András Kornai and Katalin Pajkossy

Background The Hungarian Academy of Sciences (HAS) is currently under contract to the Summer Institute of Linguistics (SIL) for developing a Digital Vitality Index (DVI) that measures the vitality of languages in the digital realm based largely on the methodology of Kornai, 2013. That work presented a broad classification into four classes, digitally thriving (T), vital (V), heritage (H), and still (S) languages, and already noted some discrepancies between the SIL data and the data Kornai obtained by means of crawling the Internet. The goal of this Report is to lay bare the principles used in resolving such discrepancies. We follow Hammarström, 2015 not so much in his specific criticisms as in his general goal of creating a transparent, objective, replicable algorithm for deciding the many issues that we encounter in practice.

1. To list or not to list Let us denote the set of Ethnologue codes (whereby we mean primarily 639-3 codes) by S, and the set of provisional HAS codes by H. From the mathematical perspective there can be only two kinds of discrepancies: in $S \setminus H$ we find those codepoints that SIL lists but HAS has no data for, and in $H \setminus S$ we find those languages for which HAS assigned a provisional code because the Ethnologue had no information. The first kind of error is practically nonexistent, because the HAS data collection effort included crawling the Ethnologue website, so if SIL has published data HAS has it (disregarding the possibility of bugs in the crawler). But the other source of error is quite significant, over 5% of H datapoints has no correspondent in S. Before going further, we illustrate the phenomenon on a class of languages that HAS has some familiarity with, Uralic (urj).

In 53 urj language candidates, S lacked codes for Seto (a dialect of Estonian); for Yazva (also known as Yodzyak) in the Komi subfamily; for the Mordvin subfamily as such (but with codes for the main varieties Erzya (myv) and Moksha (mdf). The opposite problem is seen in Nenets (yrk) which has a group code, but no individual codes for the main varieties, Tundra and Forest Nenets. Another set of languages that lacks codes involves extinct varieties such as Meshcherian, Muromian, and Kainu (Sami). For a systematic analysis, we need to consider these issues separately. We will continue using the Uralic examples, but the phenomena are obviously not restricted to this family.

1.1 Extinct languages The coverage of extinct languages in the Ethnologue is very uneven. We have found many S languages which actually died out before 1950, so the methodological principle of drawing the line at this date is no longer tenable. We propose to gradually extend 639-3 codes as reliable, editorially controlled data on such languages becomes available.

On the whole, the HAS effort is not the best source of such data, in that the digital presence of such data is minimal: the general mechanism should take Glottolog as primary, and HAS as secondary. In particular, SIL, as the maintainer of 639-3 should set up an editorial pipeline

to assimilate Linguist List and Glottolog data. Once this is in place, HAS could feed its own findings to the same pipeline. HAS is actually happy to help SIL set up a modern 'bug tracking system' to handle these cases.

While this does not solve all problems relating to extinct languages (for example Yazva or Kainu have Glottolog codes yazv1241 and kain1277 but no substantive information as of yet, while Meshcherian and Muromian don't even have this much), it is expected to lessen the problem considerably, especially for those languages where digital data is to some extent available. We note that Appendix A to Hammarström, 2015 lists over 500 missing languages, the vast majority of which are extinct. The languages listed in Section 2 have, at least for the purposes of the SIL-HAS effort, higher priority, in that we have some digital data for these, while generally only secondary sources and unsubstantiated word-lists exist for the extinct languages noted by Hammarström, 2015.

2. Systematic inventory of $H \setminus S$

- 2.1 Crubadan There is only one language missing from S in the Crubadan crawl (Scannell, 2007), Elfdalian. This has Linguist List Code qer, and The Ethnologue considers it a dialect of Swedish, even though it notes "Dalecarlian spoken in northern Dalarna Province by about 10,000 speakers. Many would actually consider this variety a language in its own right, with its own literary standard and features that are markedly different from standard Swedish. Elfdalian is considered the most archaic vernacular within Dalecarlian, preserving many features of Old Norse." Obviously, HAS has no stake in declaring Elfdalian a separate language, but a clear policy of encoding dialects (perhaps by a unique identifier outside 639-3) would be extremely useful. This matter will come up with far greater force in 2.3, where it affects hundreds of languages/dialects.
- 2.2 Languages and scripts Writing systems, both classical scripts and the 'input modes' offered in the major operating systems (MacOS, Windows, and Linux) are closely tied to languages. We not that the same language is often written in multiple scripts (this is generally a significant factor slowing down the digital ascent of a language, especially when elementary schooling involves periods of using one script followed by a change in script, as happened to several Uralic languages). Conversely, different (often genealogically distinct) languages are written by the same script. That said, for the most part it is not hard to associate a dominant language to each script, and a dominant script to each language, and it is desirable to do so for all forms of computational linguistic work. First we list keyboard layouts that should be assigned to languages:

name	source
Chinese, Traditional (Hong Kong)	mac_input
Chinese, Traditional (Taiwan)	mac_input
Berber languages	ubuntu_language_pack
Berber	$ubuntu_input$
Ogham	ubuntu_input
Coeur d'Alene Salish	$ubuntu_input$
Cameroon Multilingual	ubuntu_input

name	source
Iraqi	ubuntu_input
Indian	$ubuntu_input$
Cameroon English	$ubuntu_input$
Pannonian Rusyn	$ubuntu_input$
Berber languages	$ubuntu_input$
Chinese (Taiwan)	win10_language_pack
Chinese (Hong Kong SAR)	$win10_language_pack$
Phags-pa	$win10_input$
Chinese (Simplified, Singapore) - US keyboard	$win10_input$
Chinese (Traditional, Hong Kong S.A.R.)	$win10_input$
Chinese (Traditional Macao S.A.R.) US Keyboard	$win10_{input}$
India	$win10_input$
Thai Kedmanee	$win10_{input}$
Futhark	$win10_{input}$
Old Italic	${ m win}10$ _input
Chinese (Traditional) - US Keyboard	$win10_{input}$
Hindi Traditional	${ m win}10$ _input
Ol Chiki	$win10_{input}$
Emilian-Romagnol	hunspell
Nahuatl	hunspell
Banyumasan	hunspell
Bihari	hunspell

