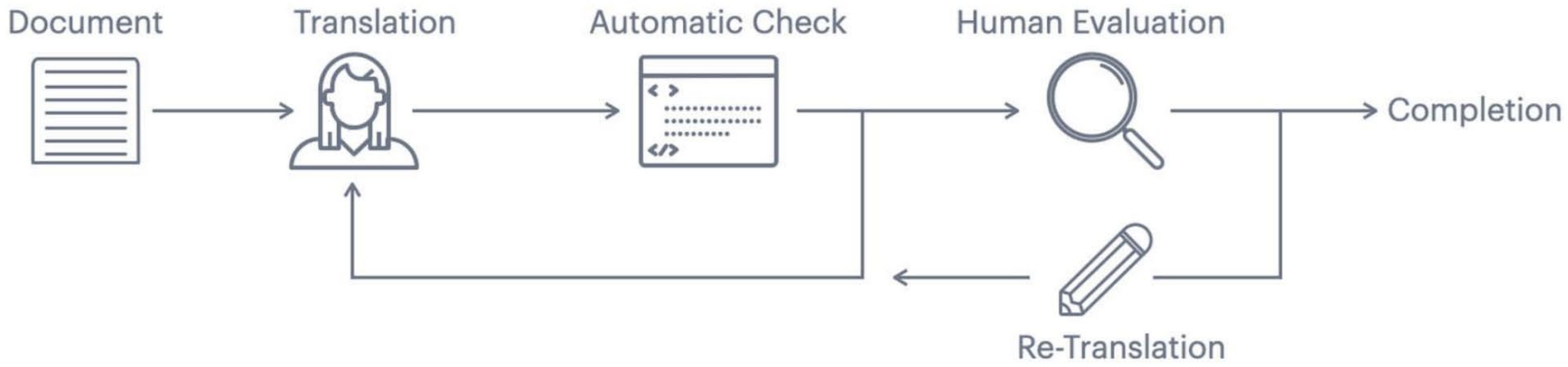


Figure 1: **Depiction of Overall Translation Workflow.**

FLoRes-101 Comparisons

	# Languages	Diverse Topics	Many to Many	Manual Alignments	Document Level	Multi modal
FLORES v1 (Guzmán et al., 2019)	2	✓	✗	✓	✗	✗
AmericasNLI (Ebrahimi et al., 2021)	10	✓	✓	✓	✗	✗
ALT (Riza et al., 2016)	13	✓	✓	✓	✗	✗
Europarl (Koehn, 2005)	21	✗	✓	✗	✓	✗
TICO-19 (Anastasopoulos et al., 2020)	36	✗	✓	✓	✗	✗
OPUS-100 (Zhang et al., 2020)	100	✓	✓	✗	✗	✗
M2M (Fan et al., 2020)	100	✗	✓	✓✗	✗	✗
FLORES-101	101	✓	✓	✓	✓	✓

Table 2: **Comparison of Various Evaluation Benchmarks**. We compare FLORES-101 to a variety of popular, existing translation benchmarks, indicating language coverage, topic diversity, whether many-to-many translation is supported, if the translations are manually aligned by humans, and if the tasks of document-level translation or multimodal translation are supported.

FLoRes-101 Quality

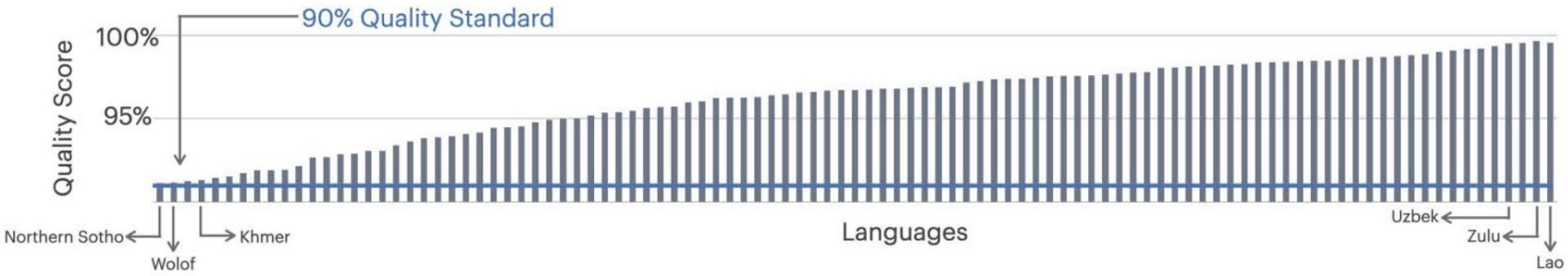


Figure 2: **Translation Quality Score across Languages.** We require the final translation quality score to be above 90% before the translation is of sufficient quality to include in FLORES-101.

