
Andras KORNAI

*Senior Scientific Advisor, Computer and Automation Research Institute,
Hungarian Academy of Sciences
(Budapest, Hungary)*

A New Method of Language Vitality Assessment

1. Background

In Kornai [2013] we demonstrated that over 95% of the world's languages are digitally still. This means there is a small pool of roughly 400 languages, many spoken in Russia and the FSU [Comrie 1981], from which a final set of digital survivors, perhaps some 200 languages, will emerge. Since at this point the digital ascent of no more than a few dozen languages is assured, we need a more detailed assessment than the simple four-way classification put forth in Kornai [2013] which distinguished only Thriving and Vital languages (neither Heritage nor Still languages can survive in the obvious sense of being actively used in communication). Figure 1 at the end of the paper, based on the data given in Table 1, shows this distribution for languages of the FSU, with the Thriving language (star) in the top right being Russian, and the Vital languages (circles) largely corresponding to the main languages of former republics. Squares are Heritage languages such as Old Church Slavonic, and smaller arrows corresponding to the remaining languages are either for Borderline (rightward pointing arrow) meaning that the current statistical method is incapable of fully resolving their status or for Still (down arrow), the majority of languages in the FSU.

2. Discussion

As the digital future of Thriving languages is assured, we use the lessons learned from the digital development of these to devise both a more detailed assessment of the digital potential of Vital languages and a strategy of maximizing the number of languages that make it across the digital divide. For the assessment we propose a simple log-linear formula that derives a single number D (digital vitality index) as a weighted sum of well-understood components such as the EGIDS ranking, (log) number of L1 speakers, (log) size of wikipedia, adjusted for quality, (log) crawl size, the existence of FLOSS spellcheckers, etc.

Some of the key factors, such as the number of speakers, represent long-range trends that are outside the immediate control of speakers. Others, such as EGIDS ratings, are set by expert judgment and no doubt carry some slight subjective element. From our perspective these are still objective, in that SIL experts also focus on long-range trends, such as literacy or official use,

that can be influenced only indirectly by the computational linguists primary responsible for digital vitality. These factors, which tend to be common for vitality in the traditional and the digital sense, are in sharp contrast to another group of factors we will call *volitional*. Whether there is a wikipedia, a blog, or a twitter community in a language depends on two factors: the availability of tools (entirely in the hands of software engineers) and the willingness/motivation of native speakers to add content. In fact, when there is a will, there is a way: a good number of native projects have already started even in the absence of language-specific tools [Scannell 2013].

For digital vitalization (as opposed to digital heritage preservation, which we see as a fallback position) we must work together with speakers who are both motivated and literate. The body of text they produce constitutes the base (Stage 0) of the following language technology pyramid: 1. Locale or i18n support for the input and output of native characters; 2. Word-level tools (spellchecker, stemmer, dictionaries); 3. Phrase- and sentence-level tools; and 4. Speech and character recognition, machine translation. Besides Stage 1 capabilities, Stage 2 requires in-depth morphological analysis and generation (which will be trivial only for isolating languages). Stage 3 (POS taggers, named entity recognizers, chunkers) presuppose Stage 2 tools, and Stage 4, the peak of the language technology pyramid, presupposes all lower levels. Measuring the maturity of tools at the various stages, and creating them as needed, is the central task of digital language vitalization.

3. Conclusions

The Information for All Programme, and UNESCO in general, can foster the vitalization process by addressing the main issues directly. On the legal front, corpora, the lifeblood of modern computational linguistics, must be unencumbered by copyright, and IFAP/UNESCO can make sure that a research exemption is enshrined in the legal framework. At the national level, projects need to make their corpora not just searchable but also downloadable by ROAMing (randomize, omit, anonymize, mix). Both for international grants and those coming from national science foundations, linguistics should follow the lead of biosciences and demand, as a precondition of funding, open access to the materials collected. Finally, a wikipedia is a necessary but insufficient condition for digital ascent (“no wikipedia, no survival”), and digital communities (not just read-only material) are also needed. Therefore we suggest to give micro-grants to small communities (literary, theatrical, etc.) to document in their native language what they are doing.

