Are languages alive? Rethinking the biological metaphor in linguistics

András Kornai

#IstandwithCEU

Hungarian Academy of Sciences

16th Ann'l Applied Ling Conf

Acknowledgements

- Judit Ács and Katalin Pajkossy (Budapest University of Technology and Economics)
- Johanna Domokos (Universität Bielefeld)
- Anna Fenyvesi (University of Szeged)
- Marianne Bakró-Nagy (HAS)
- Pushpak Bhattacharyya (IIT Bombay)
- Gary Simons (SIL)

Ács,Pajkossy

The biological metaphor

begins with Herder (1772) *Treatise on the Origin of Language*. Fully articulated in Darwin (1871) *The descent of man*:

The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously parallel. () We find in distinct languages striking homologies due to community of descent, and analogies due to a similar process of formation. The manner in which certain letters or sounds change when others change is very like correlated growth. () Languages, like organic beings, can be classed in groups under groups; and they can be classed either naturally according to descent, or artificially by other characters. Dominant languages and dialects spread widely, and lead to the gradual extinction of other tongues.

Death

- Language death is broadly studied, reasonably well understood
- Before death, a period of simplification
- After death, sometimes a curious afterlife (Latin, Sanskrit, . . .)
- Here the “cells” of the “organism” are the speakers
- Elsewhere the cells are the words
- But what are the genes?

Traditional symptoms of language death

EGIDS: 0. International; 1. National; 2 Provincial; 3 Wider communication; 4 Educational; 5 Developing; 6a Vigorous; 6b Threatened; 7 Shifting; 8a Moribund; 8b Nearly Extinct; 9 Dormant; 10 Extinct.

- Loss of function (trade, education, ...)
- Loss of prestige ("only the old folks talk like that")
- Loss of competence (younger generation doesn't know the words, grammar)
- Shrinking and aging of language community
- Identity function (I am Greek, my parents were Greek, some of my ancestors were Greek, ...)

Digital laguage death

DLD: T thriving; V vital;
H heritage; S still

- Digital function (reading-writing, commerce, system level support)
- Digital prestige if it's not on the web it doesn't exist
- Digital competence or computer-illiteracy
- Digital language community ‘digital natives’
- Digital identity ‘my language, my culture’ (see e.g. Moldavian wikipedia debate)

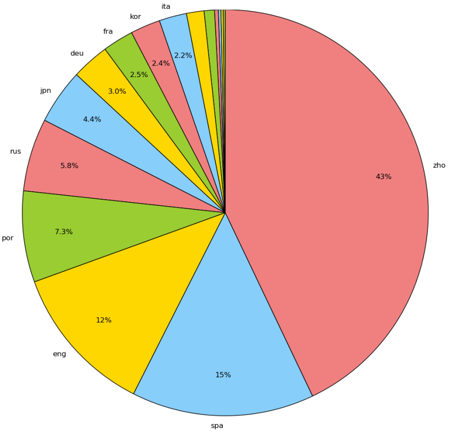
Life

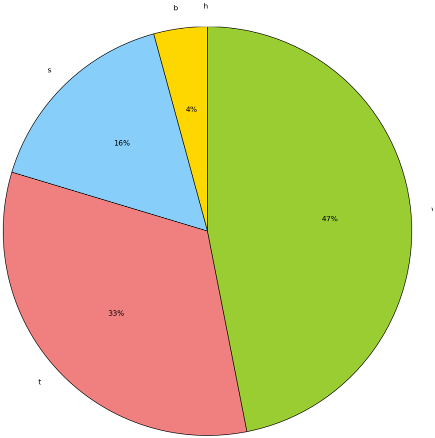
- There are no structurally (as opposed to chronologically) young languages
- There are no structurally (as opposed to chronologically) old languages
- There is no ‘aging’ (progress from ‘young/fertile’ to ‘old/senescent’)
- Do languages move? Do they metabolize? Do they react to their environment? Do they replicate?
- The metaphor has little traction (except perhaps in pre-death stage)