Back to FLEURS

FLEURS

- Test set of FLoRes-101 is NOT publicly available
- Use dev and devtest
- 2009 sentences →
  - 1509 Train
  - 150 Dev
  - 350 Test

FLEURS

- 3 Recordings by Native Speakers
- Sex Ratio 30%/70% if possible
- After the first recording, validation by other humans
- 21.5% of sentences missing (none of the 3 validated)
- All recording kept (including noise)
- 16kHz Sampling Rate
- Segments are within 30 seconds
- Disjoint speakers separation
- Aim for Train/Dev/Test of 7:1:2

FLEURS Evaluation

- Similar to Character Error Rate
- NFC and FST normalizations
- Lowercase, normalize, and remove punctuation (how?)
  - SRC_RAW
  - SRC_NORM
  - SRC_CHAR

Comparison

Table 1: *Compare FLEURS to common public multilingual speech benchmarks.*

Dataset	#Languages	Total Duration	Domains	Speech Type	Transcripts	Parallel text	Parallel speech
BABEL [13]	17	1k hours	Conversational	Spontaneous	Yes	No	No
CommonVoice [12]	93	15k hours	Open domain	Read	Yes	No	No
CMU Wilderness [15]	700	14k hours	Religion	Read	Yes	Yes	Yes
MLS [8]	8	50.5k hours	Audiobook	Read	Yes	No	No
CoVoST-2 [11]	22	2.9k hours	Open domain	Read	Yes	Yes	No
Voxlingua-107 [14]	107	6.6k hours	YouTube	Spontaneous	No	No	No
Europarl-ST [16]	6	500 hours	Parliament	Spontaneous	Yes	Yes	No
MuST-C [17]	9	385 hours	TED talks	Spontaneous	Yes	Yes	No
mTEDx [18]	9	1k hours	TED talks	Spontaneous	Yes	Yes	No
VoxPopuli [9]	24	400k hours	Parliament	Spontaneous	Partial	Partial	Partial
CVSS [19]	22	1.1k hours	Open domain	Read/Synthetic	Yes	Yes	Yes
FLEURS (this work)	102	1.4k hours	Wikipedia	Read	Yes	Yes	Yes

Table 2: *A comparison of commonly used datasets for multilingual speech representation learning, ASR, Speech Translation and Speech-LangID. CommonVoice statistics as on 24th May 2022.*

- 17 Language Families
- 27 Writing Systems

Language Families

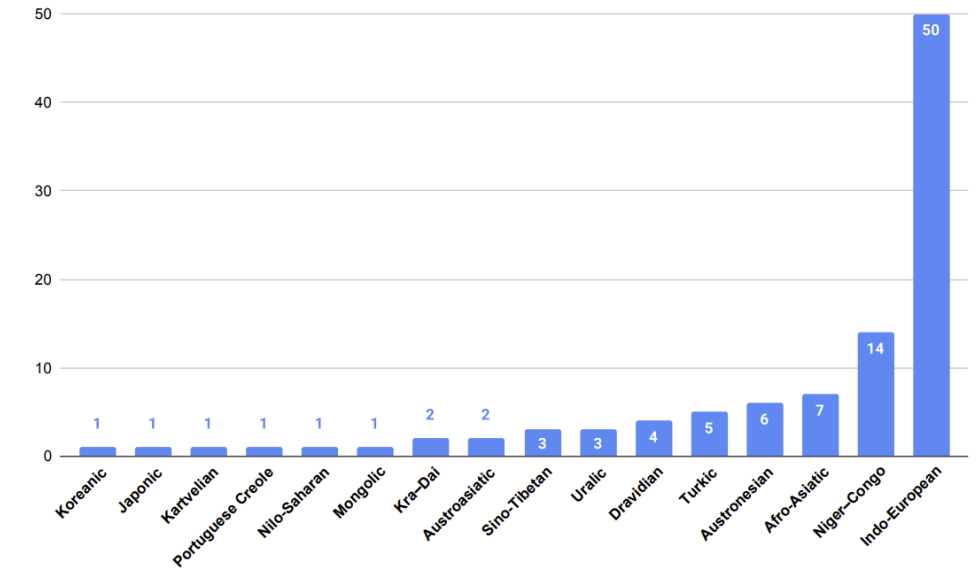


Figure 1: *Distributions of language families in FLEURS (y-axis is the count).*

Geographic Groupings

- Languages Grouped by Geographical Areas:
  - Western European (WE)
  - Eastern Europe (EE)
  - Central-Asia/MiddleEast/North-Africa (CMN)
  - Sub-Saharan Africa (SSA)
  - South Asia (SA)
  - South-East Asia (SEA)
  - Chinese, Japanese, and Korean (CJK) languages

Language Groupings

Table 3: *Statistics for speech and transcript data in FLEURS.*

Data Statistics	WE	EE	CMN	SSA	SA	SEA	CJK	All
train speech hours	231h	134h	116h	237h	124h	112h	32h	987h
dev speech hours	29h	18h	14h	24h	16h	14h	4h	120h
test speech hours	68h	43h	33h	58h	37h	35h	9h	283h
train transcript tokens	1475k	772k	630k	1072k	699k	525k	405k	5578k
dev transcript tokens	184k	107k	75k	116k	93k	65k	51k	692k
test transcript tokens	443k	260k	181k	272k	210k	158k	116k	1640k

