Acta Linguistica Academica Vol. **64** (2017) 3, 327–345
DOI: 10.1556/2062.2017.64.3.1

Digital vitality of Uralic languages

Judit Ács
Department of Automation and Applied Informatics
Budapest University of Technology and Economics
judit@mokk.bme.hu

Katalin Pajkossy
Oktafone Ltd.
pajkossy@mokk.bme.hu

András Kornai
Institute for Computer Science and Control
Hungarian Academy of Sciences
andras@kornai.com

Abstract: We investigate the digital vitality of Uralic languages and dialects, and discuss how existing approaches to language revitalization relate to this model.

Keywords: machine learning; digital language vitality; Uralic

1. Introduction

In Kornai (2013) we presented a general statistical method for assessing the digital vitality of the world's languages and dialects. Here we apply a slightly improved version of the same method, on updated data, to the Uralic family. Section 2 describes the method and the improvements, section 3 presents the results, and section 4 discusses the main implications.

Traditionally, language vitality assessment is the sole domain of the experts, who collect demographic data about the population of speakers, the level of education available in the language, and other factors (cultural, political, etc.) that affect language vitality. Their judgment is usually expressed on a standardized scale such as EGIDS (Lewis & Simons 2010) or in terms of aptly named categories such as "vulnerable, definitely endangered, severely endangered, critically endangered, extinct" (Moseley 2010). While expert judgments are in many ways superior to machine-generated results, it is generally very hard to find experts deeply familiar with many languages at the same time (a small family like Uralic is already problematic in this respect), and the process is highly subjective.

2559–8201/$ 20.00

In our work we treat the problem of assessing digital vitality as a classification task amenable to **supervised machine learning** techniques. In principle our technique is equally applicable to the task of assessing traditional vitality, but in practice we simply don't have access to the relevant (demographic, political, cultural etc.) data on a global scale in a homogeneous format that would permit cross-language comparisons and statistical model building.

While computational linguists are very familiar with machine learning, the experts of Uralic languages generally come from the more philological tradition of linguistics, and to make this work more accessible to them, we use 1.1 to outline, however sketchily, the supervised learning paradigm. Conversely, the Uralic experts will find little that is new to them in 1.2, where we describe our selection of Uralic languages and dialects in terms of making the data more easily accessible to the computational linguist.

1.1. Machine learning

By a **classification** task we mean a potentially infinite set of inputs $x_1, x_2, \ldots$, each of which must be assigned to exactly one of the classes $C_1, C_2, \ldots, C_k$. The inputs can display an infinite variety, but the number of classes is finite, typically quite small, often just two (binary classification). Imagine that the task is to find, based entirely on acoustic input, those utterances that the grammarian would classify as questions. Based on written, properly punctuated text this is trivial: sentences that end in a question mark are questions, the others are not. Based on intonation contour, word order, and the presence of *wh* words (as provided e.g., by a speech recognition system) the problem is much harder, especially if (as in real life) the output of the recognizer is not error-free.

The key idea of **supervision** is to provide the machine learner with example inputs where the output is already known, e.g., because we have access not just to the acoustic data but also to expert transcription that will unambiguously show whether the utterance is a question. For example from *Why, this is beautiful!* the learning algorithm will learn that the presence of the sentence-initial *wh*, otherwise a very strong clue to questionhood, is not an absolute indicator, just something with a high evidential weight. In the so-called **maximum entropy** learner that we apply, the algorithm finds the optimum weight of each measured factor (called **feature** in machine learning) that will make the output of the classifier maximally consistent with the human-generated supervisory data (also known as **ground truth** or **gold** data). Once the algorithm is **trained** on such data, it produces

models of the classes in terms of the evidential weights that each feature has for each class. These models can be **tested** on data for which the ground truth was known but not used during training (held back data) or simply run on novel data to obtain machine classification for those cases where the ground truth was not known.

1.2. Uralic

Table 1: Uralic languages/dialects and their ISO codes

>Enets	—	Tundra Enets	enh	Forest Enets	enf
>Estonian	est	Estonian Standard	ekk	Estonian Võro	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	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

The primary registry of languages is SIL International, which publishes the Ethnologue both on paper (now in its 18th edition, Lewis et al. 2015) and online (http://www.ethnologue.com), and maintains the 3-letter International Standards Organization code ISO 639-3 (http://www-01.sil.org/iso639-3/ download.asp). Table 1 lists all 63 candidates our survey of the literature provided, together with their code where available.

