A 1000-language Collaborative Universal Dictionary and Universal Translator

David Yarowsky, Arya D. McCarthy, Garrett Nicolai, Winston Wu, Aaron Mueller, Dylan Lewis,
Yingqi Ding, Abhinav Nigam, Emre Ozgu, Debanik Purkayastha, James Scharf and Kenneth Zheng

Johns Hopkins University

yarowsky@jhu.edu

We present JHU's Universal Dictionary and Universal Translator, covering 1000+ world languages in a broadly-accessible Android/IOS mobile phone and web-browser app, with 1,000,000+ planet-wide language pairs and 100's of under-resourced languages which have never had access to a substantial dictionary or machine translation capability. In addition to providing immediate access to a base vocabulary of 1500-20000 core vocabulary lemmas in all 1000+ languages, this novel app actively engages its users to contribute collaboratively to the universal dictionary in an easy-to-use and efficient way, with automatic suggestion of possible translations based on sound-shift transductions from related languages and pan-linguistic compositional constructions.

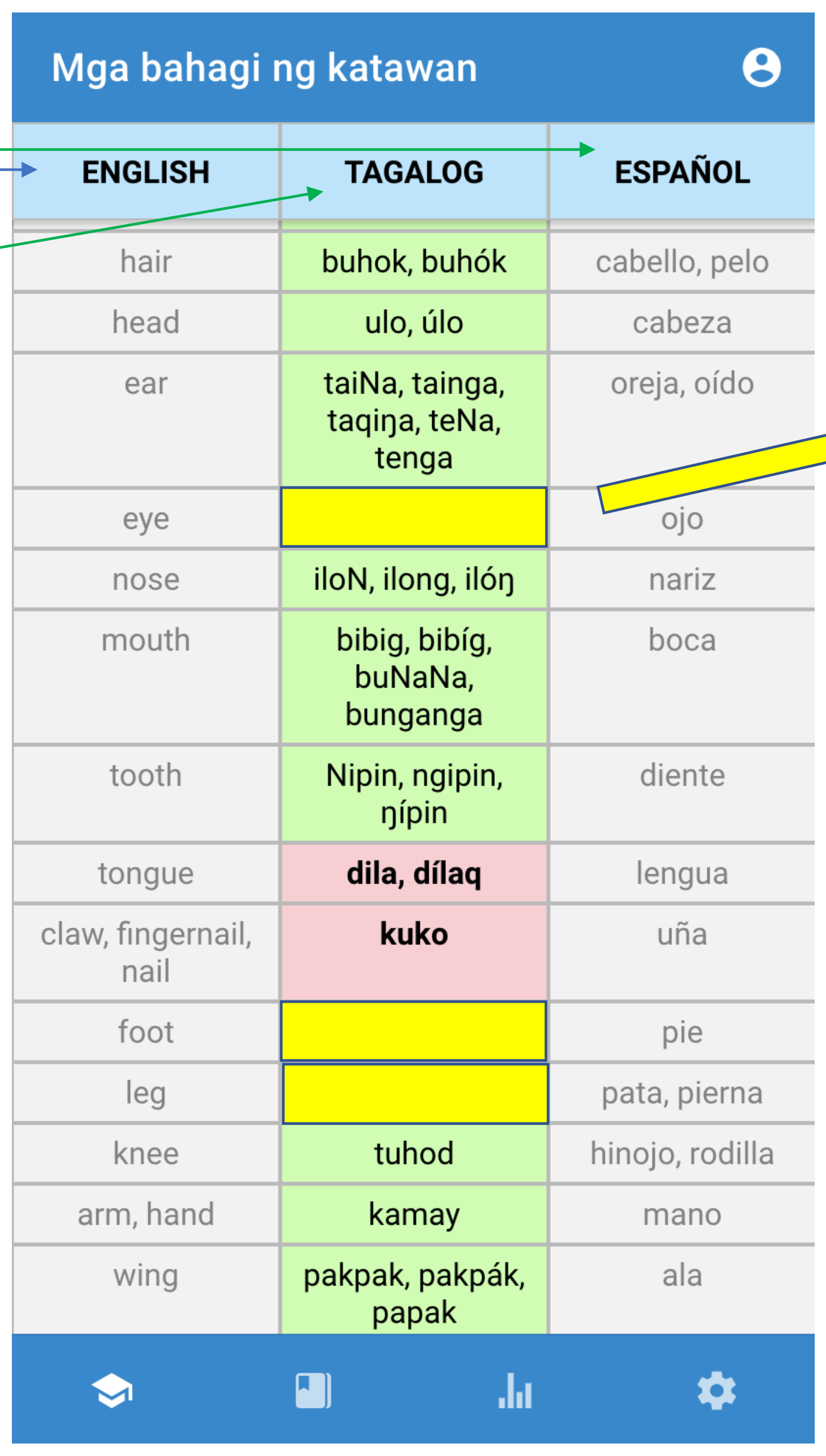
Collaborative 1000+ Language Dictionary Augmentation (Android App):

