Language death in the digital age

András Kornai
HAS Computer Science Research Institute
and
Department of Computer Science, Boston University

@CSCS-CSM

June 28 2012

Plan of the talk

1. What is language death?
2. Why do we care?

The biological metaphor

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.
The Descent of Man (Darwin 1871: 67)

Background (linguistics)

- Languages are produced by children from unstructured input in a single generation (Bickerton 1981, Kegl 1994)
- Languages change (already known to Pāṇini)
- Languages die (though rarely by the sword – Nettle and Romaine 2000, Crystal 2002)
- There are about 6,000 languages (modulo difficulties in separating dialects)
- Vitality normally measured/projected on a hundred-year timescale
- Almost half (2,500 out of 6,000) of the world's languages are endangered

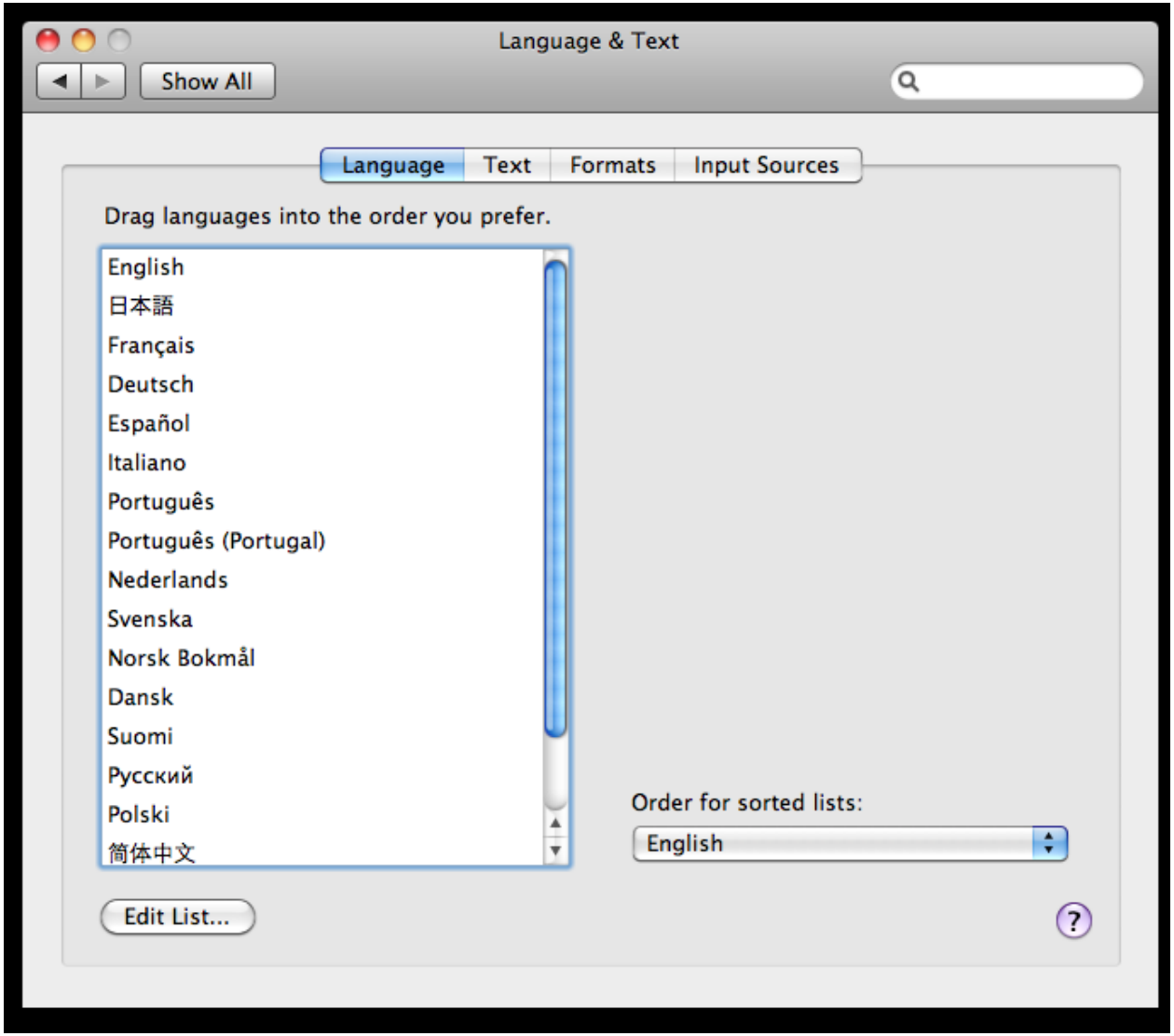
Defining language death

- Loss of function (other languages take over entire functional areas)
- Loss of prestige
- Loss of competence (emergence of semi-speakers) Menomini Bloomfield 1927, Gaelic Dorian 1981, Dyrbal Schmidt 1985

In the digital age:

- Loss of function performed digitally that is, increasingly, every function, from day to day communication (texting, email, ...) to commerce, official business, ...
- Loss of prestige If it's not on the web, it doesn't exist
- Loss of competence Can you raise a digital native in your language?

In the comfort zone



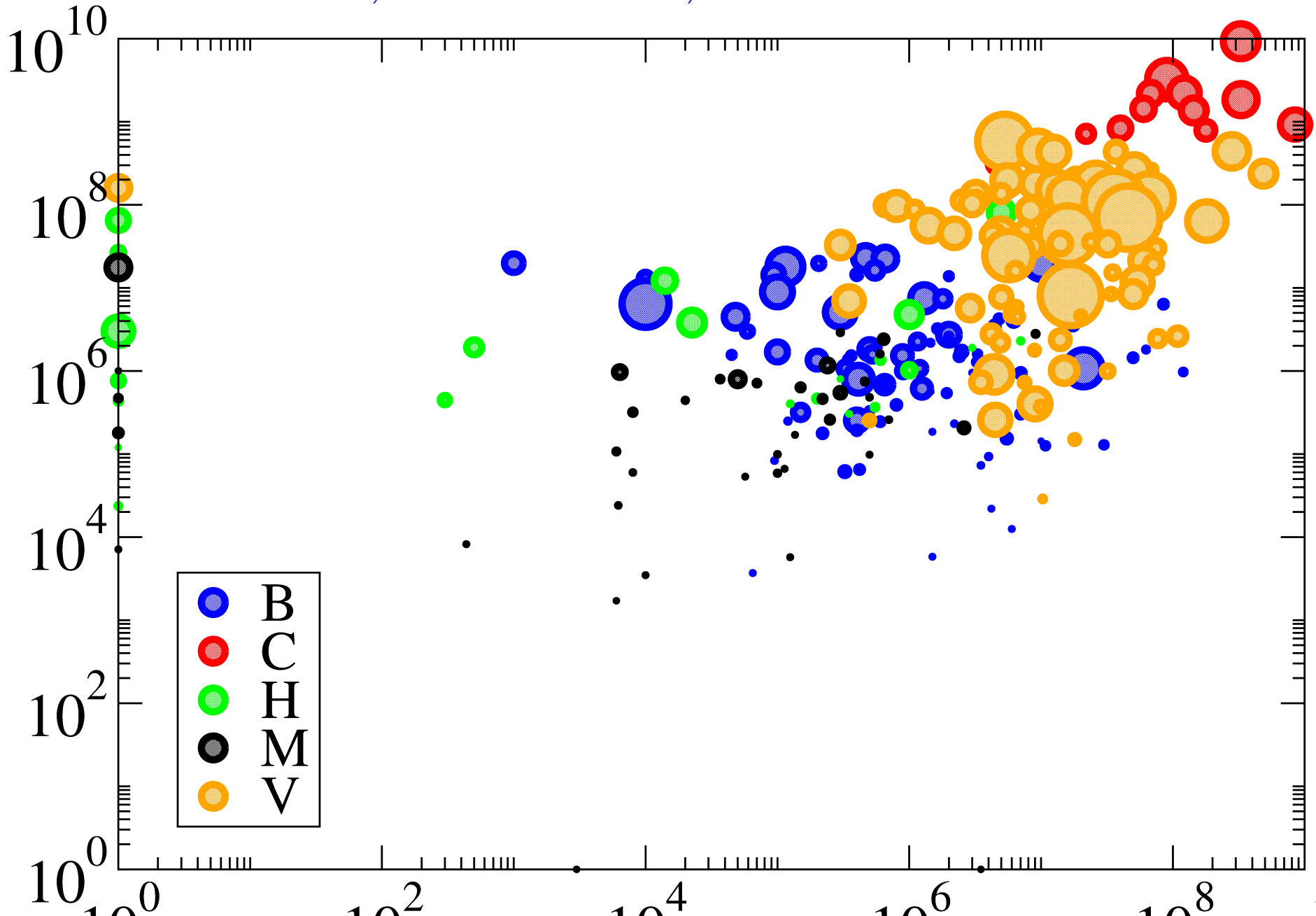
Full locale support, fonts, spellchecker, dictionaries, NLP tools