Writing Systems

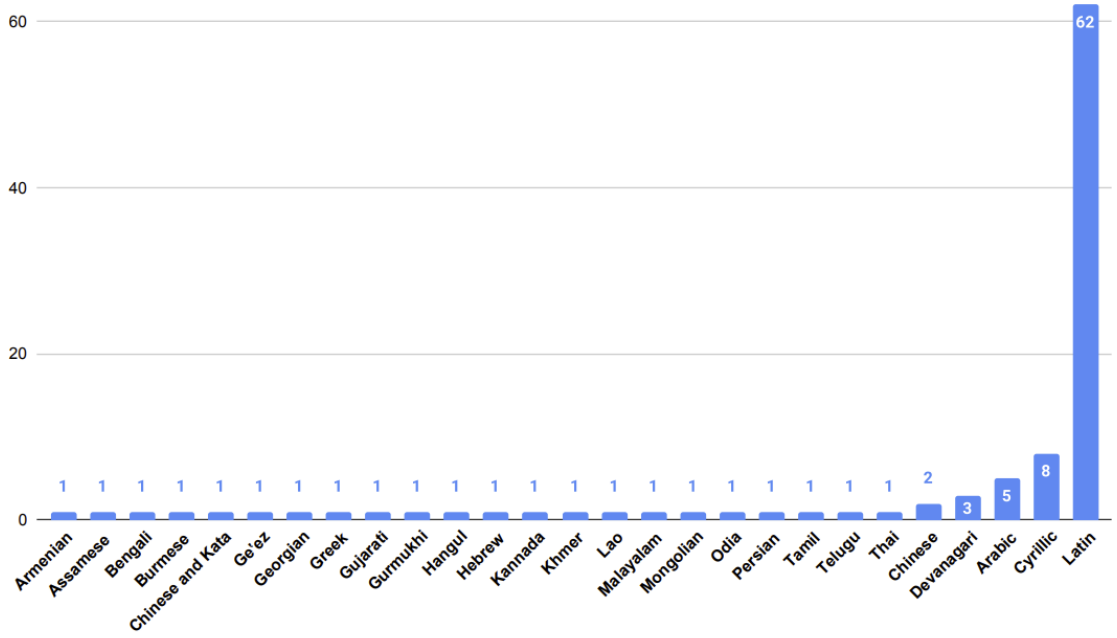


Figure 2: *Distributions of writing systems in FLEURS (y-axis is the count).*

Baselines

- Multilingual Pre-trained
  - Fine-Tune 600M wave2vec-BERT model
  - 429k hours unlabeled speech in 51 languages
  - VoxPopuli, MLS, CommonVoice, BABEL
- Multilingual Multimodel Pre-Trained
  - mSLAM 600M
  - 10TiB unlabeled speech in 101 langauges

54 Seen Languages

- **WE**: Catalan (ca), Croatian (hr), Danish (da), Dutch (nl), American English (en), Finnish (fi), French (fr), German (de), Greek (el), Hungarian (hu), Irish (ga), Italian (it), Latin American Spanish (es), Maltese (mt), Portuguese (pt), Swedish (sv), Welsh (cy)
- **EE**: Bulgarian (bg), Czech (cs), Estonian (et), Georgian (ka), Latvian (lv), Lithuanian (lt), Polish (pl), Romanian (ro), Russian (ru), Slovak (sk), Slovenian (sl), Ukrainian (uk)
- **CMN**: Arabic (ar), Kazakh (kk), Kyrgyz (ky), Mongolian (mn), Pashto (ps), Persian (fa), Tajik (tg), Turkish (tr)
- **SSA**: Ganda (lg), Swahili (sw), Zulu (zu)
- **SA**: Assamese (as), Bengali (bn), Hindi (hi), Oriya (or), Punjabi (pa), Tamil (ta), Telugu (te)
- **SEA**: Cebuano (ceb), Indonesian (id), Lao (lo), Thai (th), Vietnamese (vi)
- **CJK**: Cantonese (yue), Japanese (ja), Mandarin (cmn)

48 Unseen Languages

- **WE**: Asturian (ast), Bosnian (bs), Galician (gl), Icelandic (is), Kabuverdianu (kea), Luxembourgish (lb), Norwegian (nb), Occitan (oc)
- **EE**: Armenian (hy), Belarusian (be), Macedonian (mk), Serbian (sr)
- **CMN**: Azerbaijani (az), Hebrew (he), Sorani-Kurdish (ckb), Uzbek (uz)
- **SSA**: Afrikaans(af), Amharic (am), Fula (ff), Hausa (ha), Igbo (ig), Kamba (kam), Lingala (ln), Luo (luo), Northern-Sotho (nso), Nyanja (ny), Oromo (om), Shona (sn), Somali (so), Umbundu (umb), Wolof (wo), Xhosa (xh), Yoruba (yo)
- **SA**: Gujarati (gu), Kannada (kn), Malayalam (ml), Marathi (mr), Nepali (ne), Sindhi (sd), Urdu (ur)
- **SEA**: Filipino (fil), Javanese (jv), Khmer (km), Malay (ms), Maori (mi), Burmese (my)
- **CJK**: Korean (ko)