Finally, from the Office13 distribution we get input customization as follows:

```
தமிழ்
繁體中文
हिंदी
वांश्ला (वांश्ला ( দশ)
简体中文
```

Next we turn to Omniglot, which lists many scripts where associating a dominant language with SIL code was hard for HAS: Bagatha, Canaanite, Cypriot, Cyrillic, Elfdalian, Emilian-Romagnol, Folkspraak, Fuzhounese, Goudu, Hadhramautic, Himyaritic, Hotcak, Ifugao, Interslavic, Jewish Neo-Aramaic, Kammara, Kotia, Latino sine Flexione, Maghrebi Arabic alphabet, Makasarese, Malachim, Mayan, Mixtec, Mwangwego, Nabataean, Nushu, Old Turkic, Odia, Orkhon, Oshi Wambo, Ranjana, Romániço, Saami/Sámi, Sankethi, Shanghainese, Sharda, Shetland(ic), Siddham, Slovio, Solresol, Tepehuán, Theban, Tocharian, West Polesian, Wenzhounese, Yolngu.

Most of these, such as Peano's Interlingua (Latino sine flexione), are clearly out of scope for SIL, yet they should be proposed for 639-3 codification. Others like Caanite map to language families, yet others like Nushu exist only as scripts. The list is not that long, we could go through it together and decide on a case-by-case basis how to proceed.

2.3 Languages from the Endangered Languages Project It is here that the need for standardization is the most clear. While HAS could assimilate data from thousands of languages treated by endangeredlanguages.com, there remain 207 for which the mapping is unclear. In about a third of these (69 cases, marked by X at the end of line) this is an embarrassment of riches, in that more than one 639-3 candidate is present – we assume these can be resolved in short order by consulting with SIL area experts. In the remaining two-thirds this looks harder, and may take years.

name	SIL_code	linglist	multiple_sils
Meymai		98s	
Zefra'i		7ri	
Kuhpayi		3oz	
Jarqu'i		1xw	
Homshetsi		1ev	
Cambodian Sign Language		4rr	
Baka (Far North Region, Cameroon)		5h1	
Amami-North Okinawan	kzg, xug, ryn, okn, ams, tkn, yox		X
Hachijo			
Hill Miri			
Gyalsumdo		nmm-gya	
Buu (Nigeria)		gji-zar	
Dyarim			
Tule			
Minhe Monguor			
Kumandin			
Tuha			
Lopnor Uighur			
Khamnigan Ewenki			
Guus		say-sig	
Judeo-Isfahani			
Judeo-Kashani			
Judeo-Yazdi			
Judeo-Hamadani			
Judeo-Shirazi	•		**
Central Baja Mixtec	mks, mxa		X
Coast Mixtec	mbz, mih, mio, mjc, mtu, mxt, mza, vmj		X
Eastern Alta Mixtec	mab, mil, mqh, mtx, mxs, mxy, vmm, vmx, xtd, xtp, xts		X X
Guerrero Mixtec	mim, mxv, xta, xty		X
Northeastern Alta Mixtec Northern Alta Mixtec	mip, vmq miz, xtu		X
Northern Baja Mixtec	mii, mit, xtb		X
Southern Baja Mixtec	jmx, miy, vmc		X
Tezoatlan Mixtec	miu, mxb		X
Western Alta Mixtec	mce, mdv, meh, mib, mie, mig,		71
Western Trick Minister	mpm, mvg, xti, xtj, xtl, xtm, xtn, xtt		X
Western Baja Mixtec	1 7 37 7 37 7 7 7		
One	aun, oin, okk, onk, onr, osu		X
Ambulas	abt, wos		X
Kwaruwi Kwundi	sdk, keh		X
Bisorio	bir, bic		X
Alfendio	afk, afp		X
Triw		kuf-tri	
Dakkang		kuf-dak	
Chatong		1iu	
Ashéninka	cpc, cjo, prq, cpu, cpy, cpb		X X
Cuna	cuk, kvn	1:3	Λ
Baïnounk Gubëeher Baïnounk Gujaher		$_{ m 0tz}^{ m 1id}$	
Lapachu			
Demushbo		qa6 $1at$	
Siona-Secoya	snn, sey	lat	X
Kaiep	kbw, trb		X
Emmi	zmr, amy		X
Awiakay	•	1j1	
Xaad Kil	hdn, hax	3	X
Mardin Sign Language	•	1kz	
Huastec	hsf, hus, hva		X
Nese		08o	
Kodiak Russian Creole		$1 \mathrm{hs}$	
Teushen		0qk	
Tipai		dih-tip	
Eel River Athabaskan		qt8	
Naati		1hr	
Sosorian		119	v
Kaiwá	kgk, pta		X