Must be FOSS – if it cannot be torrented it doesn't exist

Vital languages

- No wikipedia, no survival People know this – currently 340 proposals in incubator stage
- *But how good is a WP?*
- Estimate character entropy of language e.g. based on length of parallel texts. Filter out pages with longest paragraph shorter than 450 German chars *because these are weak/fake pages*
- Proportion of remaining pages gives real ratio, total (normalized) character count of real pages is adjusted WP size
- Potemkin wikipadias (e.g. Volapük) contribute nothing to survival
- Currently less than a hundred V+C

Community, WP size, real ratio



A four-way classification of languages C,V,H,M

C **In the comfort zone:** Wherever humanity goes, this language goes with them. 16 lgs, real ratio 0.36±0.10, #speakers 145m, WP 1.6g chars

V **Vital:** Significant digital community generating online material 83 lgs, real ratio 0.36±0.19, #speakers 31m, WP 96m chars

B **Borderline:** May yet make the transition to the digital age. 90 lgs, real ratio 0.15±0.13, #speakers 7m, WP 3.8m chars

- as a vital language B → V
- as a read-only carrier of cultural heritage B → H

H **Heritage:** 22 lgs, real ratio 0.14±0.13, #speakers 870k, WP 9.5m chars

M **Moribund or dead:** Digital natives cannot be raised. 41 lgs, real ratio 0.05±0.06, #speakers 840k, WP 970k chars

Borderline languages

- No community, no survival The WP language policy states that at least five active users must edit that language regularly before a test project will be considered successful
- A group of enthusiasts can do wonders, but cannot sustain a lively community. As the digital death of Gaelic, Nynorsk, etc makes clear, the communities are voting with their smartphones
- Passive (read only) web presence (lexicons, classical literature, news services) is no substitute for active use in a broad variety of two-way contexts (social networks, business/commerce, live literature/blogs, etc) Heritage preservation has *huge* value!
- For any language pair, Google Translate likes to see gigaword monolingual corpora and megaword parallel text Reasonable goal for vitalization projects

OK, SO 97-97% DIE. WHY DO WE CARE?

Background (formal languages)

Languages

An n-letter *alphabet* is just a set Σ of n distinct symbols (it doesn't matter what they are). A *string* of length k over these is just the concatenation of k symbols (one symbol can be used more than once). The set of all strings (including the empty string λ) over Σ is denoted Σ^*. A (formal) *language* is some $L \subset \Sigma^*$.

Grammars

Grammars are finite lists of string replacement rules of the form $\alpha \to \beta$ where $\alpha, \beta \in (\Sigma \cup N)^*$.

The extra symbols in N are called *nonterminals*

Grammars

0 recursively enumerable

1 context sensitive

1.5 mildly context sensitive

2 context free

3 finite state (regular, rational)

4 non-counting

5 finite (can be listed)

A recent controversy

March 20, 2012

Angry Words

Will one researcher's discovery deep in the Amazon destroy the foundation of modern linguistics?

Illustration by Steve Brodner for The Chronicle Review

By Tom Bartlett

A Christian missionary sets out to convert a remote Amazonian tribe. He lives with them for years in primitive conditions, learns their extremely difficult language, risks his life battling malaria, giant anacondas, and sometimes the tribe itself. In a plot twist, instead of converting them he loses his faith, morphing from an evangelist trying to translate the Bible into an academic determined to understand the people he's come to respect and love.

