Improving digital vitality on the cheap

András Kornai

Hungarian Academy of Sciences

META-FORUM, July 5 2016

Acknowledgements

Katalin Pajkossy (BUTE)

Plan of the talk

- Digital vitality in Europe
- The danger zone
- The cheap way forward

What to measure

- 'European' defined geographically, broader than EU.
- Brexit notwithstanding, we consider the European idea as the only way towards a Europe that is livable both for the minorities inside, and those unfortunate to be outside the political borders of the EU.
- Geographic criterion yields 283 languages
- This number excludes historical languages like Old Norse
- 41 are sign languages, excluded from the study

How to measure

- Main idea: select seeds whose classification is known in advance.
- Only 4 classes: Thriving, Vital, Heritage, Still
- Find seeds that everybody would agree on, e.g. Spanish, German, French are thriving; Czech or Romanian are vital; Latin or Old Church Slavonic are heritage; any language with no digital footprint is still. Avoid hard cases like Basque
- Collect lots of data on standard and digital vitality
- Build classifiers by supervised machine learning
- Kornai 2013: Digital language death PLoS ONE 8(10): e77056. doi:10.1371/journal.pone.0077056

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

	Seed 0				Seed 1			
# feat	SH-VT	S-H-VT	SH-V-T	S-H-V-T	SH-VT	S-H-VT	SH-V-T	S-H-V-T
33	95.0	99.3	92.3	90.7	99.3	98.6	94.3	87.9
18	97.2	99.3	91.4	96.4	99.3	98.6	95.0	89.3
10	97.9	99.3	92.9	95.7	100.0	99.3	93.6	90.0
8	97.1	99.3	92.9	97.1	100.0	96.4	94.3	85.7
6	97.1	99.3	92.1	93.6	100.0	96.4	95.7	89.3