Where there is no ISO 639-3 code, the Linguist List code is substituted, see http://linguistlist.org/forms/langs/find-a-language-or-family.cfm.

There are several discrepancies between this list of 63 entries and the final list of 54 languages and dialects that we considered in our work. First, we excluded [urj], the code for the entire Uralic family. Such codes are of course eminently useful in identifying resources that pertain to the entire family, but correspond to no actual community of speakers whose (digital) vitality can be assessed. (We discuss the situation of the major branches in 4.2). Next there are cases like Mansi [mns] and Selkup [sel], whose constitutive dialects are treated individually and found not to be digitally viable, but digital heritage preservation efforts for the respective group as a whole are in a relatively advanced stage.

A similar problem is seen for Enets and Mordvin, for which SIL International doesn't even provide collective codes, quite justifiably in light of the lack of mutual intelligibility between the branches. Clearly, forced lumping together of Erzya and Moksha in a Mordvin superclass would create an artificial population, perhaps justifiable on historical grounds, but flying in the face of the synchronic situation where neither group particularly self-identifies as Mordvin. From the standpoint of digital vitality, the fact that Moksha has a viable Wikipedia and good foundations (dictionaries) for long-term survival in no way helps Erzya.

Where there is better mutual intelligibility, particularly between the main dialects Permyak [koi] and Zyrian [kpv] of Komi [kom], or between the Estonian macrolanguage [est] and standard Estonian [ekk], there is, perhaps inevitably, a significant risk of confusing data that pertains to the entire group and the subgroups. For this reason [est] and [ekk] score very similarly, but we interpret this result as the vitality of standard Estonian, not the Estonian macrolanguage as a whole.

Higher groupings, whether synchronically justified or not, are marked by a prefixed > in Table 1. Another prefix we use is *D* (dead) for dialects only known from historical descriptions (no native speakers). The *Ethnologue* is uneven in its coverage of extinct languages/dialects (Hammarström 2015), and three of the four lacunae we found, Mescherian, Muromian, and Kainu Sami, were long extinct before the 1950 cutoff point the *Ethnologue* editors aim at. The last missing entry is the Seto dialect of Estonian, with an estimated 12,500 native speaker population. Here the lack of ISO 639-3 is actually limiting the data collection effort, and we can provide no machine-generated result, but note that the Seto's larger cousin Võro is already borderline, so the chances of Seto acquiring digital vitality are rather slim.

2. Methodology

Here we describe the four classes and the gold data we use (2.1), the features (2.2), and how we evaluate the model (2.3). The actual numerical algorithm used to compute the weights is not discussed here, readers interested in the details should consult e.g., Ratnaparkhi (1997). The data and the software are freely available at http://github.com/kornai/langdeath.

2.1. Classification and gold data

Our chief interest is with long-term survivability in the digital domain, but a simple binary classifier (will it survive? yes/no) is not nearly as informative as the traditional language vitality classes: for example EGIDS has 13 different values, with the vital/non boundary running between 6a "vigorous" and 6b "threatened". Machine learning techniques are in principle capable of making the fine distinctions required to sort languages in 13 or even more classes, but only if the data required for doing so, e.g., political factors affecting language policies, could be put at their disposal in some clear format. As a compromise, we use only four classes: Thriving (T), Vital (V), Heritage (H), and Still (S).

Thriving languages are generally what EGIDS would class 0 "global", e.g., Spanish, German, or French. Remarkably, the whole Uralic family does not contain a single global language, quite possibly because none of the nation states where these languages dominate built a colonial empire. Vital languages are like Czech and Romanian, with clear digital survivability on the one hand, but no chance of becoming one of the lingua francas of the digital world. The Uralic family has several of these, see 3.1.

Heritage languages have good digital presence, with many texts preserved, often with computational linguistic tools already available, but no native speakers (L2 population only): Latin or Old Church Slavonic are good examples. As with thriving or vital languages, all our training examples come from outside the Uralic family, but the trained classifiers find several that fit into this class, see 3.2.

Finally, digitally Still languages exhibit no signs of digital vitality: there is no Wikipedia or any other community where these languages are used for digital communication. Again, we made sure that the training examples are chosen from outside the Uralic family, yet the results make it clear that there are several still languages within Uralic, see 3.3.

In this regard, our method is clearly objective: we have not pre-judged the matter for any language or dialect by including it in the ground truth.