Birth

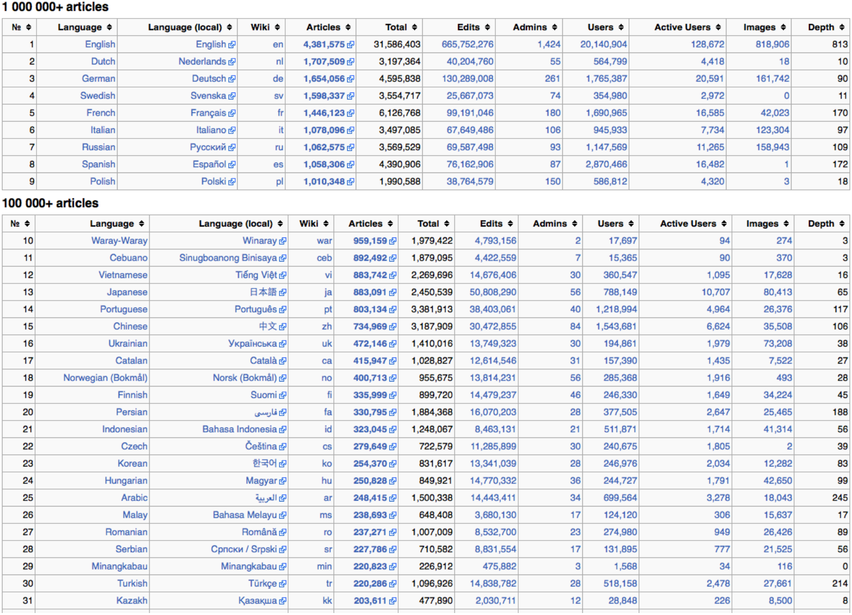
- New languages are born (e.g. Nicaraguan Sign Language)
- But not from parents! "Out of thin air" (Steven Pinker)
- Re-birth or re-constitution more frequent
- Hungarian (1770-1872), Hebrew (1882-1948), Light Warlpiri, etc.
- Perhaps *molting* would be a better metaphor than rebirth
- A fuller life cycle including not just birth-life-death but also *metamorphosis*

The distribution of languages





Wikipedias



1 000 000+ articles

№	Language	Language (local)	Wiki	Articles	Total	Edits	Admins	Users	Active Users	Images	Depth
1	English	English	en	**4,381,575**	31,586,403	665,752,276	1,424	20,140,904	128,672	818,906	813
2	Dutch	Nederlands	nl	**1,707,509**	3,197,364	40,204,760	55	564,799	4,418	18	10
3	German	Deutsch	de	**1,654,056**	4,595,838	130,289,008	261	1,765,387	20,591	161,742	90
4	Swedish	Svenska	sv	**1,598,337**	3,554,717	25,667,073	74	354,980	2,972	0	11
5	French	Français	fr	**1,446,123**	6,126,768	99,191,046	180	1,690,965	16,585	42,023	170
6	Italian	Italiano	it	**1,078,096**	3,497,085	67,649,486	106	945,933	7,734	123,304	97
7	Russian	Русский	ru	**1,062,575**	3,569,529	69,587,498	93	1,147,569	11,265	158,943	109
8	Spanish	Español	es	**1,058,306**	4,390,906	76,162,906	87	2,870,466	16,482	1	172
9	Polish	Polski	pl	**1,010,348**	1,990,588	38,764,579	150	586,812	4,320	3	18