Users typically choose 1-2 **reference languages** familiar to them and 1 **target/focus language** (e.g. Tagalog or Cebuano).

Users type their entries either directly in the tables (like a Google doc) or drill down to see other languages, pictures and definitions.

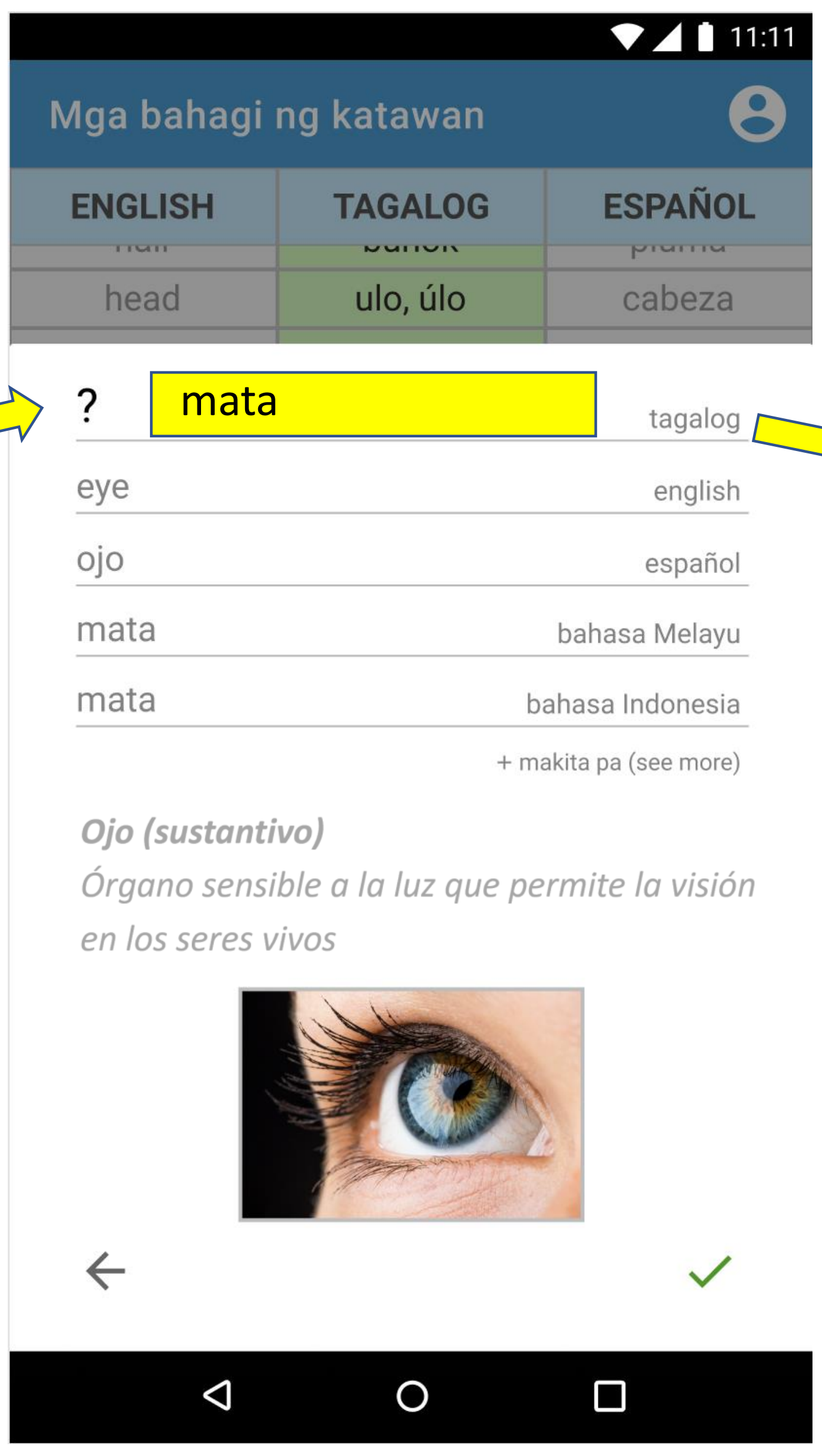
Filtering and user accuracy testing (via held-out known translations) to increase quality and reduce vandalism.