It is also objective in the sense that we give no special status to Uralic, the same methods are applicable e.g., to the languages of the Indian subcontinent (Kornai & Bhattacharyya 2014). There are other reasons for believing that the method is quite robust – these will be discussed in 2.3.

At this point, it is worth recalling from Kornai (2013) that the definitions given above are ostensive, provided primarily for concept formation, rather than criterial (extensive). We want the reader to know how we selected e.g., our Heritage training data, but it is up to the machine learner to realize that these datapoints generally involve zero L1 population. Remarkably, the learner succeeds in doing so, even if the training includes Sanskrit, for which over ten thousand native speakers are listed in our population data source, the *Ethnologue*.

2.2. Features

We use three main groups of features: those pertaining to vitality in the traditional (pre-digital) sense, those that mark the the existence of some online resource, and those pertaining to software support. Features in the first group include the number of L1 and L2 speakers, and the traditional (expert-based) assessment on the 13-point EGIDS scale (extracted from the *Ethnologue* database https://www.ethnologue.com) and from the Endangered Languages Project (http://www.endangeredlanguages.com), which uses a 10-point scale (the higher the more threatened).

Features marking online resources include those extracted from the Crúbadán Project https://crubadan.org which collects native language tweets and blogs, see http://indigenoustweets.com and http://indigenousblogs.com; from OmniGlot (https://omniglot.com), which concentrates on native scripts (writing systems); the Open Language Archives Community (http://www.language-archives.org); the World Atlas of Language Structures (http://wals.info); the Uriel compendium (http://www.cs.cmu.edu/~dmortens/uriel.html); the Leipzig Corpora (http://corpora.uni-leipzig.de/en); and Wikipedia (WP). Features related to WP proved to be particularly useful, especially the one denoting that the given language has a wikipedia incubator, and the wikipedia size in characters (adjusted by the unigram character entropy of the given language).

The final group contains features denoting software support, including operating system language packs and supported keyboard layouts (for OSX, Microsoft Windows and Ubuntu); Office 13 language pack and Firefox language tools (language packs and dictionaries); these also proved to

be useful indicators for digital vitality in a sense that will be made more precise in 3.5.

Since the digital world moves fast, it is worth emphasizing that all of our data were gathered from the most current versions of these sources (as of March 2016). Altogether, over sixty features are collected for each language, but not all are used, because we employ **feature selection**, an automatic method for deciding which of the features are actually contributing to the classification.

2.3. Evaluation

As in earlier work, we model the dependence of the classes on the features using maximum entropy (logistic regression) with first selecting the set of useful features using $l1$-norm based regularization Pajkossy (2013). The main novelty is a more refined treatment of borderline cases (see 3.4).

We evaluate our method both with regards to **external** and **internal** consistency. By external consistency we mean automatically producing results that agree well with expert judgment. This was only informally assessed (during and after the work was presented to an expert audience at IWCLUL2 in January 2016) but no major discrepancies were found. By internal consistency we mean not just that the models should test well on the training data, and using leave-one-out tests, but also robustness, whether the results are highly dependent on the exact training data, or whether they capture more general phenomena.

We report results on 200 experiments, each performed with the following setting: we randomly chose a subset of out training data (15 examples of each class), train a maximum entropy model using it, and perform classification of all our data. We aggregate the votes of the 200 classifiers, and we assign a language to the vital/heritage/still class if it has at least $\sim 95\%$ of the given vote.

We evaluated our models using cross-validation (internal consistency); the average accuracy is 0.9501, with a standard deviation of 0.0254. We report the categorization of 56 Uralic languages and dialects, from which we assign 6 to the vital (see 3.1), 14 to the heritage (see 3.2), and 20 to the still class (see 3.3). This leaves several **borderline** languages where the statistical evidence is not strong enough to make an individual determination with high confidence (see 3.4).

After $l1$-norm based feature selections the models use 10–11 features on average (see 3.5, where we list those selected in more than 10% of the cases). Internal consistency is seen from the high end of this distribution

(the same ten features are used more than half of the time), and external consistency is seen from the fact that the trained weights are easily interpretable.

3. Results on Uralic languages and dialects

3.1. Vital languages

We begin with the Vital class, which contains six languages. Perhaps the only surprise is Udmurt, but even there all indicators of vitality are strong.