5

Chorote	name	SIL_code	linglist	multiple_sils
Papabneo 20, 200 2		crq, crt		X
Daiecariana			1hx	v
Dakha Sahaga Sa				А
Kaixanas was, tqn, ums, yak ytk-for Koreak Renets tar, thh, tcu, twr, tac X Chai cti, ctu X Chai the ctu, ctu, ct, ftpl X Thombo tbu, tub, tud, tuf X Mora tbu, tub, tud, tuf Ibw Mora tbu, tub, tud, tuf X Mora tbu, tub, tud, tuf tbu Mora tbu, tub, tud, tuf tbu Mora tbu tbu Kyah tbu tbu Kyah tbu tbu Kyah tbu tbu Law tbu tbu Chain tbu tbu		dic	1hv	
Forest Strand S				
Tarabumara	Sahaptin	waa, tqn, uma, yak		X
Chol			yrk-for	
Chaj				
Calicatec				
Tapanec				
Timebo thu, tha, tnd, tnf X Mora toc, tp, tow, tku, tqt, too, tie X Mora lix Mora lix Mora lix Sterra Miwok csm, nsq, skd Chaima cly, cuo Huave hev, hue, huv, hvv X Ing kit, kjd, kld liz Garjude local and the control in the control				
Myra Sierra Mivok Sierra Mivok		tbn, tnb, tnd, tuf		
Kiyabi		toc, tlp, tos, top, tcw, tku, tqt, too, tlc		X
Sierra Mivok				
Chaima		csm nsa skd	IIIW	Y
Haave hve, hue, hnv, hvv				
Ixi	Huave			
Sengwer				
Ganjulé		ixi, ixj, ixl		X
Lowland Mise mco, mir, mzl, pxm x x stolados do Massaco lar laq stolados do Tanaru mzx, mpw x x laq mx l				
Ikr		meo mir mal pym	KCX-gan	Y
Scolado do Tanaru		inco, inii, inzi, pxiii	1kr	Λ
Mawayana				
Harakmbut hug, amr	Mawayana		•	
Mo'ang	Harakmbut			
Popoloca Popoloca Popoloca Popoloca Popoloca Popoloca Popoloca Popoloca Rup-nyj O81 Nysanjang Nysanjan				
Nyanjang		alf also and a second	$1 \mathrm{hn}$	v
Yiji OSI Ngaatjatjara 08q Putijarra 1j9 Ngaliwuru djd-nga Fjungundji ogq Karko (India) adi-kar Pani Koch kdq-ban Perion 05k Kasong 113 Kasong beq-mer Kosong lek Swoeng leu Juk lek Jakalteko jac, jai X Yinggor keq-mer K Yunggor leu Juk Jakalteko jac, jai X Yinggor key X Yanggor key X Juk Jumaytepeque Xinka qhq Y Poqomchi' poh, pob X Sumo sum, yan, ulw X Cora crn, cok If Barrada bor-bar K Gayir oht X Gayir oht X		ры, рое, pow, poe, pps, pls, pca	knn n	Λ
Nganitajitara 08q Putijarra 1j0 Ngaliwurru 1j0 Ngaliwurr				
Putjarra 1j0				
Ngaliwuru				
Karko (India) add-kar Pani Koch add-ban Puiron 05k Kasong 113 Mer (Ethiopia) beq-mer Shé beq-she Yunggor lek Swoeng leu Juk lbo-juk Jakalteko jac, jai Stawacapán Xinka qda Jumaytepeque Xinka qda Jumaytepeque Xinka qhq Joponchi' poh, pob Sumo sum, yan, ulw X Cora crn, cok 1kf Barrada bzr-bar X Barrada bcy2 Barrada Barrada bcy2 Br			djd-nga	
Pani Koch				
Puiron				
Mayor Mayo				
Mer (Ethiopia) beq-mer Seq-mer Seq-mer She beq-she swoeng 1ek Seq-she beq-she beq		119	USK	
She		113	bca-mer	
Swoeng				
Juk	Yunggor			
Jakalteko jac, jai X Triqui trs, trc, trq X Guazacapán Xinka qda Jumaytepeque Xinka yda Poqomchi' poh, pob X Sumo sum, yan, ulw X Cora crn, cok Ikf Bardada bzr-bar Barnada bzr-bar Barna Obq Boonwurrung Kapong Kapong ake, pbc Gayiri Ohf Guwar 1j0 Karamin Ohx Keramin Ohx Kolakngat Jif Mbiywom Ij4 Mbiywom Ogx Muk-Thang Jif Ngurhat Ohw Ngumbarl Ohw Ngumbarl Ohw Ngumbarl Ohw Ngumbarl Ohw Wernba-Wemba Ovd Well Marini Ijg Marki-Mathi Ijg Taruma Oj </td <td>Swoeng</td> <td></td> <td></td> <td></td>	Swoeng			
Triqui trs, trc, trq X Guazacapán Xinka qda yda Peqomchi' pob, pob X X Sumo sum, yan, ulw X X Cora crn, cok X X Wudjari lkf baz-bar Barada baz-bar baz-bar Barna 0y2 Boonwurrung X Kapong ake, pbc X Gayiri 0hf K Gayiri 0hf K Gayar 1j2 K Keramin 0hr M Kaniyang 1j2 K Keramin 0hx K Kolakngat 1j6 N Muk-Thang 1j5 N Ngarkat 1j6 N Ngintait 0hw N Ngintait 0hw N Wernba-Wemba 0pu N Well Thang 1pu N Wernba-Wemba			lbo-juk	
Guazacapán Xinka qda Jumaytepeug Xinka qbq Poqomchi' poh, pob X Sumo sum, yan, ulw X Cora crn, cok Ikf Bardada bzr-bar Barnad Barnad oy2 Boonwurrung Kapong ake, pbc X Gayiri ohf K Guwar lj0 K Eora lj0 K Kaniyang lj2 K Keramin ohx Ohx Kolakngat lj4 Mbjiywom Muk-Thang lj5 Ngarkat Kojumbarl ogx Mr Ngurhati ohw Ngumbarl Ngumbarl ohw Ngumbarl Ngumbarl ohw Ngumbarl Ramindjeri lka Uwinymil Wernambool qs4 Wemba-Wemba ulgur lp Wulguru qgu lp Taruma njg				
Jumaytepeque Xinka				Λ