Initial dictionaries auto-populated for 1000+ languages as below.



Mga bahagi ng katawan

ENGLISH	TAGALOG	ESPAÑOL
hair	buhok, buhók	cabello, pelo
head	ulo, úlo	cabeza
ear	taiNa, tainga, taqiŋa, teNa, tenga	oreja, oído
eye		ojo
nose	iloN, ilong, ilóŋ	nariz
mouth	bibig, bibíg, buNaNa, bunganga	boca
tooth	Nipin, ngipin, ŋípin	diente
tongue	**dila, dílaq**	lengua
claw, fingernail, nail	**kuko**	uña
foot		pie
leg		pata, pierna
knee	tuhod	hinojo, rodilla
arm, hand	kamay	mano
wing	pakpak, pakpák, papak	ala



11:11

Mga bahagi ng katawan

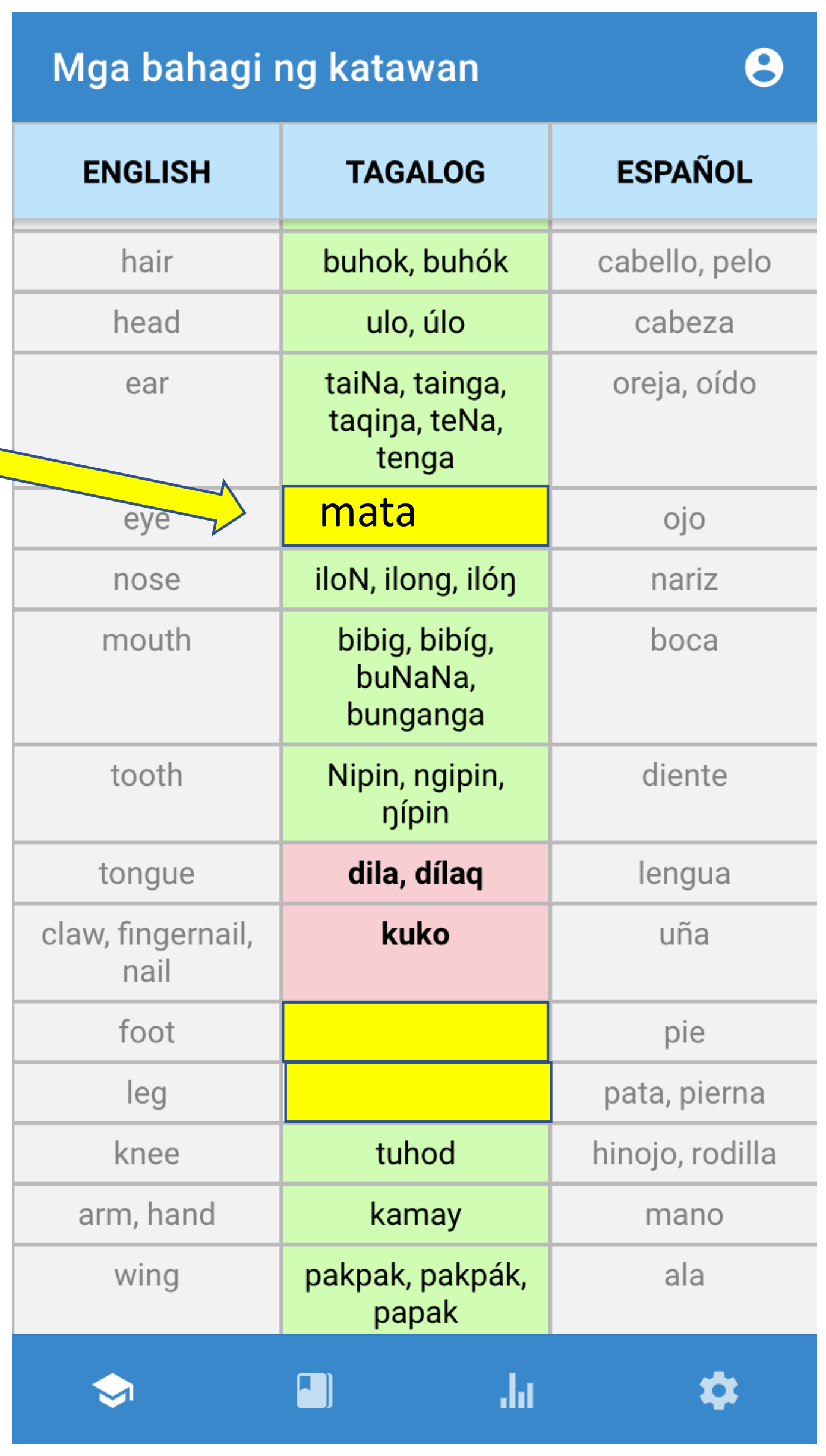
ENGLISH	TAGALOG	ESPAÑOL
head	ulo, úlo	cabeza