Table 2: Vital languages

Name	Votes
Hungarian	200
Finnish	200
Estonian	198
Northern Sami	193
Eastern Mari	190
Udmurt	188

3.2. Heritage languages

Equally clear are the heritage languages. Some of them, like Forest Enets, still have a handful of native speakers, but they are generationally older, so the L1 population is on its way out. We call attention to Karelian, where there are sufficient L1 speakers for the *Ethnologue* to consider the language viable, while its digital fate is clearly sealed.

3.3. Still languages

This group contains a large variety of languages that are dead or dormant (no living speakers): Kamas (including Koibal); Southern Khanty; Livonian; Mator; and Yurats (as well as those extinct languages where the analysis could not be carried out for lack of standard code: Mescherian, Muromian, and Kainu Sami). Also included are many digitally still languages that are at various stages of endangerment in the traditional sense

Table 3: Heritage languages

Name	Votes	Name	Votes
Votic	200	Khanty	199
Kildin Sami	200	Selkup	198
Liv	200	Nenets	198
Tornedalen Finnish	200	Karelian	197
Kven Finnish	200	Southern Sami	196
Mansi	199	Ingrian	192
Khanty	199	Forest Enets	189

– for these we list in parentheses, following Campbell & Hauk (2015), the number of L1 speakers together with the date of the census where these numbers were obtained.

The following are considered "critically endangered": Forest Enets (∼10/2011); Tundra Enets (∼30/2007); Akkala Sami (1/2013); Pite Sami (∼42/2012); Ume Sami (20/2007); Central Selkup (2/2015); Southern Selkup (1/2015); Votic (∼12/2015); Yazva (∼200/2007).

The following are "severely endangered": Ingrian (∼130/2013); Kven Finnish (2-8k/2005); Eastern Khanty (∼480/2010); Eastern Mansi (<500/2000); Kildin Sami (∼300/2007); Nganasan (500/2000); Ter Sami (30/2007) and Veps (1600/2010).

The following are "endangered": Inari Sami (∼300/2007); Northern Selkup (<600/2005); Southern Sami (600/2015). We emphasize that the above designations refer to traditional vitality assessments – digitally these languages are all still, with practically no chance of (re)vitalization.

3.4. Borderline languages

This group contains all languages (15) with no clear class; see Table 5 for the list of vote counts for each class. On the top of the list we find four languages close to the vital category with 10–20% of historic votes; these are Western Mari, Komi, Moksha and Erzya. Here the vital signs are quite strong (see section 4 for recommendations).

At the bottom panel of Table 5 we find 5 languages that congregate on the historic-still border, with Skolt Sami being closer to the still class; Livvi straddling the border (with 40%-60% in between); and 3 closer to the heritage class (Nganasan, Tundra Enets, Ter Sami). For these 5, re-

Table 4: Still languages

Name	Votes	Name	Votes
Yurats	200	Northern Mansi	200
Yazva	200	Western Mansi	200
Southern Selkup	200	Eastern Mansi	200
Central Selkup	200	Ludian	200
Northern Selkup	200	Southern Khanty	200
Akkala Sami	200	Eastern Khanty	200
Kemi Sami	200	Northern Khanty	200
Tundra Nenets	200	Kamas	200
Forest Nenets	200	Pite Sami	197
Mator	200	Ume Sami	196

vitalization efforts make little practical sense, but heritage preservation is still very much an option (see section 4).

In the middle of the table we find six languages with mixed votes; these show similarity with all sets of training examples: Lule Sami, Permyak, Veps, Võro, Inari Sami, Zyrian.

Table 5: Borderline languages

Name	Vital	Still	Historic
Western Mari	179	1	20
Komi	179	0	21
Moksha	172	3	25
Erzya	160	5	35
Lule Sami	54	57	89
Komi-Permyak	47	113	40
Veps	38	18	144
Võro	35	18	147
Inari Sami	26	87	87
Komi-Zyrian	12	179	9
Skolt Sami	0	158	42
Livvi	0	83	117
Nganasan	0	46	154
Tundra Enets	0	34	166
Ter Sami	0	21	179

While the three classes themselves are relatively clearly delineated, ranks within a class are not particularly reliable, and we caution against over-interpretation of the individual scores.

3.5. Selected features

The $l1$-norm based selection mechanism kept 10–11 features on average, with four features used in all experiments (Wikipedia incubator, number of L1 speakers, *Ethnologue* status, Crúbadán word count), and another two used in more than 80% of the experiments (WP adjusted size, Omniglot).