doi:10.1371/journal.pone.0077056.t001

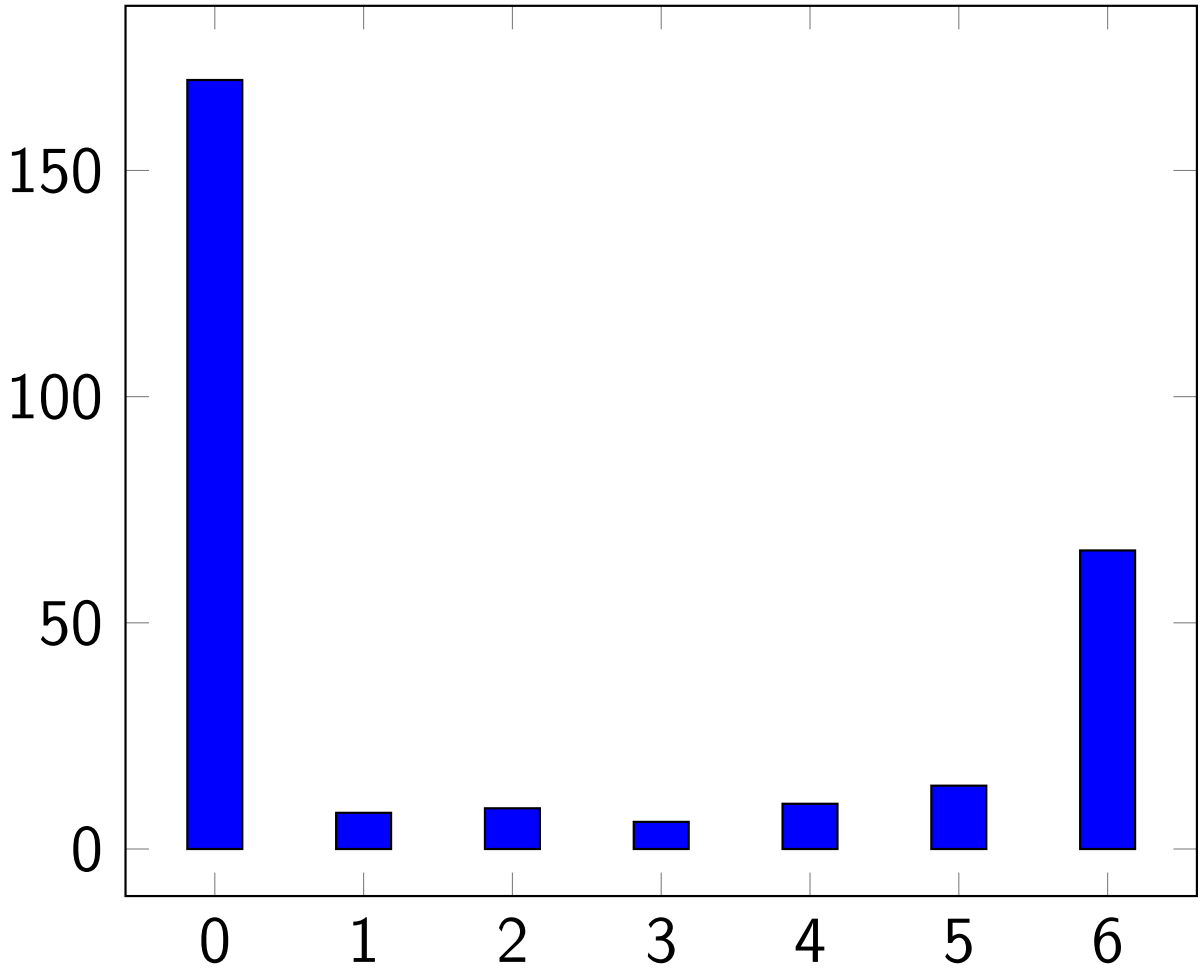
Added twist: feature selection

- 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

599	L2
526	wp_real_articles
525	cru_docs
427	wp_adjusted_size_macro
364	wp_edits
306	wp_articles
305	L1
223	indi_tweets
108	cru_words
101	indi_words
87	wp_total
32	wp_adjusted_size
25	la_oth_res_in_all
9	wp_edits_macro
3	la_primary_texts_all
1	la_primary_texts_online
1	la_oth_res_in_online

Borderline cases

- Not a category in the analysis!
- Statistical methods are hard to apply to individuals
- But we can obtain robust statistical conclusions



The quick

Bashkir	0	Bosnian	0	**Bulgarian**	0
Catalan	0	Chuvash	0	**Croatian**	0
Czech	0	**Danish**	0	**Dutch**	0
English	0	Faroese	0	**Finnish**	0
French	0	Friulian	0	**Galician**	0
German	0	**Hungarian**	0	**Icelandic**	0
Italian	0	**Lithuanian**	0	Luxembourgish	0
Macedonian	0	**Maltese**	0	**Greek**	0
Neapolitan	0	**Norwegian** Bokmal	0	**Norwegian** B+N	0
Ossetian	0	**Polish**	0	**Portuguese**	0
Romanian	0	Russian	0	**Serbian**	0
Slovak	0	**Slovenian**	0	**Spanish**	0
Swedish	0	Ukrainian	0	Venetian	0
Chechen	1	Eastern Mari	1	Lower Sorbian	1
Mirandese	1	Silesian	1	Swiss German	1
Võro	1	Yakut	1	Asturian	2
Kashubian	2	Latgalian	3	Picard	3
Scots	3	Sicilian	5	Tatar	5
Belarusian	6	**Basque**	10	Upper Sorbian	10
Walloon	10	Breton	11	Occitan	11
Piemontese	12	Lak	14	Scottish Gaelic	17
Welsh	17	Crimean Tatar	18	Western Frisian	18