100 000+ articles

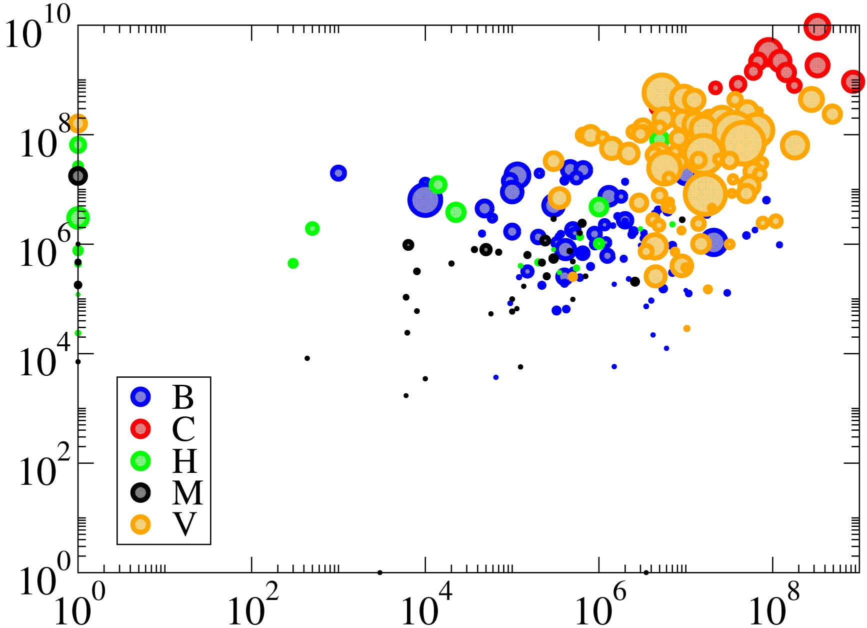
№	Language	Language (local)	Wiki	Articles	Total	Edits	Admins	Users	Active Users	Images	Depth
10	Waray-Waray	Winaray	war	**959,159**	1,979,422	4,793,156	2	17,697	94	274	3
11	Cebuano	Sinugboanong Binisaya	ceb	**892,492**	1,879,095	4,422,559	7	15,365	90	370	3
12	Vietnamese	Tiếng Việt	vi	**883,742**	2,269,696	14,676,406	30	360,547	1,095	17,628	16
13	Japanese	日本語	ja	**883,091**	2,450,539	50,808,290	56	788,149	10,707	80,413	65
14	Portuguese	Português	pt	**803,134**	3,381,913	38,403,061	40	1,218,994	4,964	26,376	117
15	Chinese	中文	zh	**734,969**	3,187,909	30,472,855	84	1,543,681	6,624	35,508	106
16	Ukrainian	Українська	uk	**472,146**	1,410,016	13,749,323	30	194,861	1,979	73,208	38
17	Catalan	Català	ca	**415,947**	1,028,827	12,614,546	31	157,390	1,435	7,522	27
18	Norwegian (Bokmål)	Norsk (Bokmål)	no	**400,713**	955,675	13,814,231	56	285,368	1,916	493	28
19	Finnish	Suomi	fi	**335,999**	899,720	14,479,237	46	246,330	1,649	34,224	45
20	Persian	فارسی	fa	**330,795**	1,884,368	16,070,203	28	377,505	2,647	25,465	188
21	Indonesian	Bahasa Indonesia	id	**323,045**	1,248,067	8,463,131	21	511,871	1,714	41,314	56
22	Czech	Čeština	cs	**279,649**	722,579	11,285,899	30	240,675	1,805	2	39
23	Korean	한국어	ko	**254,370**	831,617	13,341,039	28	246,976	2,034	12,282	83
24	Hungarian	Magyar	hu	**250,828**	849,921	14,770,332	36	244,727	1,791	42,650	99
25	Arabic	العربية	ar	**248,415**	1,500,338	14,443,411	34	699,564	3,278	18,043	245
26	Malay	Bahasa Melayu	ms	**238,693**	648,408	3,680,130	17	124,120	306	15,637	17
27	Romanian	Română	ro	**237,271**	1,007,009	8,532,700	23	274,980	949	26,426	89
28	Serbian	Српски / Srpski	sr	**227,786**	710,582	8,831,554	17	131,895	777	21,525	56
29	Minangkabau	Minangkabau	min	**220,823**	226,912	475,882	3	1,568	34	116	0
30	Turkish	Türkçe	tr	**220,286**	1,096,926	14,838,782	28	518,158	2,478	27,661	214
31	Kazakh	Қазақша	kk	**203,611**	477,890	2,030,711	12	28,848	226	8,500	8

Pilot study (2012)

Goal: to scope out the phenomena, discover methodological issues

- Measuring traditional language vitality (SIL)
- Measuring digital language vitality (WP)
- Assessing digital vitality based on expert opinion
- Heuristics, not proof!
- What if something is left out?

Results



Advanced technology and digital vitality

1. Intelligent text understanding, question answering – English only
2. Machine Translation – T-T and T-V pairs only
3. ASR – V only
4. OCR – V, H
5. Functional sentence parsing – V
6. Probabilistic lg models – V
7. Phrase-level analysis (chunking) – V
8. Word-level analysis (morphology) – V,H,S

Corpora only for the ‘densest’ lgs!