The remaining features relevant in at least 10% of the cases are the following, in order of decreasing importance. The availability of the Bible (typically only the New Testament) as listed in watchtower.org; being listed as having covered by a dictionary or lexicography projects at (https://github.com/RichardLitt/endangered-languages); having a corpus available in the Leipzig Corpora Collection (a good sign); a dedicated Language Pack in Office 13; its status assigned by Endangered Languages Project; having an input method in Windows 10; a Language Pack in FireFox; Ubuntu (linux) language pack and input support; having the Universal Declaration of Human Rights translated to the language; the availability of features in Uriel.

Table 6: Features selected in more than 10% of the experiments

wp incubator	200	office13 lp	99
L1	200	endangeredlang proj. status	92
ethnologue status	200	win10 input method	81
cru words	200	firefox lpack	50
wp adjusted size	188	ubuntu pack	45
omniglot	162	ubuntu input	36
newtestament	131	udhr	36
dic/lex. work	120	uriel feats	24
leipzig corpora	115		

Let us now inspect the weights associated with the top six features, averaged across each type of model, as listed in Table 7. We find that the model has learnt plausible weights: for example the existence of a WP incubator has a negative weight for vitality, since vital languages have already moved from the incubator to the full wikipedia stage, and positive weights for the other two classes.

The number of native speakers is positively related to vital, and very negatively to heritage status. This makes good sense since heritage languages are very unlikely to have native speakers, though Sanskrit is often claimed as an exception, and the model is presented in the training phase with L1 data from the *Ethnologue* that shows 15,770 native speakers.

Similarly, *Ethnologue* status is measured in EGIDS value, where the higher number means lower vitality, so the indication is strongly negative for vital, and positive for still and heritage classification. Word counts, be they from the Crúbadán crawl or from Wikipedia, are positive indicators of vitality, and negative indicators of still/heritage status. Finally, mention in Omniglot indicates some level of native literacy, which is a positive indicator for vital and heritage languages, but negative for still.

Table 7: Weights associated with top features

	Vital	Still	Heritage
wp incubator	−0.58	−1.3	1.56
L1 speakers	0.89	0.68	−2.0
ethnologue status	−1.8	0.54	0.3
crubadan word count	0.37	−1.59	0.92
wp adjusted size	0.56	−1.36	0.49
omniglot	0.31	−0.87	0.27

4. Current work and future directions

Altogether, the Uralic family is considerably better off than the world's languages in general: here we estimated that roughly one in 10 language is vital, while globally the ratio of vital languages is well below 5% (Kornai 2013). But 10% is not exactly a reason to celebrate: the loss will be tremendous, and we see an entire branch of Uralic, Samoyedic, as getting completely wiped out by the transition to the digital realm.

In 4.1 we distinguish preservation from (re)vitalization, and survey the current work in both. In 4.2 we turn to the dialect situation, and in 4.3 we list some "action items" we see as critical for improving the digital vitality of the Uralic family. Some general conclusions are offered in 4.4.

4.1. Current preservation and revitalization efforts

To some extent, even huge losses like that of Samoyedic are masked by brilliant philological work providing excellent grammatical sketches such as Simoncsics (1998), but these can hardly mitigate the fact that we will never have a gigaword corpus which provides the empirical lifeblood of modern linguistics, both theoretical and applied. To quote from Kornai (2013):

> "Just as the dodo is no less extinct for skeleta, drawings, or fossils being preserved in museums of natural history, online audio files of an elder tribesman reciting folk poetry will not facilitate digital ascent, and both still and heritage languages are digitally dead in the obvious sense of not serving the communication needs of a language community."

We emphasize the difference between **preservation,** moving languages from the still to the heritage category; and **(re)vitalization,** moving languages from borderline to vital. These tasks require different approaches (philological versus socio-political); take different linguistic expertise (classical versus modern); involve different technologies; etc. The central preservation effort, clearing up and digitizing old fieldwork collections, is surveyed in Campbell & Hauk (2015). The computational efforts we summarize below mix the preservation and the (re)vitalization tasks, we believe to the detriment of both.

We begin with the joint effort of Morphologic and the Hungarian Academy of Sciences to create two-level morphological analyzers for Khanty, Komi, Mansi, Udmurt, Mari, and Nganasan, of which Novák (2006) wrote: "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". Our results show a 3–3 split, with Udmurt, Komi, and Western Mari ready for (re)vitalization, but all dialects of Khanty, Mansi, and Nganasan a lost cause. This is not to deny the value of morphological analysis for cleaning up field notes and for heritage preservation in general, but as long as resources are limited, the communication needs of live communities should take priority. For these, morphological analysis is essential, since it provides stemming (thus aiding the dictionary building process and enabling information retrieval applications) and serves as the basis of all higher level language technologies we discuss in 4.3.