The dead

Abaza, Achterhoeks, Aghul, Akhvakh, Alutor, Andi, Angloromani, Arbëreshë Albanian, Archi, Arvanitika Albanian, Bagvalal, Baltic Romani, Bezhta, Botlikh, Caló, Campidanese Sardinian, Carpathian Romani, Chamalal, Chukot, Chulym, Cimbrian, Dargwa, Dido, Dolgan, Drents, Eastern Frisian, Emilian, Erromintxela, Even, Fala, Forest Enets, Gallurese Sardinian, Ghodoberi, Gilyak, Gronings, Hinukh, Hunzib, Inari Sami, Ingrian, Ingush, Istriot, Istro Romanian, Itelmen, Judeo-Italian, Judeo-Tat, Jutish, Jèrriais, Kalo Finnish Romani, Karagas, Karaim, Karata, Karelian, Ket, Khakas, Khanty, Khvarshi, Kildin Sami, Koryak, Krymchak, Kumyk, Kven Finnish, Ladin, Liv, Livvi, Logudorese Sardinian, Lower Silesian, Ludian, Lule Sami, Mainfrnkisch, Mansi, Mednyj Aleut, Megleno Romanian, Minderico, Mócheno, Nanai, Naukan Yupik, Negidal, Nenets, Nganasan, Nogai, Northern Altai, Northern Yukaghir, Oroch, Orok, Pite Sami, Polari, Prussian, Quinqui, Romagnol, Romano-Greek, Romano-Serbian, Rutul, Sallands, Selkup, Shelta, Shor, Siberian Tatar, Sinte Romani, Skolt Sami, Slavomolisano, Southern Altai, Southern Sami, Southern Yukaghir, Stellingwerfs, Swabian, Tabassaran, Tavringer Romani, Ter Sami, Tindi, Traveller Norwegian, Traveller Scottish, Tsakonian, Tundra Enets, Tuvinian, Twents, Udihe, Ulch, Ume Sami, Upper Saxon, Veluws, Vlaamse Gebarentaal, Vlax Romani, Votic, Walser, Welsh Romani, Western Yiddish, Westphalien, Wymysorys, Yeniche

In the danger zone

Romansh	21	Vlaams	23
Adyghe	24	Ligurian	26
Udmurt	26	Russia Buriat	27
Corsican	29	Aragonese	35
Macedo-Romanian	35	Komi-Permyak	37
Irish	38	Northern Sami	38
Bavarian	39	Lombard	39
Standard Latvian	42	Balkan Romani	44
Rusyn	49	**Standard Estonian**	51
Tosk Albanian	52	Northern Frisian	56
Saterfriesisch	57	Komi-Zyrian	58
Zeeuws	58	Limburgan	60
Kölsch	67	Karachay-Balkar	70
Avaric	73	**Norwegian Nynorsk**	74
Extremaduran	75	Erzya	82
Gagauz	85	Pontic	85
Western Mari	86	Kalmyk	89
Manx	92	Lezghian	93
Sassarese Sardinian	99	Low German	104
Gheg Albanian	107	Veps	110
Moksha	113	Tornedalen Finnish	113
Kabardian	114	Samogitian	115
Arpitan	124	Cornish	125
Pfaelzisch	126		

First, the dialects

Warning!

Speaker knows nothing about dialectology and has no data

- Often vigorous, but unlikely to become digitally vital
- Etymological relations are not useful for native speakers
- But remain a considerable source of regio-national identity
- Exactly one dialact in Sápmi, Northern Sami
- Exactly one dialact of Gaelic, Irish
- Perhaps more than one dialect of German?

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

The upward path

1. Coordinate bid for fundraising/crowdsourcing (EUR 500)
2. Identify speaker community (EUR 500)
3. Give 50 people smartphone subscription 200 hrs spoken + unlimited text (EUR 10k)
4. Subsidize development/tuning of language ID (EUR 2k)
5. Subsidize development/tuning of rough phoneme reco (EUR 5k)
6. Subsidize development/tuning of unsupervised morphology (EUR 3k)
7. Create lexicon development website (EUR 3k)
8. Publish results (1k)

Thank You