Character Error Rate

Table 4: ***Speech recognition*** *- Fleurs massively multilingual ASR baselines, reporting % CER (↓), by geographical group.*

Model	WE	EE	CMN	SSA	SA	SEA	CJK	Avg.
Speech recognition CER for all languages								
# languages	25	16	12	20	14	11	4	102
w2v-bert-51 (0.6B)	10.7	9.9	14.5	15.6	17.4	14.7	24.6	14.1
mSLAM (0.6B)	10.6	10.0	14.8	16.4	19.2	14.9	25.0	14.6

LangID

Table 5: ***Speech identification*** *- FLEURS langID baselines, reporting % accuracy (↑), by geographical group.*

Model	WE	EE	CMN	SSA	SA	SEA	CJK	Avg.
Speech identification accuracy for all languages								
# languages	25	16	12	20	14	11	4	102
w2v-bert-51 (0.6B)	85.3	78.4	72.9	59.1	52.0	65.7	89.7	71.4
mSLAM (0.6B)	84.6	81.3	75.9	62.2	51.7	73.4	87.8	73.3

Cross-Modal Retrieval

Table 6: ***Cross-modal Speech-Text Retrieval*** *- FLEURS massively multilingual Speech-to-Text and Text-to-Speech retrieval baselines, reporting % P@1 (↑) score, by geographical group.*

Task	WE	EE	CMN	SSA	SA	SEA	CJK	Avg.
P@1 for all languages								
# languages	25	16	12	20	14	11	4	102
Speech-to-Text Retrieval	87.6	91.1	79.4	83.9	67.7	54.8	4.7	76.9
Text-to-Speech Retrieval	83.7	88.3	77.1	83.5	61.4	55.4	4.7	74.4

No Language Left Behind

No Language Left Behind: Scaling Human-Centered Machine Translation

NLLB Team, Marta R. Costa-jussà[*], James Cross[*], Onur Çelebi[*], Maha Elbayad[*], Kenneth Heafield[*], Kevin Heffernan[*], Elahe Kalbassi[*], Janice Lam[*], Daniel Licht[*], Jean Maillard[*], Anna Sun[*], Skyler Wang[*,§], Guillaume Wenzek[*], Al Youngblood[*]
Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran
Pierre Andrews[†], Necip Fazil Ayan[†], Shruti Bhosale[†], Sergey Edunov[†], Angela Fan[†,‡], Cynthia Gao[†], Vedanuj Goswami[†], Francisco Guzmán[†], Philipp Koehn[†,¶], Alexandre Mourachko[†], Christophe Ropers[†], Safiyyah Saleem[†], Holger Schwenk[†], Jeff Wang[†]

Meta AI, [§]UC Berkeley, [¶]Johns Hopkins University

FLoRes → FLORES

The innovation

We extended 2x the coverage of FLORES, a human-translated evaluation benchmark, to now cover 200 languages. Through automatic metrics and human evaluation support, we’re able to extensively quantify the quality of our translations.

“The Journey”

- https://ai.facebook.com/research/no-language-left-behind/#research-milestones

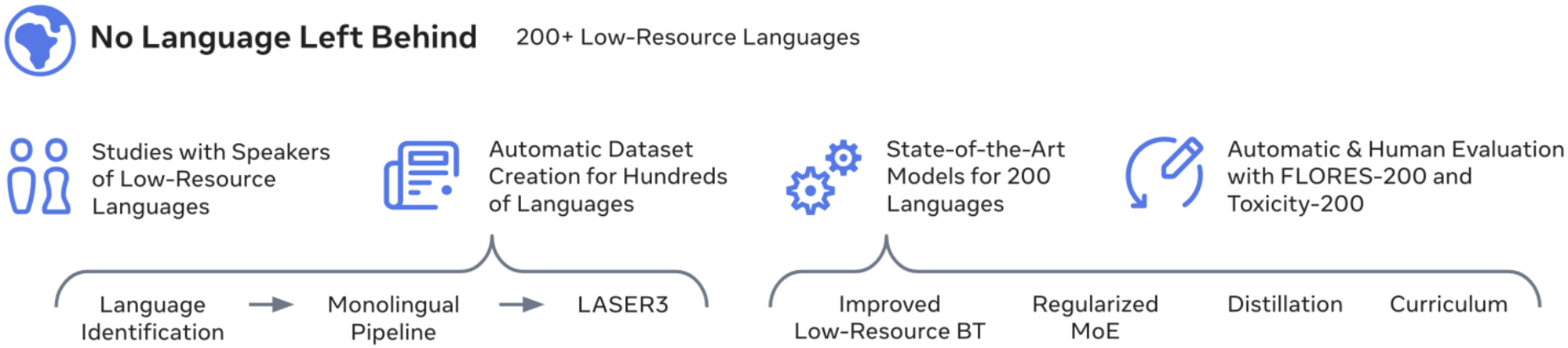


Figure 1: **No Language Left Behind:** Our low-resource translation effort focuses on four cornerstones. (1) We strive to understand the low-resource translation problem from the perspective of native speakers. (2) We study how to automatically create training data to move low-resource languages towards high-resource. (3) We utilize this data to create state-of-the-art translation models. (4) We evaluate every language we aim to translate.