For revitalization, we mention the Medvedeva/Arkhangelskiy Udmurt corpus http://web-corpora.net/UdmurtCorpus. Remarkably, there is a great deal of heritage corpus work, such as EuroBabel for Khanty and

Mansi http://www.babel.gwi.uni-muenchen.de; the work on Tavda Mansi http://norbertszilagyi91.wix.com/tawdamansi; and work on Nganasan http://www.slm.uni-hamburg.de/ifuu/forschung/forschungsprojekte.html.

Another project suffering from mixing the heritage work (on Mansi) with genuine revitalization (of Udmurt) is FinUgRevita http://www.ieas-szeged.hu/finugrevita, and the same can be said about the standard-setting Giellatekno work that seems to dedicate the same effort to all languages/dialects spoken in Sápmi, of which only Northern Sami (as opposed to the Southern, Skolt, Kildin, Ter, and Pite dialects they also cover) is digitally survivable. http://giellatekno.uit.no has sections devoted to the borderline vital Komi, Erzya, and Moksha, but also to Kven. There are many efforts for standard Estonian, Finnish, and Hungarian, but as these languages are vital we see no need to survey these.

4.2. Dialects and koinés

The major languages of the internet all underwent the process of koiné formation: English from the 15th century, Spanish and German from the 16th, French from the 18th, Italian from the 19th, and so on. In the digital age, even the mega-community of Chinese speakers found it necessary to pave over the differences between the Guóyǔ (Traditional) and Pǔtōnghuà (Simplified) versions of Mandarin so that a unified pluri-centric Wikipedia can be maintained. From a linguistic perspective, the opposite process, manifested e.g., in the emergence of regional language/dialect WPs, is a huge advance. Communities are formed, speakers are (re)discovered, and data is accumulating at an unprecedented rate.

Yet as we assess the digital vitality of the major branches, the picture emerging is far less cheerful. The WPs are often Potemkin villages, with actually contentful articles vastly outnumbered by template-based, often machine-generated pages. As an example, consider the two languages at the middle of Table 7 (borderline languages), Võro and Veps. The raw number of pages (5,417 and 5,176 respectively) look good, but the ratio of substantive to short articles is below 5%, and the number of edits, a good measure of activity, is less than 3% of the Estonian (resp. Finnish) Wikipedias.

We already discussed the complete loss of the Samoyedic branch, and our conclusions in regards to Sami are only slightly less pessimistic: we believe that there is exactly one vital dialect, Northern Sami. Koiné formation is a painful process, requiring the sacrifice of much that gives distinctiveness to the minor dialects, but in this case we see no option: either

an effort is made to create a digital koiné, or speakers of the other dialects are forever banned from accessing the digital realm in a near-native fashion.

In the Permian branch, Udmurt is vital, and Komi (Permyak, but not Zyrian) can perhaps be vitalized. This echoes our conclusion from Kornai (2013):

> "Evidently, what we are witnessing is not just a massive die-off of the world's languages, it 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. As an example, consider Komi, with two wikipedias corresponding to the two main varieties (Permyak, 94,000 speakers and Zyrian, 293,000 speakers), both with alarmingly low (<2%) real ratios. Given that both varieties have several dialects, some already extinct and some clearly still, the best hope is for a koiné to emerge around the dialect of the main city, Syktyvkar. Once the orthography is standardized, the university (where the main language of education is Russian) can in principle turn out computational linguists ready to create a spellchecker, an essential first step toward digital literacy (Prószéky & Novák 2005). But the results will benefit the koiné speakers, and the low prestige rural Zyrian dialects are likely to be left behind.
>
> What must be kept in mind is that the scenario described for Komi is optimistic. There are several hundred thousand speakers, still amounting to about a quarter of the local population. There is a university. There are strong economic incentives (oil, timber) to develop the region further. But for the 95% of the world's languages where one or more of these drivers are missing, there is very little hope of crossing the digital divide."

The Mordvin branch is handicapped by two salient facts: first, that one cannot really hope for a koiné to pave over the differences between Erzya and Moksha, and second, that the language policies of the Russian Federation, quite supportive e.g., in Ugra, are more restrictive in Mordovia. We discuss some possible action items in 4.3.

