Nyelvi életképesség a digitális világban

Kornai András

BME Algebra Tszk és MTA SZTAKI

2018.2.19

Köszönet és hála

- Bakró-Nagy Marianne (MTA Nyelvtudományi Intézet)
- Domokos Johanna (Universität Bielefeld)
- Fenyvesi Anna (Szeged)
- Pajkossy Katalin, Ács Judit (BME)

A nyelvi életképesség hagyományos értelmezése

- Alapja a biológiai metafora: „a nyelv élőlény” (a gondolat Herderre és Darwinra megy vissza)
- Komolyan kutatott terület, hatalmas nyelvészeti irodalma van: olyan neves nyelvészek foglalkoztak a kérdéssel mint Bloomfield, Crystal, Dixon, Dorian, Dressler, Fishman, Hagège, Hale, Hill, Krauss, Moseley, Nettle, Romaine, Thomason
- Az alapvető probléma a *generációk közti szakadás* (intergenerational disruption) amikor a felnövekvő generació már nem veszi át teljesen a nyelvet (Fishman 1991).
- A hagyományos időskála: 100 év, ennyire próbáljuk meg előrejelezni a veszélyeket
- 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.

A legfontosabb tényezők

- A nyelvközösség mérete és korosztályi összetétele
- A nyelv presztizse (így csak a szolgák beszélnek, így csak a parasztok beszélnek, így csak az öregek beszélnek)
- A nyelv mint az önazonosság tényezője *magyar vagyok, román vagyok, *ógörög vagyok* (bantu nyelvekben ez módszeresen feltűnik: Bugandában (ez Uganda egy tartománya) lakik a Baganda törzs, melynek tagjai a Muganda emberek, akik a Luganda nyelvet beszélik, és Kiganda szokásrendet tartanak
- Funkcionális területek (kereskedelem, oktatás, egészségügy, jog, vallás . . .)
- Írásbeliség, „magas” kultúra

Életképesség a digitális világban

- Ugyanezen kritériumok digitális megfelelői, pl. nyelvközösség helyett *hálózati (online) nyelvközösség*
- Kevésbé finom beosztás: 13 helyett csak 4 kategória:
  - Thriving (T) ‘viruló’
  - Vital (V) ‘életképes’
  - Heritage (H) ‘örökségi’
  - Still (S) ‘halott’
- Módszer: felügyelt (gépi) tanulás, maximum entrópia modell
- Eredmények: a 8,426 vizsgált nyelvből kevesebb mint 5% életképes, 95% (és ez még optimista) halott. Nem ‘veszélyeztetett, kihalásra ítélt’, hanem már ma (2013) digitálisan halott.
- Az indiai szubkontinensen 634 nyelv, ebből 36 életképes, 1 örökségi, 576 halott (2014)

Az urali nyelvcsalád

- 54 nyelv/nyelvjárás, ebből 6 élő (magyar, finn, észt, északi számi, keleti mari, udmurt), 14 örökségi, 20 halott (2017)
- Sok adatforrást (40+) használunk, de azt már maga az algoritmus dönti el, ezekből mennyit és milyen súllyal használ fel:

wp incubator	200	office 13 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		

Honnan tudjuk, hogy az algoritmus helyes eredményt ad?

- Ahol vannak ismereteink ott nagyon egybevág ezekkel
- Belső konzisztencia
- Értelmes súlyok

	élő	halott	örökségi
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

A digitális életképesség alapvető tényezői

bu-	országhatárok	*nem*
ba-	törzsi tudat	*igen*
mu-	egyéni elkötelezettség	*igen*
lu-	közös nyelv	*igen*
ki-	közös kultúra	*IGEN*


Két nyelvre van szükséged!
I
English
I
PYTHON
SPACE APPS

A digitális írásbeliség alapjai

- Valamennyire szabványosított helyesírás „Mindegy, hogy hogy, csak nehogy sehogy”
- i18n (**I**nternationalizatio**N**)
- nyelv-szintű támogatás
- operációs rendszer szintű támogatás

A nyelvtechnológiai piramis

technológia	életképességi szint
szövegértés, kérdés-megválaszolás	T
gépi fordítás	csak T-T és T-V párokra
beszédfelismerés	V
optikai karakterfelismerés	V H
funkcionális mondatelemzés	V
valószínűségi nyelvmodellezés	V
sekély mondattani elemzés	V
névelem-felismerés	V
szó-szintű elemzés (morfológia)	V H S

Köszönöm a figyelmet