Language	Largest corpus	tokens (M)	Reference
Catalan	CUCWeb	166	Boleda et al. 2006
Croatian	Croatian Nat. Corpus	100	Tadic 2002
Czech	Czech National Corpus	1300	Kucera 2002
Danish	KorpusDK	56	n/a
Dutch	Dutch Parallel Corpus	10	Paulussen et al. 2006
Finnish	Finnish Text Collection	180	various
Indonesian	SEALang Library	5	n/a
Lithuanian	Corpus of Lithuanian	180	Marcinkevičienè 2004
Norwegian	noWaC	700	Guevara 2010
Polish	Polish National Corpus	1200	Przepiórkowski 2008
Portuguese	Corpus do Português	45	Davies & Ferreira 200
Romanian	Romanian Corpus	50	n/a
Serbian	CSL	11	Kostić 2001
Slovak	Slovak National Corpus	719	Horák et al. 2004
Spanish	Corpus del Espanol	100	Davies 2001
Swedish	Korp	910	various

CORPUS	DOWNLOAD	SEARCH
CUCWeb	NO	YES
Czech National Corpus	NO	YES
Croatian National Corpus	NO	YES
KorpusDK	NO	YES
Dutch Parallel Corpus	NO	NO
Finnish Text Collection	SOME	YES
SEALang Library	NO	YES
Corpus of Lithuanian	NO	YES
noWaC	NO	YES
Polish National Corpus	NO	YES
Corpus do Português	NO	YES
Romanian Corpus	NO	NO
CSL	NO	NO
Slovak National Corpus	NO	YES
Corpus del Español	NO	YES
Korp	SOME	YES

The process

Stage	%	Av (GB)	Stdev (GB)
Crawl		97.4	46.4
HTML, boilerplate	100.0	14.2	5.1
Sentence filtering	67.9	9.7	4.0
Language detection	44.8	6.4	3.2
Duplicate filtering	43.5	6.2	3.0
Near-duplicate filt	37.4	5.3	2.4
Morphological analysis		5MB	

Method

- Collect indicator data (existence/size of corpora is just one indicator)
- Select unquestionable ‘gold’ instances manually (training set)
- Build maximum entropy classifiers (machine learning)
- Strong automated feature selection (leaving 6-8 features out of 35)
- Internal (cross)validation
- Perturbation of train set

On the whole, how many languages can be analyzed?

1. Dictionary, grammar: < 6000
2. Standardized orthography: < 1500
3. Keyboard/input method: < 400
4. Word-level analysis: < 150

Language databases

- Summer Institute of Linguistics 7,776
- Open Language Archives Community 7,478
- Catalogue of Endangered Languages 3,175
- An Crúbadán 1,322
- Omniglot 696

Three studies (2013,2014,to appear)

Goal: furnish proof, open methodology, open data sets

- How good are these four classes (T/V/H/S)? very good
- How much data cleaning is needed? practically none
- How much can expert opinion be eliminated from the method? almost entirely
- How many languages are covered? practically all
- How reproducible? entirely
- 2014 Indian survey (with Pushpak Bhattacharya) 634 lgs, 36 V, 21 B, 1 H, 576 S
- Ongoing Uralic survey (with Judit Ács and Katalin Pajkossy)

Much data from many sources

- Traditional linguistic community (L1, L2) – SIL
- Digital linguistic community (WP, OLAC, CEL, crawls)
- Software environment (Microsoft, Apple, open source spellcheckers)
- Expertise (EGIDS, ELP)
- Over 30 parameters, trivially encoded (class numbering, log)
- For current data (February 2017) see http://hlt.bme.hu/en/dld/search

How do you know that the classifiers are any good?

- Internal consistency: tests well on train data
- Robustness: does not depend on seeds
- Correlates well with other classifiers
- Trained weights make sense
- External consistency: results agree well with expert judgement

Classification accuracy (10-fold crossvalidation)

	Seed 0				Seed 1			
#f	SH-VT	S-H-VT	SH-V-T	S-H-V-T	SH-VT	S-H-VT	SH-V-T	S-H
33	95.0	99.3	92.3	90.7	99.3	98.6	94.3	8
18	97.2	99.3	91.4	96.4	99.3	98.6	95.0	8
10	97.9	99.3	92.9	95.7	100.0	99.3	93.6	9
8	97.1	99.3	92.9	97.1	100.0	96.4	94.3	8
6	97.1	99.3	92.1	93.6	100.0	96.4	95.7	8