? mata — tagalog

eye	english
ojo	español
mata	bahasa Melayu
mata	bahasa Indonesia

+ makita pa (see more)

Ojo (sustantivo)
Órgano sensible a la luz que permite la visión en los seres vivos

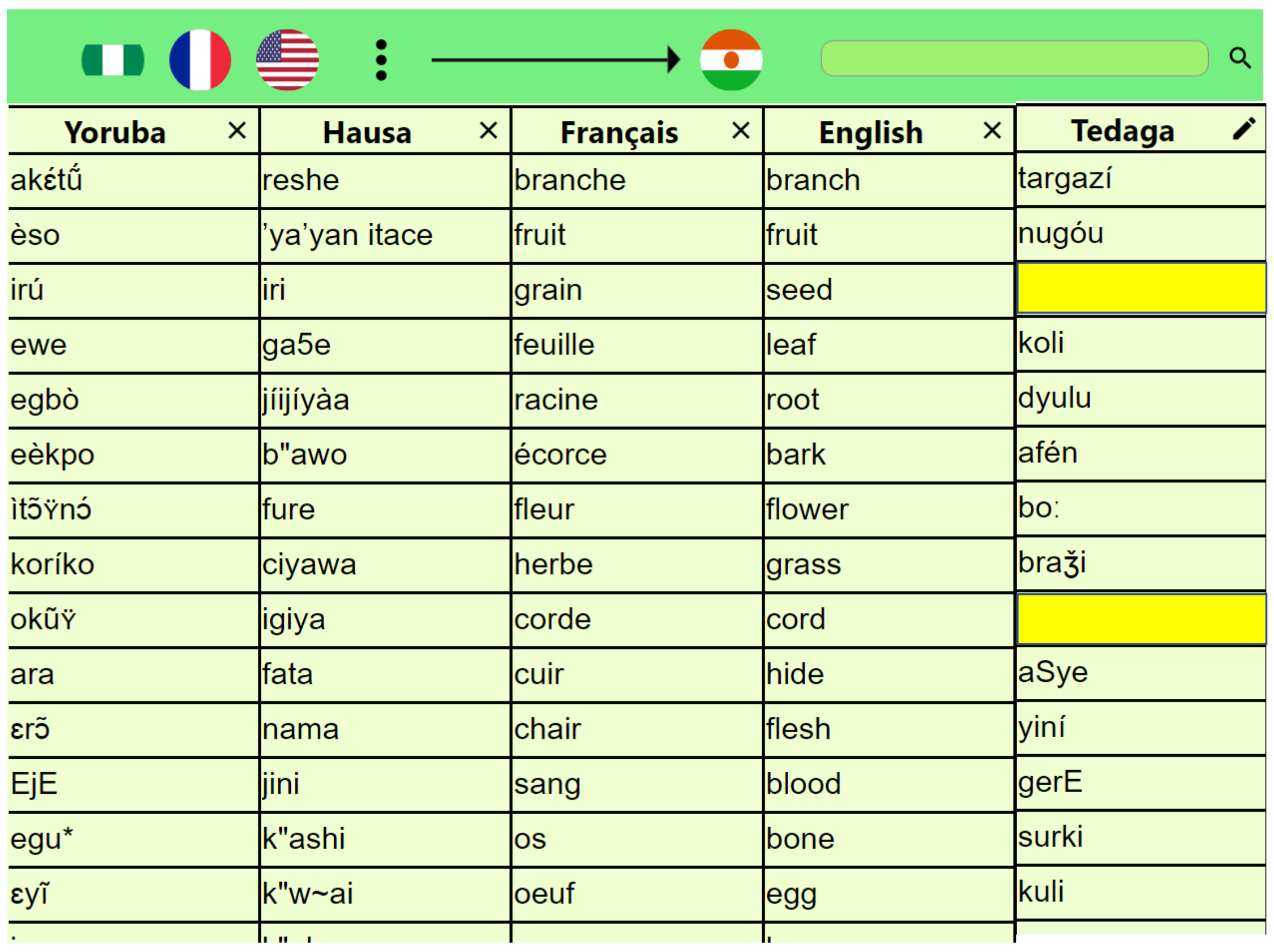


Mga bahagi ng katawan

ENGLISH	TAGALOG	ESPAÑOL
hair	buhok, buhók	cabello, pelo
head	ulo, úlo	cabeza
ear	taiNa, tainga, taqiŋa, teNa, tenga	oreja, oído