Pogonchi' poh, pob X Sumo sum, yan, ulw X Cora crn, cok X Wudjari lkf Barna 0y2 Boonwurrung Ohq Kapong ake, pbc X Gayiri Ohf Guwar Ohf Eora 1j0 Kaniyang 1j2 Keramin Ohx Kolakngat Ohr Majljarra 1j4 Mbiywom 0gx Muk-Thang 1j5 Ngarkat 0hw Ngunbarl 0s Peramangk 0hw Ngunbarl 0w Ngumbarl 0w Peramangk 0w Wemba-Wemba 0w Wulguru qgu Yilpun Sign Language 1a4 Halipur Sign Language 1a4 Hujur Sign Language 1kx Brunca Sign Language 1kx Jakarta Sign Language 1kx <td></td> <td></td> <td></td> <td></td>				
Cora crn, cok X Wudjari lkf Barada bzr-bar Barna 0y2 Kanyang ake, pbc X Gayiri 0hf Guwar 0hf Eora 1j0 Kaniyang 1j2 Keramin 0hx Kolakngat 0hr Molk-Tang 0gx Maljijura 0gx Muk-Thang 1j6 Ngarkat 1j6 Nginthait 0hw Ngumbarl 0hw Peramangk 0vd Ramindjeri 1ke Uwinymil 1ec Warrambool 0bo Wemba-Wemba 0g5 Mathi-Mathi 1jg Taruma 1kp Alipur Sign Langu				X
Mudjari				
Barda bzr-bar Barna 0y2 Kapong ake, pbc X Gayiri 0hf Guwar 0hf Eora 1j0 Kaniyang 1j2 Keramin 0hx Kolakngat 0hr Manjiljarra 1j4 Mbijwom 0gx Muk-Thang 1j5 Ngarkat 0hw Ngintait 0hw Ngumbarl 08s Peramangk 0vd Ramindjeri 1ka Uwinymil qs4 Wemba-Wemba 0ho Wullguru qgu Yilba 0g5 Mathi-Mathi 1g Taruma qoi Original Costa Rican Sign Language 1k Piupany 1kp Alipur Sign Language 1kv Purnea Sign Language 1kv Purnea Sign Language 1kv Purapecha 1kx Salvadoran Lenca <t< td=""><td></td><td>crn, cok</td><td></td><td>X</td></t<>		crn, cok		X
Banna 0y2 0hq Kapong 0hq Kapong X Gayiri 0hf Kapong 0hf Kapong X Obf Gayiri 0hf Description Description Midland Description Description Description Description Midland Description <				
Boomwirung				
Kapong ake, pbc X Gayiri 0h5 0hf Eora 1j0 K Kaniyang 1j2 K Keramin 0hx Nhx Kolakngat 0hr Mhdsangailjarra 1j4 Mbiywom 0gx Mwk-Thang 1j5 Ngarkat 1j6 Ngintait 0hw Ngumbarl Ngumbarl 0 0ms Ngumbarl 0ks Neramangk 0vd Ramindjeri 1ka Uwinymil 1ec Warrambool 0who Wuguru qs4 Wemba-Wemba 0mho Wuguru qgu Yiba 0g5 Mathi-Mathi 1jg Taruma 1jg Taruma 1jg Taruma 1jg Taruma 1jg Taruma 1kp Allpur Sign Language 1kp Allpur Sign Language 1kp Allpur Sign Language 1kx Original Bangkok Sign Language 1kx Original Bangkok Sign Language 1kx Original Eanguage 1kx Original Eanguage 1kx Original Eanguage 1kx				
Gayiri 0h5 Guwar 0hf Eora 1j0 Kaniyang 1j2 Keramin 0hx Kolakngat 0hr Manjiljarra 1j4 Mbiywom 0gx Muk-Thang 1j5 Ngarkat 0hw Ngintait 0hw Ngintait 0hw Ngumbarl 08s Peramangk 0vd Ramindjeri 1ka Uwinymil 1ec Warrnambool 0ho Wulguru qs4 Yilba 0g5 Mathi-Mathi 1jg Taruma 0j Original Costa Rican Sign Language 1ka Jipany 1kp Alipur Sign Language 1kt Bribri Sign Language 1kt Bribri Sign Language 1kv Jakarta Sign Language 1kv Salvadoran Lenca 1kv Purepecha mxq, neq X Kaqchike		ake, pbc	01	X
Eora 1j0 Kaniyang 1j2 Keramin 0hx Kolakngat 0hr Manjiljarra 1j4 Mbijwom 0gx Muk-Thang 1j5 Ngarkat 1j6 Ngintait 0hw Ngintait 0hw Ngumbarl 0vd Peramangk 0vd Ramindjeri 1ka Uwinymil 1ec Warnambool qs4 Wemba-Wemba 0ho Wulguru qgu Yilba 0g5 Mathi-Mathi 1jg Taruma 0g5 Original Costa Rican Sign Language 1a4 Tjupany 1kp Alipur Sign Language 1kt Bribri Sign Language 1kv Jakarta Sign Language 1kv Sulvadoran Lenca 1kv Purepecha tsz, pua X Widland Mixe mxq, neq X Kaqchikel cak, ckk, cke, ckc, ck			0h5	
Kaniyang Keramin	Guwar		Ohf	
Keramin Ohx Kolakngat Ohr Manjiljarra 1j4 Mbiywom Ogx Muk-Thang 1j5 Ngarkat 1j6 Ngintait Ohw Ngumbarl 08s Peramangk Ovd Ramindjeri 1ka Uwinymil 1ec Warrnambool qs4 Wemba-Wemba 0ho Wulguru qgu Yilba 0g5 Mathi-Mathi 1jg Taruma qoi Original Costa Rican Sign Language 1a4 Tjupany 1kp Alipur Sign Language 1ku Bribri Sign Language 1ku Brunca Sign Language 1ku Brunca Sign Language 1kv Yogyakarta Sign Language 1kx Original Bangkok Sign Language 1kw Yogyakarta Sign Language 1kw Purepecha txz, pua Midland Mixe mxq, neq X Kaqchi				
Kolakngat Ohr Manjiljarra 1j4 Manjiljarra 0gx Muk-Thang 1j5 Ngarkat 1j6 Ngintait 0hw Ngumbarl 08s Peramangk 0vd Ramindjeri 1ka Uwinymil 1ec Warrnambool qs4 Wemba-Wemba 0ho Wulguru qgu Yilba 0g5 Mathi-Mathi 1jg Taruma qoi Original Costa Rican Sign Language 1a4 Tjupany 1kp Alipur Sign Language 1kt Bribri Sign Language 1ku Brunca Sign Language 1kv Jakarta Sign Language 1kv Jakarta Sign Language 1kw Yogyakarta Sign Language 1kw Salvadoran Lenca 062 Purepecha tsz, pua Midland Mixe mxq, neq Cak, ckk, cke, ckc, cki, ckj, ckd, ckf, ckw, cbm X <t< td=""><td></td><td></td><td></td><td></td></t<>				