Along the way, the former missionary discovers that the language these people speak doesn't follow one of the fundamental tenets of linguistics, a finding that would seem to turn the field on its head, undermine basic assumptions about how children learn to communicate, and dethrone the discipline's long-reigning king, who also happens to be among

What is the Everett-Chomsky controversy?

Chomsky (1957, 1965, 1981, 1993, ...) Languages are essentially recursive (type 1.5 according to Stabler 2000)

Everett (2004) Pirahã has no recursion

HCF (2002) [Faculty of Language] comprises only the core computational mechanisms of recursion as they appear in narrow syntax and the mappings to the interfaces

Everett (2009) Other languages without recursion: Hixkaryana (600), Warlpiri (2670), Teiwa (spoken in th village of Lebang in mountainous interior of northeastern Pantar, as well as by residents of Madar, just south of Kabir), Dutch Sign Languag (20,000) all digitally moribund

The standard counting argument

The alphabet

Σ is *not* the set of words (which is generally considered infinite, cf. *great***grandfather, leges***legnagyobb*) but rather the set of lexical categories (Noun, Verb, Adjective, ...) Instead of *colorless green ideas sleep furiously* take Adj Adj N.PL V_I.3SG Adv

Used widely in generative linguistics:

> *(...) the traditional view (assumes) the (Latin) proto-language could not have complex sentences. (...) If this assumption were realistic, and the proto-language actually could not embed sentences inside others, it could easily be shown that this proto-language had only a finite number of sentences, unlike any natural language known to linguistics. (Lakoff 1968:5)*

Are languages finite?

> *Any grammar of a language will project the finite and somewhat accidental corpus of observed utterances to a set (presumably infinite) of grammatical utterances. (Chomsky 1957:15)*

Large English Corpora

BNC British National Corpus about a hundred million (10^8) words
W1T Google Corpus – summary stats of about 1T (10^{12}) words

BNC: 175 subcorpora 7k to 2m words, average sentence lengths from 6.1 to 28.8 (not counting punctuation as separate words) or from 7.1 to 32.2 (counting punctuation as contributing to sentence length). The grand average is 17.0 (or 19.2) words per sentence in the BNC, and 9.8 words/sentence (counting punctuation) in W1T. The smallest average sentence length is found in transcriptions of spoken material, the largest in legal texts. Sentences over 150 words not that rare

Subsentence length

Subsentence defined as comma-separated stretch. Average length is 7.6 (without punctuation) in BNC, close to the average sentence length in spoken materials. Even the legal material has only 8.6 words per subsentence, spoken material transcribed by slightly different conventions regarding the placement of commas can go as high as 12.7! English has 3.2 subsentences per sentence, Pirahã maybe 1.

> *In general, the assumption that languages are infinite is made made in order to simplify the description of these languages. If a grammar does not have recursive devices (...) it will be prohibitively complex. If it does have recursive devices of some sort, it will produce infinitely many sentences. (Chomsky 1957:23-24)*

A toy grammar

Two equiprobable states, one producing *great-* and the other *grandfather was a true pioneer.* Infinite language, produces *great*^{n} *grandfather was a true pioneer* with probability $1/2^{n+1}$ Average sentence length is 6. Grammar expresses an infinite set of meanings.

Now let's exclude from consideration every sentence with 9 or more repetition of *great*s. New grammar still characterizes 99.9% of the data, but more complicated (10 states instead of 2). More complex grammar only expresses finite set of meanings.

If the loss is truly this momentous, all the visceral reaction to Everett's interpretation of Pirahã is easy to understand, as it is our unsurpassed abilities at communication, many would say our very humanity, that is at stake.

Information content

2 bits on average per sentence generated by recursive grammar, 1.976 bits for max 9 grammar. English words (infinite stock because of recursion like *great-great-....-grandfather* carry 12.7 bits on average. A finite vocabulary of 7,000 equiprobable words carries more.

It is simply not true that a finite language must have smaller information carrying capacity than an infinite one.

Conclusions

- 97.5% of the world's languages are dying (150 out of 6,000 may make it to the digital age)
- Recursion is just a simple (but important) example, the scientific loss is devastating (not to speak of the general cultural loss)
- This is not somebody else's problem, it hits many Uralic languages and minority languages in Hungary as well

THANK YOU