eye	mata	ojo
nose	iloN, ilong, ilóŋ	nariz
mouth	bibig, bibíg, buNaNa, bunganga	boca
tooth	Nipin, ngipin, ŋípin	diente
tongue	**dila, dílaq**	lengua
claw, fingernail, nail	**kuko**	uña
foot		pie
leg		pata, pierna
knee	tuhod	hinojo, rodilla
arm, hand	kamay	mano
wing	pakpak, pakpák, papak	ala

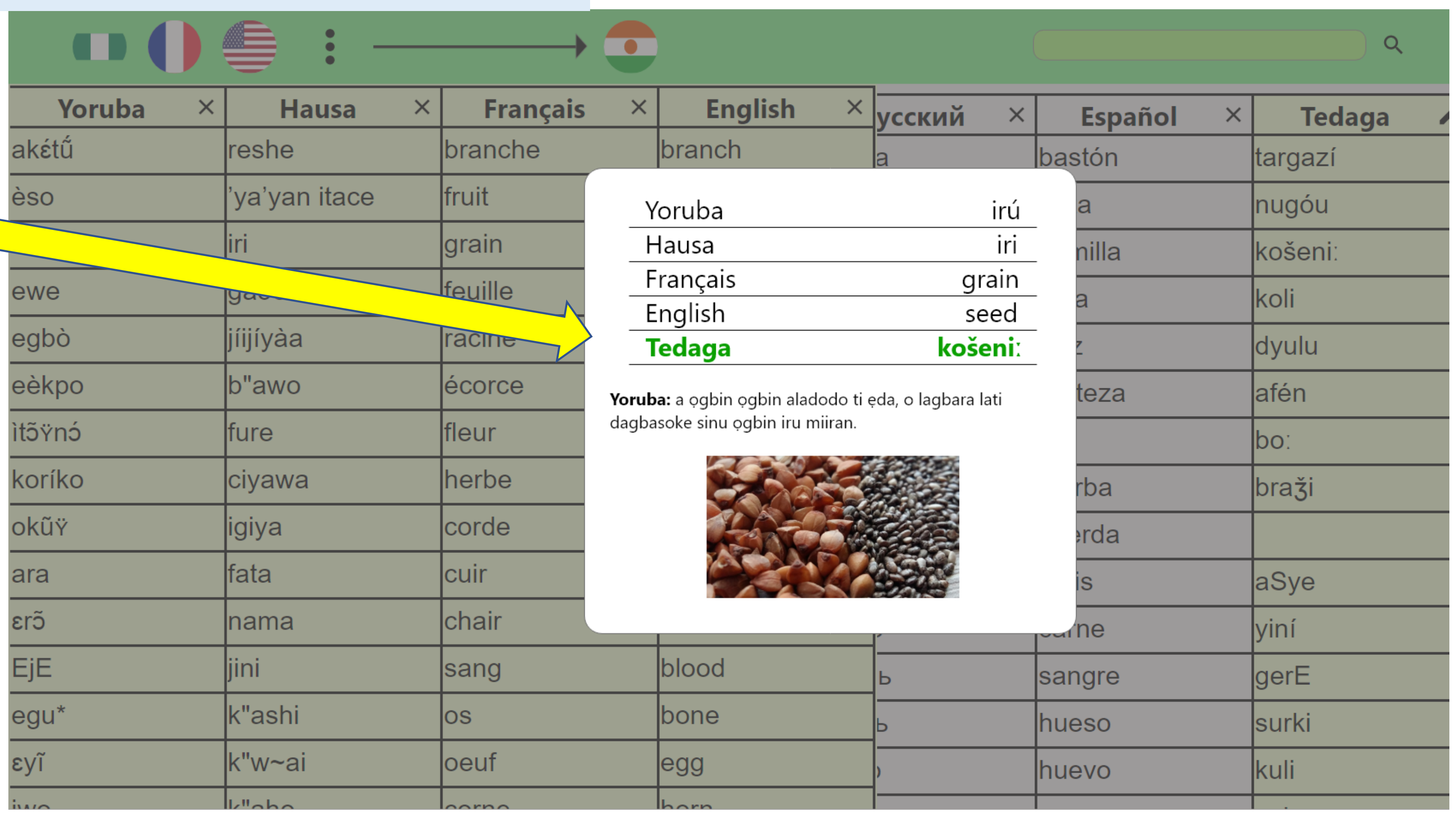
Collaborative 1000+ Language Dictionary Augmentation (Web App):