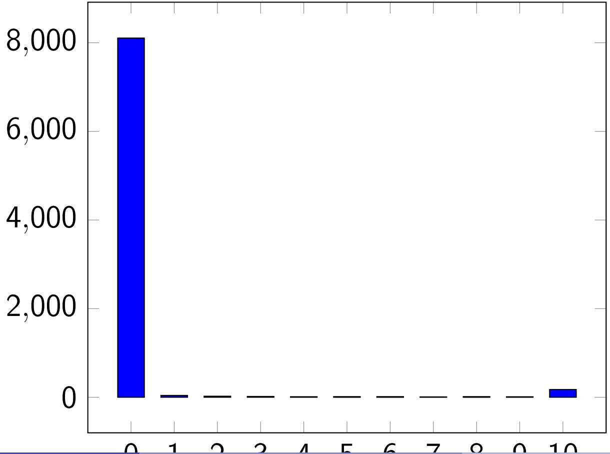
“Brain surgery” (LeCun 1990, Pajkossy 2013): we look at the weights of logistic models, leave out those small in absolute value, retrain

Added twist: feature selection

- So far we made sure we don't depend on the seeds
- Let's also eliminate data selection bias
- We collect over 30 measures of vitality such as population, EGIDS ranking, size of Wikipedia, number of docs in OLAC, etc etc.
- Leave it to the system to decide which of these actually matter
- Result: 6 or 8 feature are all it takes to build reliable classifiers

Borderline cases

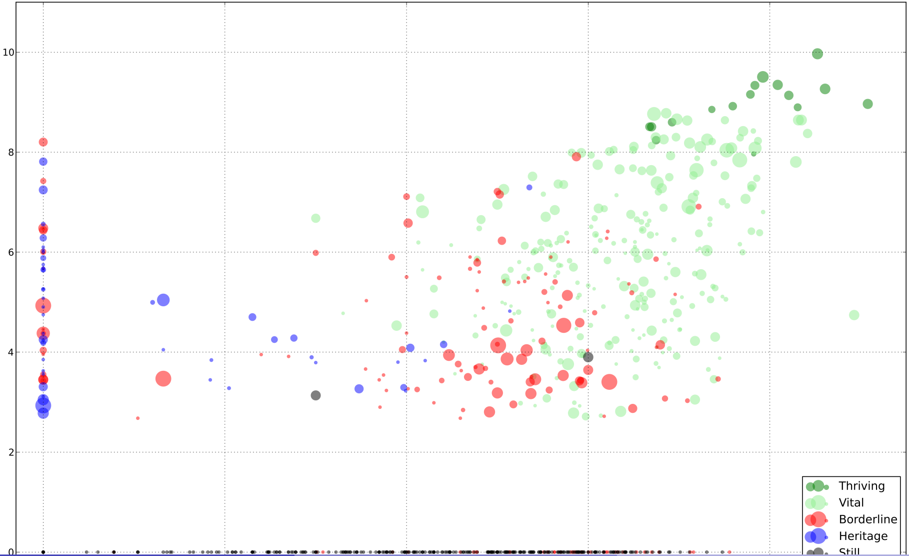
- Not a category in the analysis!
- Statistical methods are hard to apply to individuals
- But we can obtain robust statistical conclusions by “bagging” (Breiman 1996)



Perturbation of training data

- Brain surgery does not always pick the same features, e.g. L2 or WP incubator status
- The average 8-dim classifier has 0.958 ± 0.021 accuracy (as measured by crossvalidation)
- At 80% train set replacement the correlation across the classifiers is 0.889±0.04,
- At 100%, 0.823±0.088
- Based on the 80% independent ones the number of dead languages is 8,049±36, based on the 100% independent ones 8,008±69.

The results based on machine classification



Conclusions from 2013

- Tribal languages are caught in a pincer movement – the old folks can't be bothered to learn the computer, the young folks no longer care about the old traditions
- These 8,000 languages are not going to be digitally still, they already are
- Because of our conservative methodology we speak of 420 potential survivors, in reality we can expect 200 or less
- This is the final act of the Neolithic Revolution, with the urban agriculturalists moving on to a different, digital plane of existence leaving the hunter-gatherers and nomad pastoralists behind