Open-Source Releases

- **Human-Translated Datasets**

  FLORES-200: Evaluation dataset in 204 languages

  NLLB-SEED: Seed training data in 39 languages

  NLLB-MD: Seed data in different domains in 6 languages to assess generalization

  Toxicity-200: wordlists to detect toxicity in 200 languages

- **Tools to Create Large Scale Bitext Datasets**

  Language Identification for more than 200 languages

  LASER3: sentence encoders for identifying aligned bitext for 148 languages

  `stopes`: a data mining library that can be used to process and clean monolingual data, then create aligned bitext

  Training data recreation: Scripts that recreate our training data

- **Translation Models covering 202 languages**

  NLLB-200: A 54.5B Sparsely Gated Mixture-of-Experts model

  3.3B and 1.3B Dense Transformer models

  1.3B and 600M Dense transformer models distilled from NLLB-200

  Training and generation scripts to reproduce our models

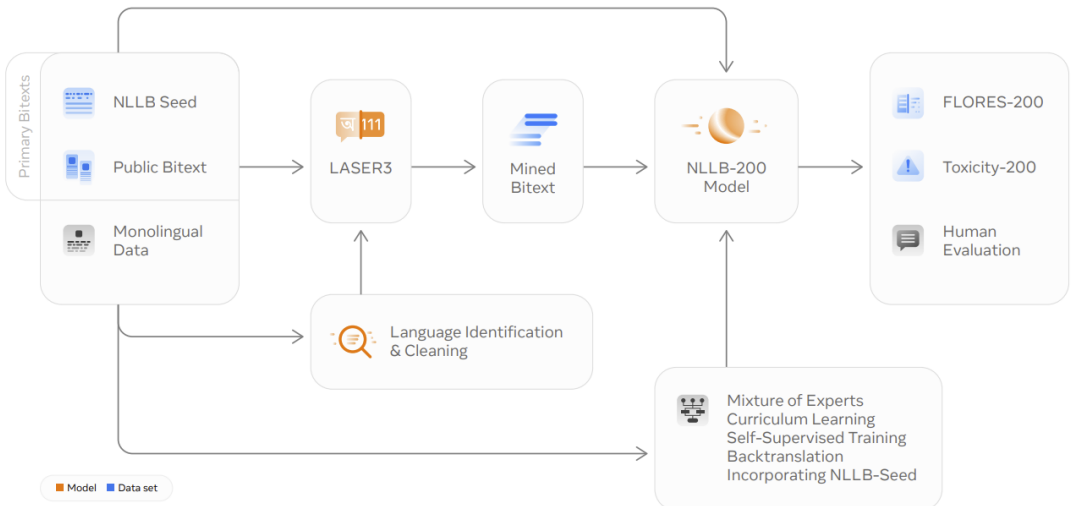


Figure 2: **How the Pieces Fit Together, a Bird's-Eye View:** We depict the technical components of No Language Left Behind and how they fit together. We display the interaction between data, how data is utilized in the models we develop (orange), and how models are evaluated. Datasets shown in blue are novel datasets created in No Language Left Behind.

