SIL-HAS 1st Report

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	win10_input
Chinese (Traditional) - US Keyboard	win10_input
Hindi Traditional	win10_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
বাংলা (বাংলাদেশ)
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, Hotcąk, 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
Guerrero Mixtec	mim, mxv, xta, xty		X
Northeastern Alta Mixtec	mip, vmq		X
Northern Alta Mixtec	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, mpm, mvg, xti, xtj, xtl, xtm, xtn, xtt		X
Western Baja Mixtec			
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
Cuna	cuk, kvn		X
Baïnounk Gubëeher		1id	
Baïnounk Gujaher		0tz	
Lapachu		qa6	
Demushbo		1at	
Siona-Secoya	snn, sey		X
Kaiep	kbw, trb		X
Emmi	zmr, amy		X
Awiakay		1j1	
Xaad Kil	hdn, hax		X
Mardin Sign Language		1kz	
Huastec	hsf, hus, hva		X
Nese		08o	
Kodiak Russian Creole		1hs	
Teushen		0qk	
Tipai		dih-tip	
Eel River Athabaskan		qt8	
Naati		1hr	
Sosorian		1l9	
Kaiwá	kgk, pta		X

name	SIL_code	linglist	multiple_sils
Chorote	crq, crt		X
Navwien		1hx	
Papabuco	zte, zpw, zpz		X
Dalecarlian	dlc		
Dukha		1hv	
Kaixana		08c	
Sahaptin	waa, tqn, uma, yak		X
Forest Nenets		yrk-for	
Tarahumara	tar, thh, tcu, twr, tac		X
Chol	cti, ctu		X
Chuj	cac, cnm		X
Cuicatec	cux, cut		X
Tlapanec	tpx, tpc, tcf, tpl		X
Tunebo	tbn, tnb, tnd, tuf		X
Totonac	toc, tlp, tos, top, tcw, tku, tqt, too, tlc		X
Mpra		1ix	
Kuyabi		1hw	
Sierra Miwok	csm, nsq, skd		X
Chaima	ciy, cuo		X
Huave	hve, hue, huv, hvv		X
Inga	inb, inj		X
Ixil	ixi, ixj, ixl		X
Sengwer		1iz	
Ganjulé		kcx-gan	
Lowland Mixe	mco, mir, mzl, pxm		X
Isolados do Massaco		1kr	
Isolado do Tanaru		1kq	
Mawayana	mzx, mpw		X
Harakmbut	hug, amr		X
Samatu		1hp	
Mo'ang		1hn	
Popoloca	pbf, pbe, pow, poe, pps, pls, pca		X
Nyanjang		knp-nyj	
Yiiji		081	
Ngaatjatjara		08q	
Putijarra		1j9	
Ngaliwurru		djd-nga	
Tjungundji		0gq	
Karko (India)		adi-kar	
Pani Koch		kdq-ban	
Puiron		05k	
Kasong	1l3		
Mer (Ethiopia)		bcq-mer	
Shé		bcq-she	
Yunggor		1ek	
Swoeng		1eu	
Juk		lbo-juk	
Jakalteko	jac, jai		X
Triqui	trs, trc, trq		X
Guazacapán Xinka	qda		
Jumaytepeque Xinka	qhq		
Poqomchi'	poh, pob		X
Sumo	sum, yan, ulw		X
Cora	crn, cok		X
Wudjari		1kf	
Barada		bzr-bar	
Barna		0y2	
Boonwurrung		0hq	
Kapong	ake, pbc		X
Gayiri		0h5	
Guwar		0hf	
Eora		1j0	
Kaniyang		1j2	
Keramin		0hx	
Kolakngat		0hr	
Manjiljarra		1j4	
Mbiywom		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		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	
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		1lj	
Nyagrong Minyag		nm0	
Lower Umpqua		sis-low	
Figuig		qb8	
Tetserret		1nl	
Northern Khanty		1of	
Southern Khanty		1og	
Eastern Khanty		1ok	
Northern Selkup		1oo	
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, həńɋəminńəḿ, 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	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	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 resoltion 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.