Finding the languages/dialects

>Enets	—	Tundra Enets	enh	Forest Enets	enf
>Estonian	est	Estonian Standard	ekk	Estonian Voro	vro
Estonian Seto	—	Finnish	fin	Hungarian	hun
Ingrian/Izhorian	izh	>Karelian	krl	Khanty	kca
Khanty Northern	1of	Khanty Southern	1og	Khanty Eastern	1ok
>Komi	kom	Komi Zyrian	kpv	Komi Permyak	koi
Komi Yazva/Yodzyak	kpv-yaz	Finnish Kven	fkv	Finnish Meänkieli	fit
Karelian Livvi	olo	Karelian Ludic	lud	Mansi	mns
Mansi Northern	1nt	Mansi Eastern	1nu	Mansi Western	1od
>Mari	chm	Hill Mari	mrj	Meadow Mari	mhr
>Mordvin	—	Mordvin Erzya	myv	Mordvin Moksha	mdf
>Nenets	yrk	Tundra Nenets	yrk-tun	Forest Nenents	yrk-for
Nganasan	nio	>Selkup	sel	Selkup Northern	1oo
Selkup Central	1op	Selkup Southern	1or	Sami Inari	smn
Sami Kildin	sjd	Sami Lule	smj	Sami Northern	sme
Sami Pite	sje	Sami Skolt	sms	Sami Southern	sma
Sami Ter	sjt	Sami Ume	sju	Udmurt	udm
Veps	vep	Votic	vot	D Yurats	rts
D Kamassian	xas	D Mator	mtm	D Meshcherian	—
D Muromian	—	D Sami Akkala	sia	D Sami Kainu	—
D Sami Keni	sjk	D Livonian	liv	Uralic	urj

hun	100	100	100	100	100
fin	100	100	100	100	100
sme	99	100	99	99	99.25
mdf	95	96	96	94	95.25
est	98	96	91	93	94.5
udm	63	75	70	78	71.5
ekk	56	63	59	59	59.25
mhr	48	64	58	64	58.5
mrj	39	55	47	56	49.25
myv	37	53	49	56	48.75
koi	35	55	41	57	47
krl	32	44	39	49	41
vro	24	41	28	41	33.5
kom	26	35	27	36	31
fit	14	22	19	21	19
smn	7	8	12	10	9.25
fkv	6	9	11	10	9
kpv	5	7	4	7	5.75
vep	2	5	4	7	4.5
yrk	1	3	4	7	3.75
sjd	0	2	3	3	2
smj	1	2	2	2	1.75
sel	0	2	1	3	1.5
sma	0	2	1	2	1.25
chm	1	0	2	1	1
vot	0	1	0	2	0.75
sms	0	1	1	1	0.75
mns	0	0	1	2	0.75
liv	0	1	0	2	0.75
kca	0	0	1	2	0.75
sjt	0	0	1	0	0.25
olo	0	0	1	0	0.25
nio	0	0	1	0	0.25
lud	0	0	1	0	0.25
izh	0	0	1	0	0.25
zkb	0	0	0	0	0
xas	0	0	0	0	0
siu	0	0	0	0	0

Digitally vital

lg	sil	pop	EGIDS	wp
Finnish	fin	5.4 m	1	386k (22)
Hungarian	hun	12.6 m	1	378k (23)
Estonian	est+ekk	1.2 m	1	141k (43)
Northern Sami	sme	20k	2	7.2k (139)
Moksha	mdf	300k	5	1.1k (237)

Sámi has support from Tromsø. Moksha has good community (WP), and good foundations (dictionaries) but should really be placed at the top of the borderline range, requiring action, rather than at the bottom of the vital range.

Digitally still

Dead/dormant: Kamas (incl Koibal); Khanty Southern; Livonian; Mator; Mescherian; Muromian; Yurats
Critically endangered: Enets Forest ($\sim$ 10/2011); Enets Tundra ($\sim$30/2007); Sami Akkala (1/2013); Sami Pite ($\sim$42/2012); Sami Ume (20/2007); Selkup Central (2/2015); Selkup Southern (1/2015); Votic ($\sim$12/2015); Yazva ($\sim$200/2007)
Severly endangered: Enets Forest ($\sim$10/2011); Ingrian ($\sim$130/2013); Finnish Kven (2-8k/2005); Khanti Eastern ($\sim$480/2010); Mansi Eastern ($<$500/2000); Sami Kildin ($\sim$300/2007); Nganasan (500/2000); Sami Ter (30/2007); Veps (1600/2010)
Endangered: Sami Inari ($\sim$300/2007); Selkup Northern ($<$600/2005); Sami South (600/2015)

Digitally still $\neq$ stop working!