Manjiljarra 1j4 Mbiywom 0gx Muk-Thang 1j5 Ngarkat 1j6 Ngintait 0hw Ngumbarl 0x0d Peramangk 0v0d Ramindjeri 1ka Uwinymil 1ec Warrnambool qs4 Wemba-Wemba 0ho Wulguru qgu Yilba 0g5 Mathi-Mathi 1jg Taruma qoi Original Costa Rican Sign Language 1a4 Tjupany 1kp Alipur Sign Language 1kt Bribri Sign Language 1ku Brunca Sign Language 1kv Jakarta Sign Language 1kv Original Bangkok Sign Language 1kx Original Bangkok Sign Language 1kx Purepecha tsz, pua X Midland Mixe mxq, neq X Kaqchikel cak, ckk, cke, ckc, cki, ckj, ckd, ckf, ckw, cbm X Conchucos Quechua qva, qwa, qwa, qxn, qxo <td></td> <td></td> <td></td> <td></td>				
Mbiywom 0gx Muk-Thang 1j5 Ngarkat 1j6 Ngintait 0hw Ngintait 08s Peramangk 0vd Ramindjeri 1ka Uwinymil 1ec Warrnambool 0ho Wemba-Wemba 0p5 Mathi-Mathi 1jg Taruma 0g5 Original Costa Rican Sign Language 1a4 Tjupany 1kp Alipur Sign Language 1kt Bribri Sign Language 1ku Brunca Sign Language 1kv Jakarta Sign Language 1kv Yogyakarta Sign Language 1kx Original Bangkok Sign Language 1kx Purepecha tsz, pua 1kw Yogyakarta Sign Language 1ky Salvadoran Lenca 062 Purepecha tsz, pua X Midland Mixe mxq, neq X Kaqchikel cak, ckk, cke, ckc, cki, ckj, ckd, ckf, ckw, cbm X Conchucos Quechua				
Muk-Thang 1j5 Nagrakat 1j6 Ngarkat 1j6 Ngarkat 1j6 Ngarkat 1j6 Ngarkat 0hw Ngarkat 0hw Ngarkat 0hw Ngarkat 0hw Ngarkat 0hw Ngarkat 0kg Ngarkat 0vd Ramindjeri 1ka Uwinymil 1ka Uwinymil 1ka Uwinymil 1kc Oko Owo Owo Wulguru 0kl Owo	Mbiywom			
Ngintait 0hw Ngumbarl 08s Peramangk 0vd Ramindjeri 1ka Uwinymil qs4 Warrnambool qs4 Wemba-Wemba 0ho Wulguru qgu Yilba 0g5 Mathi-Mathi 1jg Taruma qoi Original Costa Rican Sign Language 1a4 Tjupany 1kt Alipur Sign Language 1kt Bribri Sign Language 1ku Brunca Sign Language 1kv Jakarta Sign Language 1kv Yogyakarta Sign Language 1kw Yogyakarta Sign Language 1kw Salvadoran Lenca 1kw Purepecha tsz, pua X Midland Mixe mxq, neq X Kaqchikel cak, ckk, cke, ckc, cki, ckj, ckd, ckf, ckw, cbm X Conchucos Quechua qwa, qwa, qxn, qxo X Alto Pativilca qva, qxh Y	Muk-Thang		1j5	
Ngumbarl 08s Peramangk 0vd Ramindjeri 1ka Uwinymil 1ec Warrnambool qs4 Wemba-Wemba 0ho Wulguru qgu Yilba 0g5 Mathi-Mathi 1jg Taruma qoi Original Costa Rican Sign Language 1a4 Tjupany 1kp Alipur Sign Language 1kt Bribri Sign Language 1ku Brunca Sign Language 1kv Jakarta Sign Language 1kv Jakarta Sign Language 1kx Original Bangkok Sign Language 1kx Salvadoran Lenca 062 Purepecha tsz, pua X Midland Mixe mxq, neq X Kaqchikel cak, ckk, cke, ckc, cki, ckj, ckd, ckf, ckw, cbm X Conchucos Quechua qwa, qwa, qwa, qxn, qxo X Alto Pativilca qux-lar				
Peramangk				
Ramindjeri				
Uwinymil 1ec Warrnambool qs4 Wemba-Wemba 0ho Wulguru qgu Yilba 0g5 Mathi-Mathi 1jg Taruma qoi Original Costa Rican Sign Language 1a4 Tjupany 1kp Alipur Sign Language 1kt Bribri Sign Language 1ku Brunca Sign Language 1kv Jakarta Sign Language 1kv Jakarta Sign Language 1kx Original Bangkok Sign Language 1kw Yogyakarta Sign Language 1ky Salvadoran Lenca 062 Purepecha tsz, pua X Midland Mixe mxq, neq X Kaqchikel cak, ckk, cke, ckc, cki, ckj, ckd, ckf, ckw, cbm X Conchucos Quechua qwa, qwa, qwa, qxn, qxo X Alto Pativilca qva, qxh X				
Warrnambool qs4 Wemba-Wemba 0ho Wulguru qgu Yilba 0g5 Mathi-Mathi 1jg Taruma qoi Original Costa Rican Sign Language 1a4 Tjupany 1kp Alipur Sign Language 1kt Bribri Sign Language 1kv Bribri Sign Language 1kv Original Bangkok Sign Language 1kx Original Bangkok Sign Language 1kx Yogyakarta Sign Language 1ky Salvadoran Lenca 1ky Purepecha tsz, pua X Midland Mixe mxq, neq X Kaqchikel cak, ckk, cke, ckc, cki, ckj, ckd, ckf, ckw, cbm X Conchucos Quechua qwa, qwa, qwa, qxo X Alto Pativilca qva, qxh qux-lar				
Wemba-Wemba qgu Wilguru 0g5 Mathi-Mathi 1jg Taruma qoi Original Costa Rican Sign Language 1a4 Tjupany 1kp Alipur Sign Language 1kt Bribri Sign Language 1kv Brunca Sign Language 1kv Jakarta Sign Language 1kx Original Bangkok Sign Language 1kx Yogyakarta Sign Language 1ky Salvadoran Lenca 1ky Purepecha tsz, pua X Midland Mixe mxq, neq X Kaqchikel cak, ckk, cke, ckc, cki, ckj, ckd, ckf, ckw, cbm X Conchucos Quechua qwa, qwa, qxn, qxo X Alto Pativilca qva, qxh qux-lar				