The Mari branch gives rise to a more optimistic assessment: both Western (Hill) and Eastern (Meadow) Mari are low vital (or high borderline) cases, and koiné formation (based on Eastern Mari) is already taking place.

The Khantyic branch has the advantage that Ugra is a showcase of economic development, but this also works to its disadvantage, in that the population is increasingly Russified, in spite of good state-level support for native language education. Our model predicts heritage status for all dialects, a loss comparable to that of Samoyedic.

The Hungarian branch, with loss of peripheral dialects, is vital. The Finnic branch has two vital languages, Finnish and Estonian, but no vital dialects, in spite of the efforts (Veps and Võro in particular) that we discussed above.

4.3. Possible action items

Given the widespread cultural changes, the gradual dissolution of nomadic lifestyles, and the destruction of habitat, some 90% of Uralic languages and dialects are digitally still, and the only ones automatically making the transition are the national languages of modern industrial states: Finnish, Estonian, and Hungarian. We have identified three other candidates where a great deal of effort can make a difference: Sami (a koiné based on the Northern dialect), Mari (a koiné based on the Eastern dialect), and Udmurt. To get there, we need to make a few advances, well within the reach of current technology.

Perhaps surprisingly, the foundation stone of all digital vitality is a mundane piece of software in everyday use by billions of people, the spellchecker. Literacy is essential, and will remain essential because the higher stages of the natural language processing software stack presented in Table 8 remain unreachable without building the lower levels first.

Table 8: The NLP hierarchy

NLP capacity	vitality required
Intelligent text understanding, question answering	T
Machine Translation	T–T and T–V pairs only
Automatic Speech Recognition	V
Optical Character Recognition	V H
Functional sentence parsing	V
Probabilistic language models	V
Phrase-level analysis (chunking)	V
Word-level analysis (morphology)	V H S

Starting at the top, imagine that we wish to bring the medical advice capabilities of IBM's Watson system to, say, Meänkieli speakers. Since the system will justify its answers by citing the relevant medical literature, written predominantly in English, we need Machine Translation from English to Meänkieli, and of course if we have high quality MT in both directions we don't have to build another Watson.

However, high quality MT depends on functional sentence parsing, which in turn depends on both probabilistic language modeling and on chunking. For languages with complex inflectional morphology (not just

Uralic) we need to combine the morphological information with the probabilistic model (Bilmes & Kirchhoff 2003) for every task such as chunking (for a Hungarian example see Recski 2014) and of course spellchecking already benefits from morphological analysis (Hajic & Drózd 1990; Oflazer & Guzey 1994; Németh et al. 2004). Indeed, spellchecking and stemming are the primary motivational examples behind the two-level phonology and morphology mechanism (Koskenniemi 1983) that figures prominently in several of the projects surveyed in 4.2.

Standardized spelling is a big part of koiné formation, but even the heritage preservation projects must find methods to homologize the transcription systems used by different workers in the field. For (re)vitalization, elicited texts are methodologically secondary, and orders of magnitude smaller, than "live" production such as blogs, tweets, and WP articles, and to make use of the live material we must find a way to normalize the words.

Our first proposed action concerns text production at the lowest level: we need input methods, and especially for those Uralic languages and dialects that use some kind of extended Cyrillic, we need to develop a keyboard layout. Perhaps the simplest would be to design a joint Uralic Cyrillic+ keyboard, one that contains all special symbols (typically, accented versions of standard Cyrillic characters) used in writing these, rather than one keyboard per language. Once this is in place, a reasonable second step is to get any language/dialect whose digital vitalization is a concern be supported by FireFox.

Another line of action concerns audio. As researchers we could provide native speakers with cellphones and free time on condition that whatever they say will be recorded and used for research purposes – the Linguistic Data Consortium already makes available several CALLHOME corpora collected by this method. For now, the primary goal should be collecting the data, transcription efforts can come later. We estimate a good collection effort, yielding several hundred hours of untranscribed audio, could be run on EUR 10k per language, considerably less than the costs of getting a single trained linguist in the field.

Obviously, it is not our place to prioritize exactly which languages and dialects should be targeted by live data collection efforts of the kind suggested above, but we hope our work helps to raise consciousness and focus the attention of national governments and funding agencies on this urgent problem.