4 reference languages familiar in Nigeria
1 Target/focus language (Tedaga)



Yoruba ✕	Hausa ✕	Français ✕	English ✕	Tedaga
akɛ́tű̃	reshe	branche	branch	targazí
èso	'ya'yan itace	fruit	fruit	nugóu
irú	iri	grain	seed	
ewe	ga5e	feuille	leaf	koli
egbò	jíijíyàa	racine	root	dyulu
eèkpo	b"awo	écorce	bark	afén
ìtɔ̃ÿnɔ́	fure	fleur	flower	bo:
koríko	ciyawa	herbe	grass	braǯi
okũÿ	igiya	corde	cord	
ara	fata	cuir	hide	aSye
ɛrɔ̃	nama	chair	flesh	yiní
EjE	jini	sang	blood	gerE
egu*	k"ashi	os	bone	surki
ɛyĩ	k"w~ai	oeuf	egg	kuli

Drill-down window (with photo etc.)



Yoruba ✕	Hausa ✕	Français ✕	English ✕	усский ✕	Español ✕	Tedaga
akɛ́tű̃	reshe	branche	branch	a	bastón	targazí
èso	'ya'yan itace	fruit			a	nugóu
	iri	grain			nilla	košeni:
ewe		feuille			a	koli
egbò	jíijíyàa				z	dyulu
eèkpo	b"awo	écorce			teza	afén
ìtɔ̃ÿnɔ́	fure	fleur				bo:
koríko	ciyawa	herbe			rba	braǯi
okũÿ	igiya	corde			rda	
ara	fata	cuir			is	aSye
ɛrɔ̃	nama	chair			ne	yiní
EjE	jini	sang	blood	ь	sangre	gerE
egu*	k"ashi	os	bone	ь	hueso	surki
ɛyĩ	k"w~ai	oeuf	egg		huevo	kuli

Yoruba	irú
Hausa	iri
Français	grain
English	seed
Tedaga	**košeni:**

Yoruba: a ọgbin ọgbin aladodo ti ẹda, o lagbara lati dagbasoke sinu ọgbin iru miiran.