Wulguru qgu Yilba 0g5 Mathi-Mathi 1jg Taruma qoi Original Costa Rican Sign Language 1a4 Tjupany 1kp Alipur Sign Language 1kt Bribri Sign Language 1ku Brunca Sign Language 1kv Jakarta Sign Language 1kx Original Bangkok Sign Language 1kw Yogyakarta Sign Language 1ky Salvadoran Lenca 062 Purepecha tsz, pua X Midland Mixe mxq, neq X Kaqchikel cak, ckk, cke, ckc, cki, ckj, ckd, ckf, ckw, cbm X Conchucos Quechua qwa, qwa, qwa, qxo X Alto Pativilca qva, qxh qux-lar	Wemba-Wemba			
Mathi-Mathi 1jg Taruma qoi Original Costa Rican Sign Language 1a4 Tjupany 1kp Alipur Sign Language 1kt Bribri Sign Language 1ku Brunca Sign Language 1kv Jakarta Sign Language 1kx Original Bangkok Sign Language 1kw Yogyakarta Sign Language 1ky Salvadoran Lenca 062 Purepecha tsz, pua X Midland Mixe mxq, neq X Kaqchikel cak, ckk, cke, ckc, cki, ckj, ckd, ckf, ckw, cbm X Conchucos Quechua qwa, qwa, qwa, qxo X Alto Pativilca qva, qxh x Laraos Quechua qux-lar	Wulguru	qgu		
Taruma qoi Original Costa Rican Sign Language 1a4 Tjupany 1kp Alipur Sign Language 1kt Bribri Sign Language 1ku Brunca Sign Language 1kv Jakarta Sign Language 1kx Original Bangkok Sign Language 1kw Yogyakarta Sign Language 1ky Salvadoran Lenca 062 Purepecha tsz, pua X Midland Mixe mxq, neq X Kaqchikel cak, ckk, cke, ckc, cki, ckj, ckd, ckf, ckw, cbm X Conchucos Quechua qwa, qwa, qwa, qxn, qxo X Alto Pativilca qva, qxh x				
Original Costa Rican Sign Language 1a4 Tjupany 1kp Alipur Sign Language 1kt Bribri Sign Language 1ku Brunca Sign Language 1kv Jakarta Sign Language 1kx Original Bangkok Sign Language 1kx Yogyakarta Sign Language 1kw Salvadoran Lenca 062 Purepecha tsz, pua X Midland Mixe mxq, neq X Kaqchikel cak, ckk, cke, ckc, cki, ckj, ckd, ckf, ckw, cbm X Conchucos Quechua qwa, qwa, qwa, qxo X Alto Pativilca qva, qxh qux-lar				
Tjupany 1kp Alipur Sign Language 1kt Bribri Sign Language 1ku Brunca Sign Language 1kv Jakarta Sign Language 1kx Original Bangkok Sign Language 1kw Yogyakarta Sign Language 1ky Salvadoran Lenca 062 Purepecha tsz, pua X Midland Mixe mxq, neq X Kaqchikel cak, ckk, cke, ckc, cki, ckj, ckd, ckf, ckw, cbm X Conchucos Quechua qwa, qwa, qwa, qxo X Alto Pativilca qva, qxh X Laraos Quechua qux-lar				
Alipur Sign Language				
1				
Brunca Sign Language 1kv Jakarta Sign Language 1kx Original Bangkok Sign Language 1kw Yogyakarta Sign Language 1ky Salvadoran Lenca 062 Purepecha tsz, pua X Midland Mixe mxq, neq X Kaqchikel cak, ckk, cke, ckc, cki, ckj, ckd, ckf, ckw, cbm X Conchucos Quechua qwa, qws, qxn, qxo X Alto Pativilca qva, qxh x Laraos Quechua qux-lar				
Original Bangkok Sign Language 1kw Yogyakarta Sign Language 1ky Salvadoran Lenca 062 Purepecha tsz, pua X Midland Mixe mxq, neq X Kaqchikel cak, ckk, cke, ckc, cki, ckj, ckd, ckf, ckw, cbm X Conchucos Quechua qwa, qws, qxn, qxo X Alto Pativilca qva, qxh X Laraos Quechua qux-lar	Brunca Sign Language			
Yogyakarta Sign Language 1ky Salvadoran Lenca 062 Purepecha tsz, pua X Midland Mixe mxq, neq X Kaqchikel cak, ckk, cke, ckc, cki, ckj, ckd, ckf, ckw, cbm X Conchucos Quechua qwa, qws, qxn, qxo X Alto Pativilca qva, qxh X Laraos Quechua qux-lar				
Salvadoran Lenca 062 Purepecha tsz, pua X Midland Mixe mxq, neq X Kaqchikel cak, ckk, cke, ckc, cki, ckj, ckd, ckf, ckw, cbm X Conchucos Quechua qwa, qws, qxn, qxo X Alto Pativilca qva, qxh X Laraos Quechua qux-lar				
Purepecha tsz, pua X Midland Mixe mxq, neq X Kaqchikel cak, ckk, cke, ckc, cki, ckj, ckd, ckf, ckw, cbm X Conchucos Quechua qwa, qws, qxn, qxo X Alto Pativilca qva, qxh X Laraos Quechua qux-lar				
Midland Mixe mxq, neq X Kaqchikel cak, ckk, cke, ckc, cki, ckj, ckd, ckf, ckw, cbm X Conchucos Quechua qwa, qws, qxn, qxo X Alto Pativilca qva, qxh X Laraos Quechua qux-lar		tsz pua	062	X
Kaqchikel cak, ckk, cke, ckc, cki, ckj, ckd, ckf, ckw, cbm X Conchucos Quechua qwa, qwa, qxn, qxo X Alto Pativilca qva, qxh X Laraos Quechua qux-lar				
Conchucos Quechua qwa, qws, qxn, qxo X Alto Pativilca qva, qxh X Laraos Quechua qux-lar				
Alto Pativilca qva, qxh X Laraos Quechua qux-lar				
	Alto Pativilca			
Apuri () wachua				
	Apurí Quechua		qux-apu	
Lincha Quechua qux-lin				
Chocos Quechua qux-cho Madeán Quechua qux-mad				