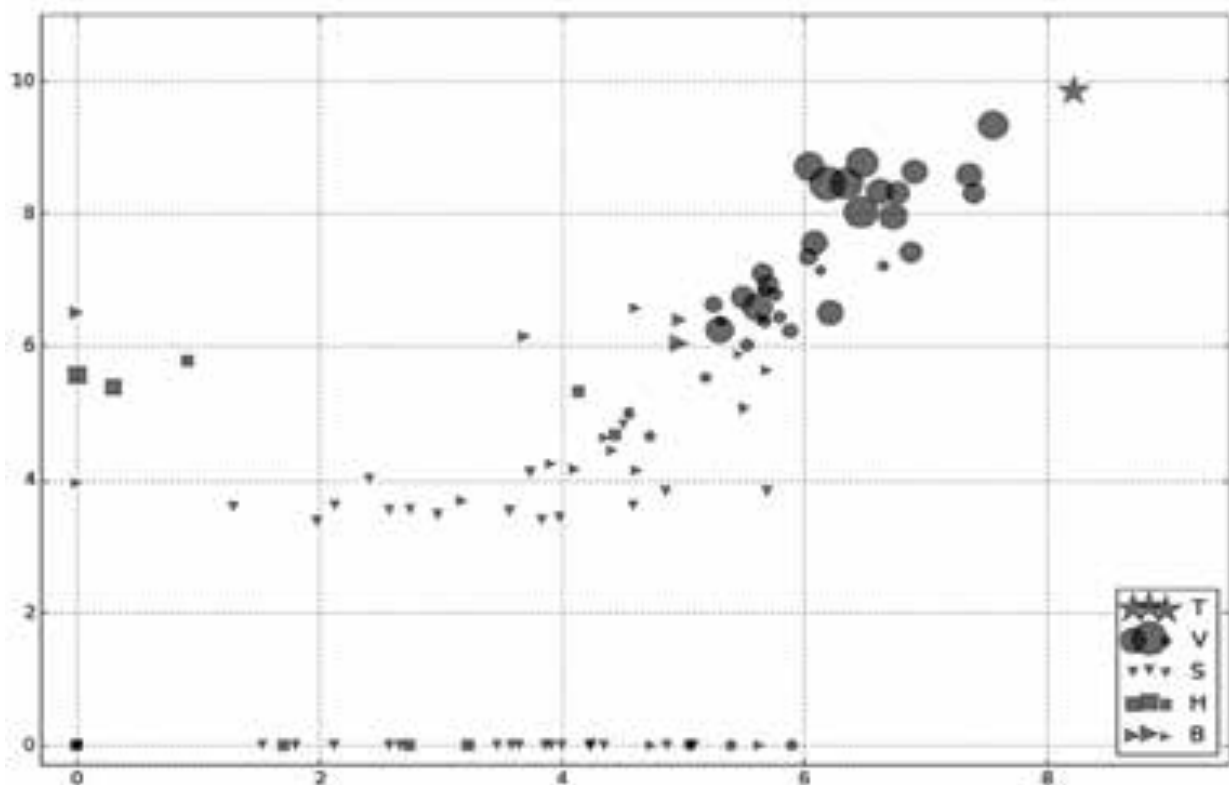


Figure 1. *Rough vitality assessment of languages in the Former Soviet Union*
(x: log population; y: log Wikipedia size; circle diameter: WP quality)

Table 1. Main vitality figures for languages in the Former Soviet Union

Language	SIL code	Vitality status	Population	Norm. WP size	WP quality
Abaza	abq	s	38,732	4,169	n/a
Abkhaz	abk	v	112,741	n/a	n/a
Adyghe	ady	b	491,801	431,288	n/a
Aghul	agx	b	22,677	43,811	n/a
Altay	alt	v	35,745	100,571	n/a
Alutor	alr	s	257	10,442	n/a
Armenian	hye	v	5,902,971	207,272,470	0.26
Avar	ava	v	761,961	1,696,067	0.11
Azerbaijani	aze	v	23,000,000	380,596,055	0.29
Bashkir	bak	v	1,221,341	36,074,594	0.29

Belarusian	bel	v	2,220,001	282,784,660	0.47
Buriat	bua	b	n/a	n/a	n/a
Chechen	che	v	1,361,001	13,900,569	0.01
Chukot	ckt	b	8,184	17,668	n/a
Chulym	clw	s	131	n/a	n/a
Chuvash	chv	v	1,077,421	22,366,496	0.15
Crimean Tatar	crh	v	475,541	2,295,102	0.06
Dargwa	dar	s	492,491	6,797	n/a
Dolgan	dlg	s	3,691	3,404	n/a
Dungan	dng	b	41,624	14,092	n/a
Enets	enf	s	33	n/a	n/a
Erzya	myv	v	336,315	1,041,262	0.08
Estonian	est	v	1,100,000	510,622,720	0.41
Even	eve	s	7,295	n/a	n/a
Gagauz	gag	v	178,024	4,238,180	0.14
Georgian	kat	v	4,237,711	215,886,780	0.33
Gilyak	niv	s	559	3,657	n/a
Gothic	got	h	n/a	369,638	0.14
Ingrian	izh	s	374	n/a	n/a
Ingush	inh	b	322,901	118,823	n/a
Itelmen	itl	s	133	4,191	n/a
Juhuri	jdt	s	17,156	n/a	n/a
Kabardian	kbd	v	1,628,501	3,187,050	0.31
Kalmyk	xal	b	291,794	750,654	0.03
Karachay-Balkar	krc	v	310,731	5,502,178	0.25
Karaim	kdr	s	94	2,431	n/a
Karakalpak	kaa	v	410,411	3,841,866	0.39
Karelian	krl	v	53,141	46,409	n/a