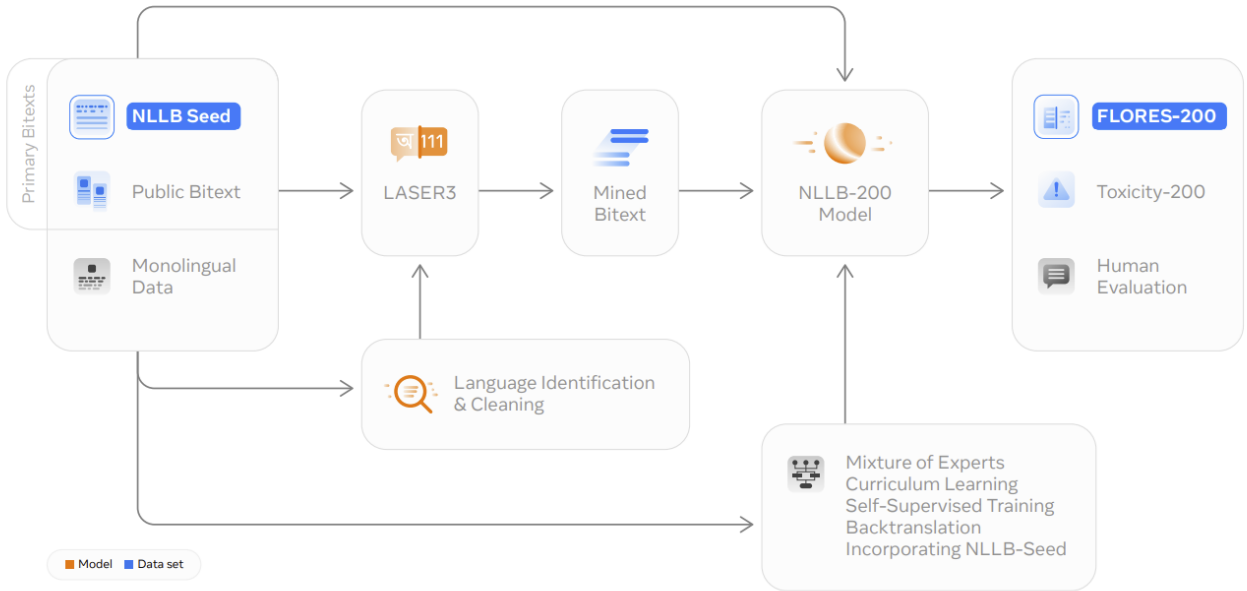


Figure 3: **Human-Translated Dataset Contributions of No Language Left Behind**: As highlighted, these datasets enable model training and evaluation.

Workflow (aka different colored image)

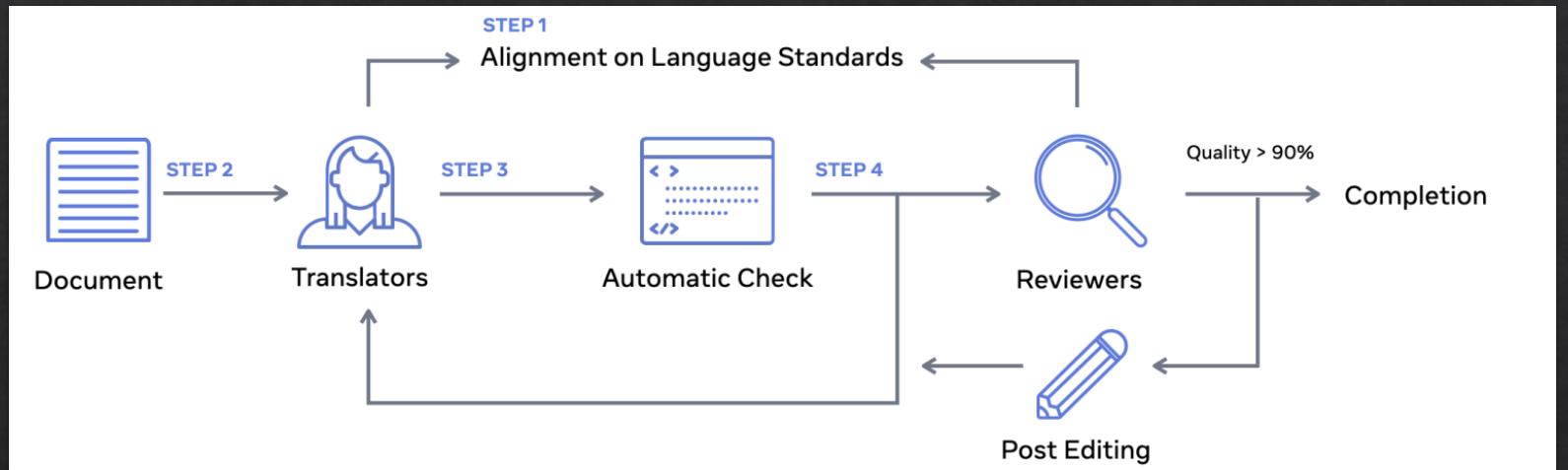


Figure 4: **FLORES-200 Translation Workflow:** We created a complex, multi-step process to ensure quality. First, professional translators and reviewers aligned on language standards. Next, translators translated the full set of FLORES-200 sentences, followed by automated checks. Subsequently, the group of independent reviewers reviewed the quality, and based on their assessment, we sent some translations out for post-editing. If the quality assessment indicated that the quality is above 90 percent, the language is considered ready for inclusion in FLORES-200.