name	SIL_code	linglist	multiple_sils
Gansu Bonan		1li	
Qinghai Bonan		1lh	
Khamnigan Mongol		1li	
Nyagrong Minyag		nm0	
Lower Umpqua		sis-low	
Figuig		qb8	
Tetserret		1nl	
Northern Khanty		1of	
Southern Khanty		log	
Eastern Khanty		lok	
Northern Selkup		100	
Southern Selkup		1or	
Amur Nivkh		1ot	
New Bargut		1no	
Old Bargut		1np	
Sinkiang Dagur		1nq	
Ongkor Solon		1ns	
Tzeltal	tzb, tzh		X
Tzotzil	tzc, tze, tzu, tzs, tzo, tzz		X
Gascon		oci-gsc	
Penange		1qa	
Ixtlán	zaa, zpd, zae		X
Rincón	zar, zsr		X
Ocotlán	zac, zpv, zpn		X
Coatlán Zapotec	zps, zpt, ztp, zao, zam, zpr, zap, ztg, ztl, zpo, zpb		X
Buu (Cameroon)		8uu	
Villalta	zad, zav, zpu, zpq, ztc, zat		X
Tlacolula	zab, ztt, ztj, zaw, zaq, zpf		X
Zimatlán	zph, zpp, zpl		X
Geviya		gev	
Urmia Northeastern Neo-Aramaic		08g	
Central Jewish Neo-Aramaic		0xp	
Northern Northeastern Neo-Aramaic		0zn	
Aširat Northeastern Neo-Aramaic		08a	
Southern Northeastern Neo-Aramaic		0xs	
Kaera		08y	
Alabugat Tatar		nog-tat	
Vivaro-Alpine		08e	
Gardiol		1h9	
Chiquimulilla Xinka		2df	
Lower West-Central Chinantec	cuc, cnt		X
Central Chinantec	cuc, cvn, csa, cle, cpa		X
Sierra Chinantec	chq, cco, cvn		X