Initial Dictionary Population (for 1000+ languages)

- Swadesh list
- Panlex
- Wiktionary
- Unimorph
- Universal Declaration of Human Rights
- Bible bitext (~3000+ core dictionary entries)

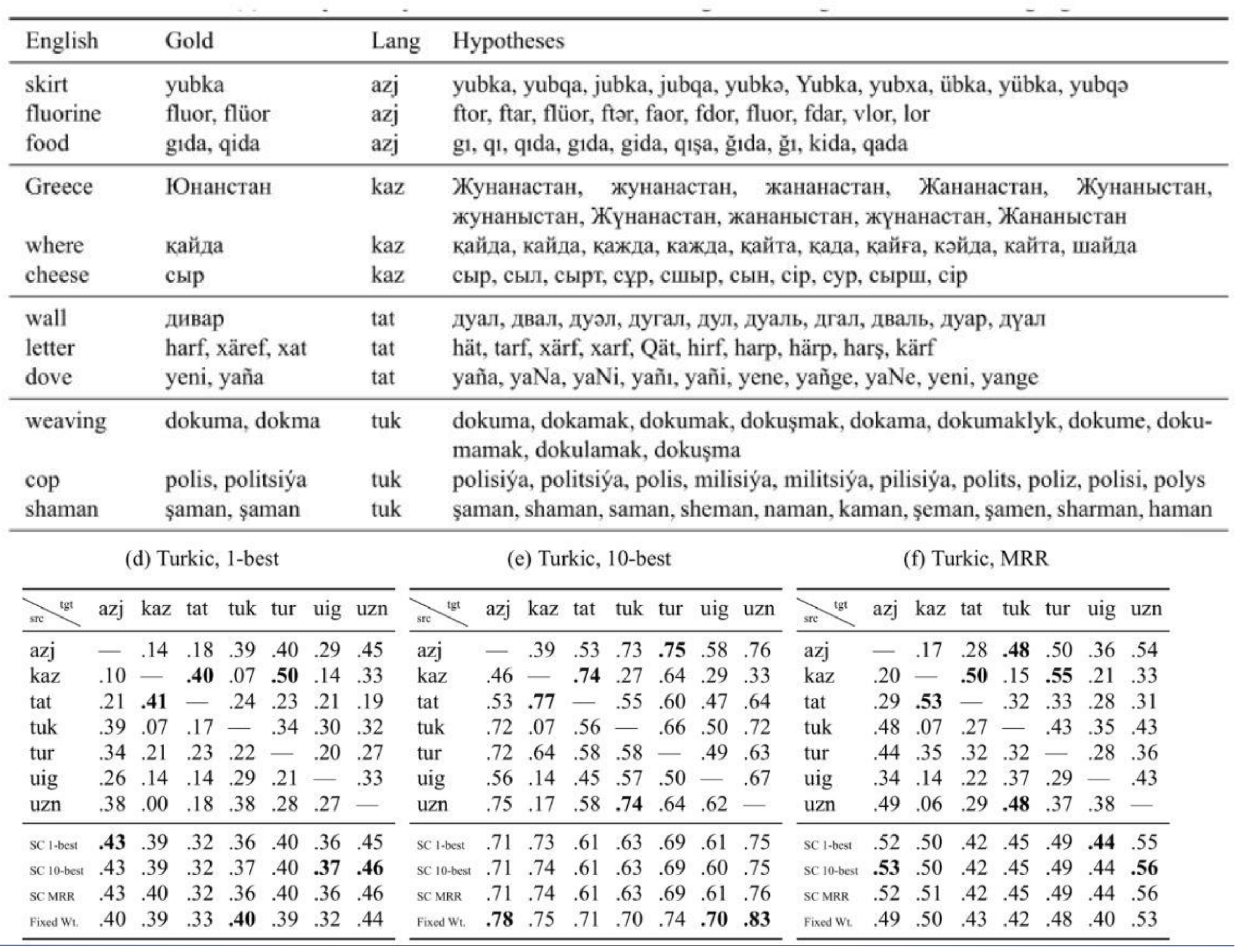
Multiple models of word formation:

(1) cognate and sound shift models
(2) borrowing models
(3) derivational models (i.e. compositional translation of digger from dig +V:N(AGT) [and +V:N(INST)]
(4) universalized compositional models:

hospital = sick|sickness|disease + house|place|institution
hospital = doctor|medicine + house|place|institution

da	syge|hus	=> hospital
sv	sjuk|hus	=> hospital
hu	kór|házi	=> hospital
hu	kór|ház	=> hospital
no	syke|hus	=> hospital
de	kranken|haus	=> hospital
nl	gast|huis	=> hospital
zh	病院	=> hospital

Automatic compositional prediction:



English	Gold	Lang	Hypotheses
skirt	yubka	azj	yubka, yubqa, jubka, jubqa, yubkə, Yubka, yubxa, übka, yübka, yubqə
fluorine	fluor, flüor	azj	ftor, ftar, flüor, ftər, faor, fdor, fluor, fdar, vlor, lor
food	gıda, qida	azj	gı, qı, qıda, gıda, gida, qışa, ğıda, ğı, kida, qada
Greece	Юнанстан	kaz	Жунанастан, жунанастан, жананастан, Жананастан, Жунаныстан, жунаныстан, Жүнанастан, жананыстан, жүнанастан, Жананыстан
where	қайда	kaz	қайда, кайда, қажда, кажда, қайта, қада, қайға, кәйда, кайта, шайда
cheese	сыр	kaz	сыр, сыл, сырт, сұр, сшыр, сын, сір, сур, сырш, сір
wall	дивар	tat	дуал, двал, дуәл, дугал, дул, дуаль, дгал, дваль, дуар, дұал
letter	harf, xäref, xat	tat	hät, tarf, xärf, xarf, Qät, hirf, harp, härp, harş, kärf
dove	yeni, yaña	tat	yaña, yaNa, yaNi, yañı, yañi, yene, yañge, yaNe, yeni, yange
weaving	dokuma, dokma	tuk	dokuma, dokamak, dokumak, dokuşmak, dokama, dokumaklyk, dokume, doku-mamak, dokulamak, dokuşma
cop	polis, politsiýa	tuk	polisiýa, politsiýa, polis, milisiýa, militsiýa, pilisiýa, polits, poliz, polisi, polys
shaman	şaman, şaman	tuk	şaman, shaman, saman, sheman, naman, kaman, şeman, şamen, sharman, haman

(d) Turkic, 1-best

src tgt	azj	kaz	tat	tuk	tur	uig	uzn
azj	—	.14	.18	.39	.40	.29	.45
kaz	.10	—	**.40**	.07	**.50**	.14	.33
tat	.21	**.41**	—	.24	.23	.21	.19
tuk	.39	.07	.17	—	.34	.30	.32
tur	.34	.21	.23	.22	—	.20	.27
uig	.26	.14	.14	.29	.21	—	.33
uzn	.38	.00	.18	.38	.28	.27	—
SC 1-best	**.43**	.39	.32	.36	.40	.36	.45
SC 10-best	.43	.39	.32	.37	.40	**.37**	**.46**
SC MRR	.43	.40	.32	.36	.40	.36	.46
Fixed Wt.	.40	.39	.33	**.40**	.39	.32	.44

(e) Turkic, 10-best

src tgt	azj	kaz	tat	tuk	tur	uig	uzn
azj	—	.39	.53	.73	**.75**	.58	.76
kaz	.46	—	**.74**	.27	.64	.29	.33
tat	.53	**.77**	—	.55	.60	.47	.64
tuk	.72	.07	.56	—	.66	.50	.72
tur	.72	.64	.58	.58	—	.49	.63
uig	.56	.14	.45	.57	.50	—	.67
uzn	.75	.17	.58	**.74**	.64	.62	—
SC 1-best	.71	.73	.61	.63	.69	.61	.75
SC 10-best	.71	.74	.61	.63	.69	.60	.75
SC MRR	.71	.74	.61	.63	.69	.61	.76
Fixed Wt.	**.78**	.75	.71	.70	.74	**.70**	**.83**

(f) Turkic, MRR

src tgt	azj	kaz	tat	tuk	tur	uig	uzn
azj	—	.17	.28	**.48**	.50	.36	.54
kaz	.20	—	**.50**	.15	**.55**	.21	.33
tat	.29	**.53**	—	.32	.33	.28	.31
tuk	.48	.07	.27	—	.43	.35	.43
tur	.44	.35	.32	.32	—	.28	.36
uig	.34	.14	.22	.37	.29	—	.43
uzn	.49	.06	.29	**.48**	.37	.38	—
SC 1-best	.52	.50	.42	.45	.49	**.44**	.55
SC 10-best	**.53**	.50	.42	.45	.49	.44	**.56**
SC MRR	.52	.51	.42	.45	.49	.44	.56
Fixed Wt.	.49	.50	.43	.42	.48	.40	.53

Universal Translator: Pre-train 1000 full-text translation models (to English Pivot) using above and Bible bitext.