Quality of FLORES-200

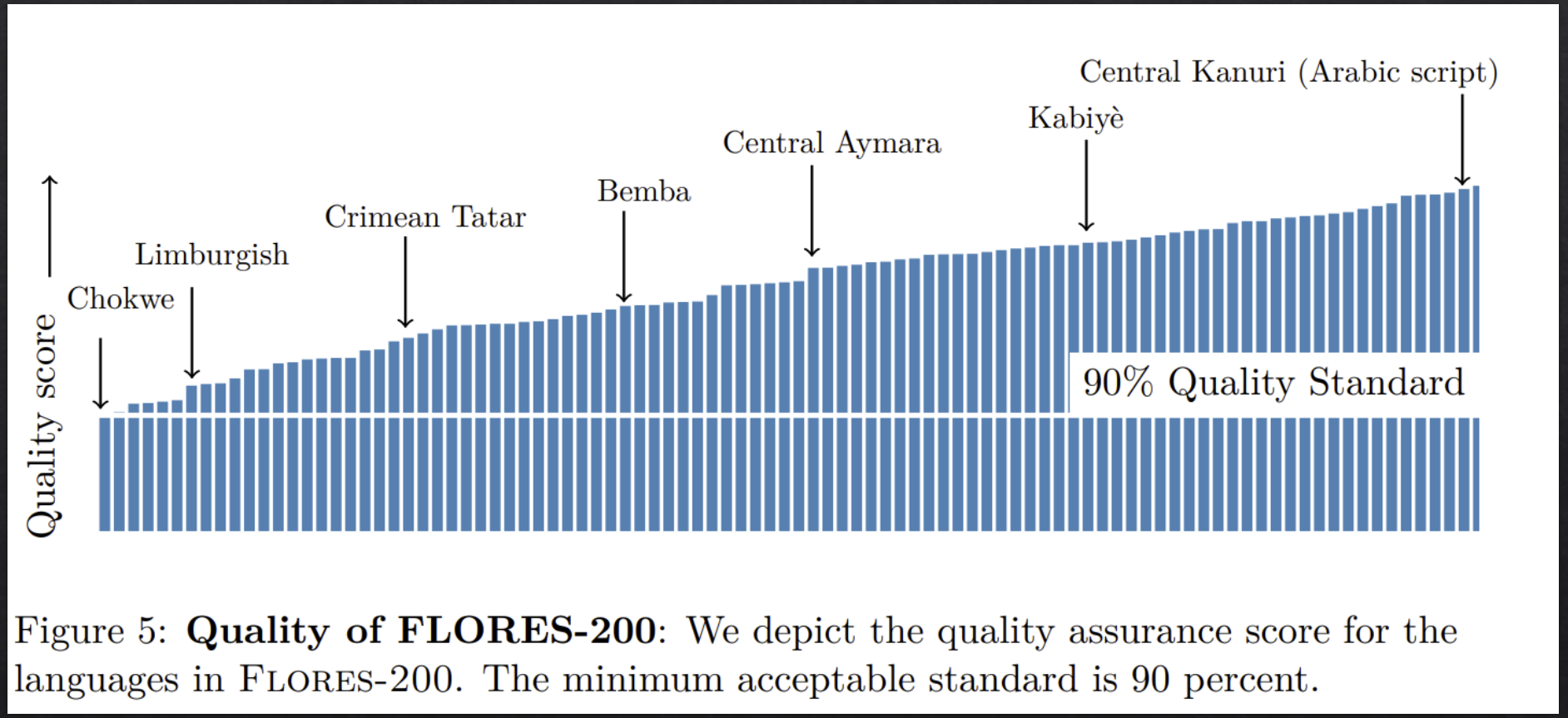


Figure 5: **Quality of FLORES-200**: We depict the quality assurance score for the languages in FLORES-200. The minimum acceptable standard is 90 percent.

FLORES-200

- Flores-200 consists of translations from 842 distinct web articles
- 3001 sentences
- dev, devtest, and test (hidden)
- On average, sentences are approximately 21 words long.

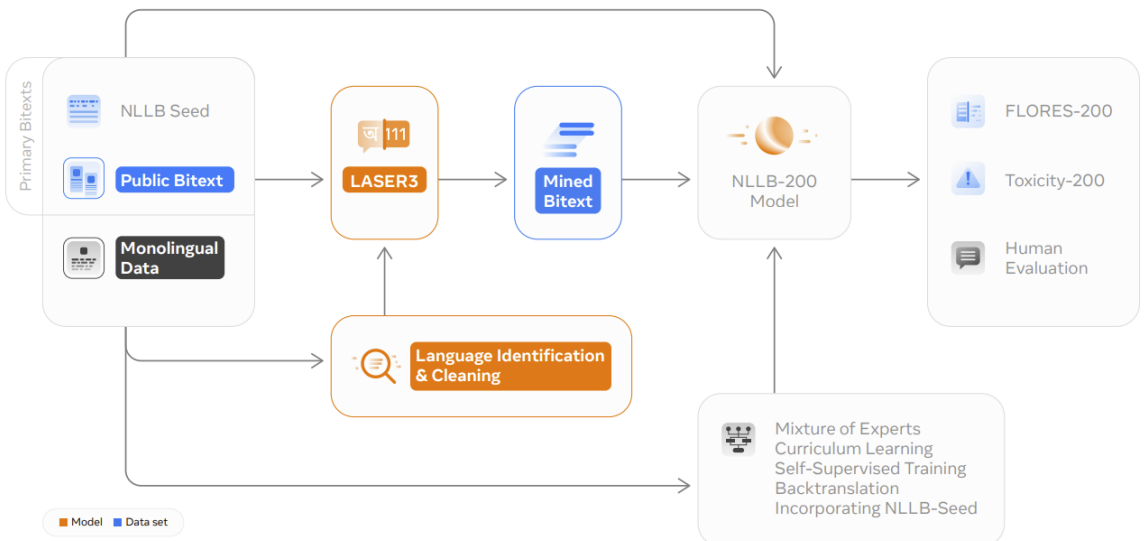


Figure 6: **Automatic Dataset Creation Contributions of No Language Left Behind**: As highlighted, we create language identification and a monolingual data cleaning process, then describe the training of LASER3 to produce large-scale mined bitext for hundreds of languages.