Still → **Heritage** is possible.

Excellent grammatical sketches can be provided (e.g. Peter Simoncsics' description of Kamas in Abondolo 1998).

There is much to be done in clearing up/digitizing old fieldwork collections (Campbell and Hauk 2015 provides a survey).

Please please please collect audio. It doesn't matter you have no time to transcribe it. Give people cellphones.

First, the dialects

Warning!

Speaker knows nothing about dialectology and has no data

- Some nonstandard Finnish dialects: Kven, Meänkieli
- One nonstandard Estonian dialect: Võro
- Exactly one dialact in Sápmi, Northern Sami
- Komi Permyak/Zyrian data confused

Borderline languages/dialects

lg	sil	ELcat	pop	E	wp	wpcorr
Erzya	myv	threat	250k (2007)	5	2.8k (192)	420
Moksha	mdf	threat	200k (2007)	5	1.1k (237)	323
Karelian	krl	threat	63k (2007)	5	incubator	n/a
Permyak	koi	vuln	110k (2007)	5	3.4k (181)	728
Zyrian	kpv	vuln	110k (2007)	5	4.5k (166)	?
Meadow Mari	mhr	vuln	500k (2007)	4	8.7k (135)	1167
Hill Mari	mrj	vuln	<50k (2007)	5	10k (132)	1380
Udmurt	udm	threat	324k (2010)	5	3.7k (175)	522

Given 4 digitally vital and these 8 potentially capable of digital ascent, the Uralic situation is far better than the global situation

The computational effort

- HAS/Morphologic Udmurt Khanty, Komi Mansi Mari Nganasan "due to the nature of Russian minority policy, the school system, the great degree of dispersion, the low esteem of the ethnic language and culture and the total lack of an urban culture of their own, they all are endangered" (Novák 2006)
- Medvedeva/Arkhangelskiy
  http://web-corpora.net/UdmurtCorpus
- EuroBabel Khanty, Mansi
  http://www.babel.gwi.uni-muenchen.de
- Tavda Mansi
  http://norbertszilagyi91.wix.com/tawdamansi
- Nganasan
  https://www.slm.uni-hamburg.de/ifuu/forschung/forschu
- FinnUgReviat Udmurt Mansi
  http://www.ieas-szeged.hu/finugrevita

Assessment

- If you had an app that spoke a dead language, what would be its impact?
- (Question asked of 8 people under 16, all totally absorbed in their smartphones)
- Answer 1: nothing, what could I do with it? (5)
- Answer 2: how would you even know it wasn't phony? (3)
- **Giallatekno** Sami (Northern Southern, Skolt, Kildin, Ter, Pite), Komi, Kven, Erzya, Moksha
- Many efforts for standard Estonian, Finnish, Hungarian
- Now see https://acl-sigur.github.io/matrix.html

Judging the quick and the dead

Preservation (Still → Heritage) versus **(Re)Vitalization** (Borderline → Vital). These tasks require different approaches (philological versus socio-political); take different linguistic expertise (classical versus modern); involve different technologies; etc.

Most of the efforts summarized so far mix these two tasks to the detriment of both.

Cultural background

lg	sil	orth	tweets	wp-act	mac
Erzya	myv	cyr	no	16	yes
Moksha	mdf	cyr	no	16	yes
Karelian	krl	no	no	inc	yes
Permyak	koi	cyr+	no	15	no
Zyrian	kpv	cyr+	no	1	no
Meadow Mari	mhr	cyr	no	24	no
Hill Mari	mrj	cyr	no	16	no
Udmurt	udm	cyr+	yes	16	yes
Northern Sami	sme	lat	yes	24	yes

Immediate action items: build joint Urali Cyrillic+ keyboard, get FireFox support, explore smartphone usage issue

Uralic conclusions

- Uralic fares a lot better than languages in general: 6 languages (10%) vital out of 62
- Estonian is sitting pretty
- Effort on borderline languages is worthwhile!

Metaphors

- The biological metaphor is flexible
- Can be saved by adding a stage of pupation/transformation between life and death
- Primary use of the metaphor is from the the death stage, cf ‘heat death’ (entropy wins) in complex systems
- Can one day your computer converese with you in Language X? This is what is at stake here. A dead language with a good cuneiform script has a better chance than one currently on the digital periphery

Thank you for your attention