- **2.4 Language Archives** Only one unidentified language, Himachali. This is spoken in a politically hypersensitive area (Kashmir), but the contents of the
- **2.5 Indigeneous tweets/blogs** Ayuujk, Emiliàn e Rumagnòl, Cántabru, Ripuarische euvergangs, Hñähñu, həńqəminnəm, Ooslimbörgs, Centraal-Limbörgs, Nahua (X)
- 2.6 WALS Allentiac Ayomán, Berber (Figuig), Betoi, Chasta Costa, Colac, Cuica, Cuitlatec, Esmeraldeño, Guaque, Jeli, Juat, Kasong, Kenyah (Uma' Lung), Kriol (Fitzroy Crossing), Kualan, Lughat al-Isharat al-Lubnaniya, Madimadi, Maipure, Máku, Mixe (Ayutla), Mongol (Khamnigan), Nahuatl (Huauchinango), Nahuatl (Milpa Alta), Nahuatl (Pochutla), Romani (Sepecides), Russian-Chinese Pidgin (Birobidjan), Tasmanian (Oyster Bay to Pitwater), Warrnambool, Western Desert (Ooldea), Xiriana, Yurimangí.

2.6 Wikipedia (incl. incubators)

We begin with the full-blown Wikipedias that belong to languages. Here we would rather not speak of dialects even though this is clearly the case with some, in that establishing a working WP is a major feat and can be thought of as carving out a significant foothold among languages:

name	wp-code
Emilian-Romagnol	eml
Zamboanga Chavacano	$\operatorname{cbk-zam}$
Aramaic	arc
Kabardian Circassian	kbd

Next we have the incubators (only the unidentified are listed here):

name	wiki_inc_code
Eranadan	aaf
Maroccan Arabic	ary
Chin	chi
Creole Spanish	crp
Hindko	hnd
Proto-Indo-European	ine
Mixtec	mxt
Mayan	myn
Otomi	ote
Old Turkic	otk
Pothowari	phr
Polisakart'	pls
Proto-Germanic	gem-pro
Dari (Afghanistan)	prs
Qivorina	qvs
Tamang	taj
Tarahumara	tar
Classical Tagalog	tgl
Old Tupi	tpn
East Franconian German	vmf
Wringinian	wra

2.7 Uriel

Finally, data has been taken from the Uriel typological database. This covers several languages with three-letter codes that are not part of 639-3: dtn, dwu, dwy, esg, fnb, gjr, ilm, ilp, itd, jka, mis, mjb, mul, ntd, olu, pgz, rsm, rzh, tdm, und, wsg, xak, yro, zxx. We don't have names for these, and it will take some special effort to identify what languages are actually meant. But for most of the Uriel languages in $H \setminus S$ we have at least a vernacular name:

name	$uriel_code$
Ammonite	qgg
Arara do Acre	adc
Dalecarlian	qer
Eka	ekb
Greek (Calabria)	gre
Günün Yajich	gny
Jamtska	$_{ m jmk}$
Limonese Creole	qlm
Old Indic	qmx
Old Kannada	qkn
Old Khmer	qok
Old Latin	qbb
Paisaci Prakrit	qpp
Parkateje	qpt
Pisamira	psx
Sabellic	qhr
Salasaca Quichua	qqs
Sanskrit (Vedic)	vsn
Scanian	scy
Situ	tzi
Southern Lalo	svl
Tapachultec	qcs
Wulguru	qgu
Xuzhang Lalo	lxu
Yangliu	lly

Summary

On the whole, the 400+ discrepancies identified here should not be too hard to resolve. More important than the details of the resolution is the creation of a sustainable method for resolving them: HAS suggests the use of GitHub both as a means of transferring the software from HAS to SIL and for the tracking of both software and data bugs. The lowest-hanging fruit in the resolution process concerns Heritage languages, which were defined in Kornai, 2013 as follows:

Since digital(ized) data persists long after the last speaker is gone, we cannot simply equate failure to [digitally] ascend with lack of online data. We will make a distinction between digital *heritage* status, where material is available for research and documentation purposes, but the language is not used by native speakers (L1) for communication in the digital world, and digitally *still* status, characterized by lack of even foreign user (L2) digital presence.

There are several languages on the lists above that obviously fall in the Heritage category, and for any project on digital vitality it is important to keep track of these. We therefore suggest

for SIL to look into the possibility of adding 639-3 codepoints for these in the next update, even if the scope and charter of The Ethnologue remains unaffected by this.

References

Hammarström, Harald (2015). "Ethnologue 16/17/18th editions: A comprehensive review". In: Language 91.3, pp. 723–737.

Kornai, András (2013). "Digital language death". In: *PloS ONE* 8.10, DOI 10.1371/journal.pone.0077056. URL: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0077056.

Scannell, Kevin P (2007). "The Crúbadán Project: Corpus building for under-resourced languages". In: Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop. Vol. 4, pp. 5–15.