Kazakh	kaz	v	8,077,771	431,343,123	0.29
Ket	ket	s	376	3,546	n/a
Khakas	kjh	s	31,903	68,969	n/a
Khanty	kca	s	9,581	2,739	n/a
Khinalugh	kjj	h	1,668	n/a	n/a
Kildin Sami	sjd	h	551	n/a	n/a
Komi	kom	b	n/a	3,216,590	0.09
Komi-Permyak	koi	b	93,543	2,472,803	0.09
Koryak	kpy	s	2,916	n/a	n/a
Krymchak	jct	h	13,627	206,329	n/a
Kumyk	kum	b	426,551	n/a	n/a
Kyrgyz	kir	v	2,941,931	105,618,112	0.51
Lak	lbe	v	153,171	334,573	0.05
Latgalian	ltg	v	200,001	1,740,184	0.35
Lezgi	lez	v	788,721	n/a	n/a
Lithuanian	lit	v	3,001,861	581,134,721	0.45
Livonian	liv	h	7	607,201	n/a
Mansi	mns	s	941	3,026	n/a
Mari	mhr	v	475,874	7,053,484	0.09
Mari	mrj	b	40,531	3,759,145	0.06
Mari (Russia)	chm	s	n/a	n/a	n/a
Mingrelian	xmf	v	500,001	8,274,826	0.21
Moksha	mdf	b	92,765	1,097,308	0.16
Nanai	gld	s	3,843	n/a	n/a
Nenets	yrk	h	27,393	48,920	n/a
Nganasan	nio	s	461	n/a	n/a
Nogai	nog	s	73,305	n/a	n/a
Northern Altai	atv	b	12,728	14,652	n/a

Old Church Slavonic	chu	h	1	242,749	0.12
Old Georgian	oge	b	n/a	9,011	n/a
Ossetian	oss	v	577,451	5,982,030	0.09
Russian	rus	t	167,332,231	7,019,024,883	0.56
Rusyn	rue	v	623,501	2,736,759	0.08
Rutul	rut	b	25,923	28,060	n/a
Samogitian	sgs	h	n/a	n/a	n/a
Selkup	sel	b	1,501	4,918	n/a
Shor	cjs	s	6,811	2,510	n/a
Shughni	sgh	s	71,588	6,740	n/a
Southern Yukaghir	yux	s	62	n/a	n/a
Standard Latvian	lvs	v	1,552,261	282,525,870	0.58
Svan	sva	s	17,171	n/a	n/a
Tabassaran	tab	s	113,529	n/a	n/a
Tajik	tgk	v	4,479,651	16,106,757	0.06
Talysh	tly	v	206,196	2,359,896	n/a
Tat	ttt	s	17,320	n/a	n/a
Tatar	tat	v	5,406,111	90,404,247	0.36
Ter Sami	sjt	s	18	3,987	n/a
Tindi	tin	s	4,440	n/a	n/a
Tsakhur	tkr	s	22,188	n/a	n/a
Tsez	ddo	s	9,986	n/a	n/a
Turkmen	tuk	v	7,560,561	26,082,541	0.23
Tuvinian	tyv	v	248,429	n/a	n/a
Udi	udi	s	5,464	13,144	n/a
Udmurt	udm	b	467,156	2,556,163	0.10
Ukrainian	ukr	v	36,048,891	2,168,400,162	0.40

Ukrainian Sign Language	ukl	s	n/a	n/a	n/a
Urum	uum	s	122,654	n/a	n/a
Uzbek	uzb	v	25,000,000	203,427,158	0.21
Veps	vep	b	4,917	1,398,616	0.08
Võro	vro	b	54,773	n/a	n/a
Votic	vot	h	49	n/a	n/a
Yagnobi	yai	s	8,124	n/a	n/a
Yakut	sah	v	450,001	1,2642,821	0.20

References

1. Comrie, B. (1981). *The Languages of the Soviet Union*. Cambridge University Press.
2. Kornai, A. (2013). Digital language death. *PloS ONE*, 8(10): DOI 10.1371/journal.pone.0077056.
3. Scannell, K. P. (2013). Indigeneous tweets, indigeneous blogs (website).