LangID

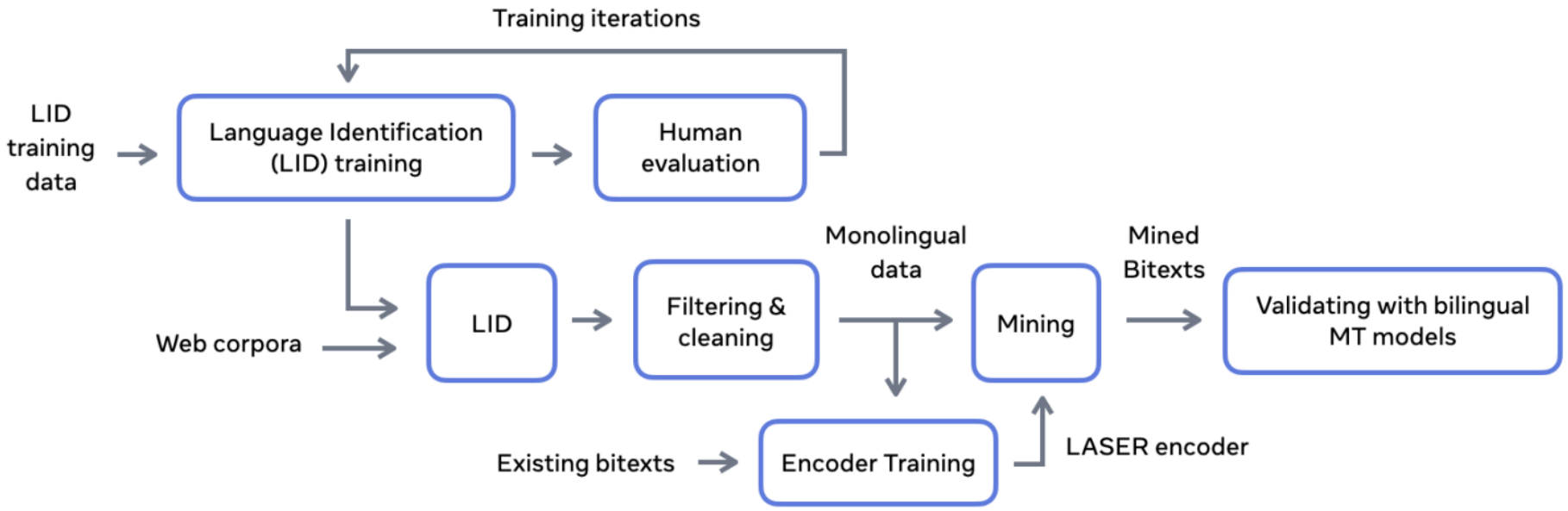
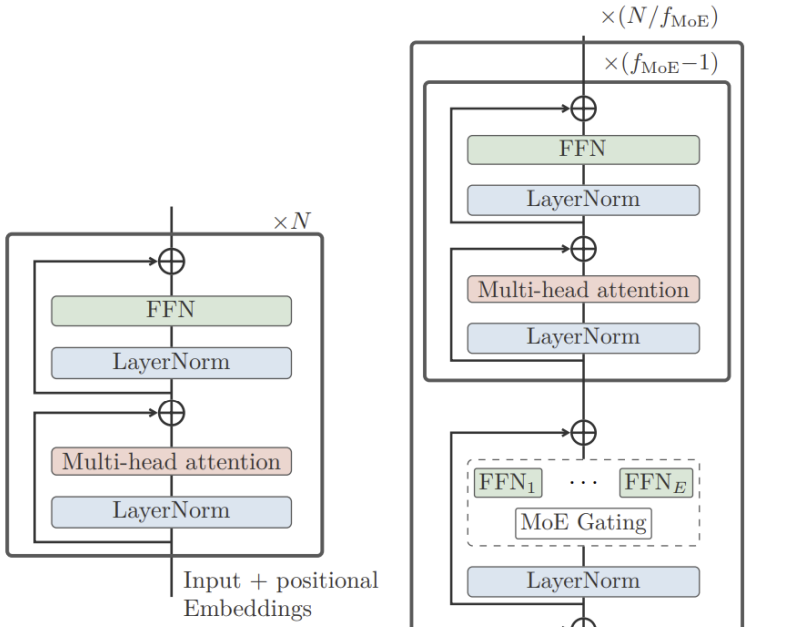


Figure 7: **Overview of our Bitext Mining Pipeline.** Language identification is applied on web corpora to extract monolingual sentences. Aligned pairs are later identified with LASER3.

Mixture of Experts



(a) Dense Transformer

(b) MoE Transformer

Figure 16: **Illustration of a Transformer encoder with MoE layers** inserted at a 1:f_{MoE} frequency. Each MoE layer has E experts and a gating network responsible for dispatching tokens.