5. Conclusions

Studies of digital vitality assessment share few of the central concerns of sociolinguistics. We have no data on social stratification (in a few cases, we see a clear urban-rural split), prestige, variation, linguistic change, etc. in the digital realm, and consider the digital variant unified, except when they are machine-distinguishable, as Nynorsk and Bokmål in the earlier study. Distinguishability is perhaps the lowest bar to cross in terms of digital survivability, and it should be kept in mind that in spite of the concentrated effort of the Crúbadán Project, we can identify only about 200 languages by standard tools like TextCat or CLD2 for lack of data. That said, in principle our method could be put to use on actual digital sociolinguistic data, and the "logic of variable rules" (Kay & McDaniel 1979) makes clear the conceptual relatedness of maximum entropy and logistic regression, so there is a connection at the level of mathematical tools as well.

Acknowledgements

We would like to thank Marianne Bakró-Nagy (HAS Research Institute for Linguistics), Johanna Domokos (Universität Bielefeld), and Anna Fenyvesi (University of Szeged) for advice on the subtler aspects of Uralic. Kornai is grateful to the organizers of IWCLUL2 for giving him the opportunity to give the keynote address whose final version appears here. We thank the anonymous referees whose comments led to many substantive improvements.

References

Bilmes, Jeff A. and Katrin Kirchhoff. 2003. Factored language models and generalized parallel backoff. In Proceedings of HLT/NACCL. 4–6.

Campbell, Lyle and Bryn Hauk. 2015. Language endangerment and endangered Uralic languages. In H. Mantila, K. Leinonen, S. Brunni, S. Palviainen and J. Sivonen (eds.) Proceedings of the Congressus Duodecimus Internationalis Fenno-Ugristarum. Oulu: University of Oulu. 7–38.

Hajic, Jan and Janus Drózd. 1990. Spelling-checking for highly inflective languages. In COLNG 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics. 358–360.

Hammarström, Harald. 2015. Ethnologue 16/17/18th editions: A comprehensive review. Language 91. 723–737.

Kay, Paul and Chad K. McDaniel. 1979. On the logic of variable rules. Language in Society 8. 151–187.

Kornai, András. 2013. Digital language death. PloS ONE 8. DOI 10.1371/journal.pone.0077056.

Kornai, András and Pushpak Bhattacharyya. 2014. Indian subcontinent language vitalization. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, Iceland: European Language Resources Association (ELRA). 24–27.

Koskenniemi, Kimmo. 1983. Two-level morphology: A general computational model for word-form recognition and production. Helsinki: University of Helsinki.

Lewis, M. Paul and Gary F. Simons. 2010. Assessing endangerment: Expanding Fishman's GIDS. Revue roumaine de linguistique 2. 103–119.

Lewis, Paul, Gary Simons and Charles Fennig (eds.). 2015. The ethnologue. Dallas, TX: Summer Institute of Linguistics.

Moseley, Christopher (ed.). 2010. Atlas of the world's languages in danger. Third edition. Paris: UNESCO Publishing. http://www.unesco.org/languages-atlas

Németh, László, Viktor Trón, Péter Halácsy, András Kornai, András Rung and István Szakadát. 2004. Leveraging the open source ispell codebase for minority language analysis. In J. Carson-Berndsen (ed.) Proceedings of the SALTMIL Workshop at LREC 2004. 56–59.

Novák, Attila. 2006. Morphological tools for six small Uralic languages. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06).

Oflazer, Kemal and Cemaleddin Guzey. 1994. Spelling correction in agglutinative languages. In Proceedings of the Fourth Conference on Applied Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics. 194–195.

Pajkossy, Katalin. 2013. Studying feature selection methods applied to classification tasks in natural language processing. Msc thesis. Eötvös Loránd University.

Prószéky, Gábor and Attila Novák. 2005. Computational morphologies for small Uralic languages. In A. Arppe, L. Carlson, K. Linden, J. Piitulainen, M. Suominen, M. Vainio, H. Westerlund and A. Yli-Jyrä (eds.) Inquiries into words, constraints and contexts (Festschrift in the honour of Kimmo Koskenniemi on his 60th birthday). Stanford: CSLI Publications. 116–125.

Ratnaparkhi, Adwait. 1997. A simple introduction to maximum entropy models for natural language processing. IRCS Report 97-08. Philadelphia: University of Pennsylvania.

Recski, Gábor. 2014. Hungarian noun phrase extraction using rule-based and hybrid methods. Acta Cybernetica 21. 461–479.

Simoncsics, Peter. 1998. Kamassian. In D. Abondolo (ed.) The Uralic languages. London & New York: Routledge. 580–601.