



A nyelvtechnológia alapjai

2.

A szövegek kódolása és rendezése (meg egy kis nyelvstatisztika)

2019/2020. tanév, I. félév



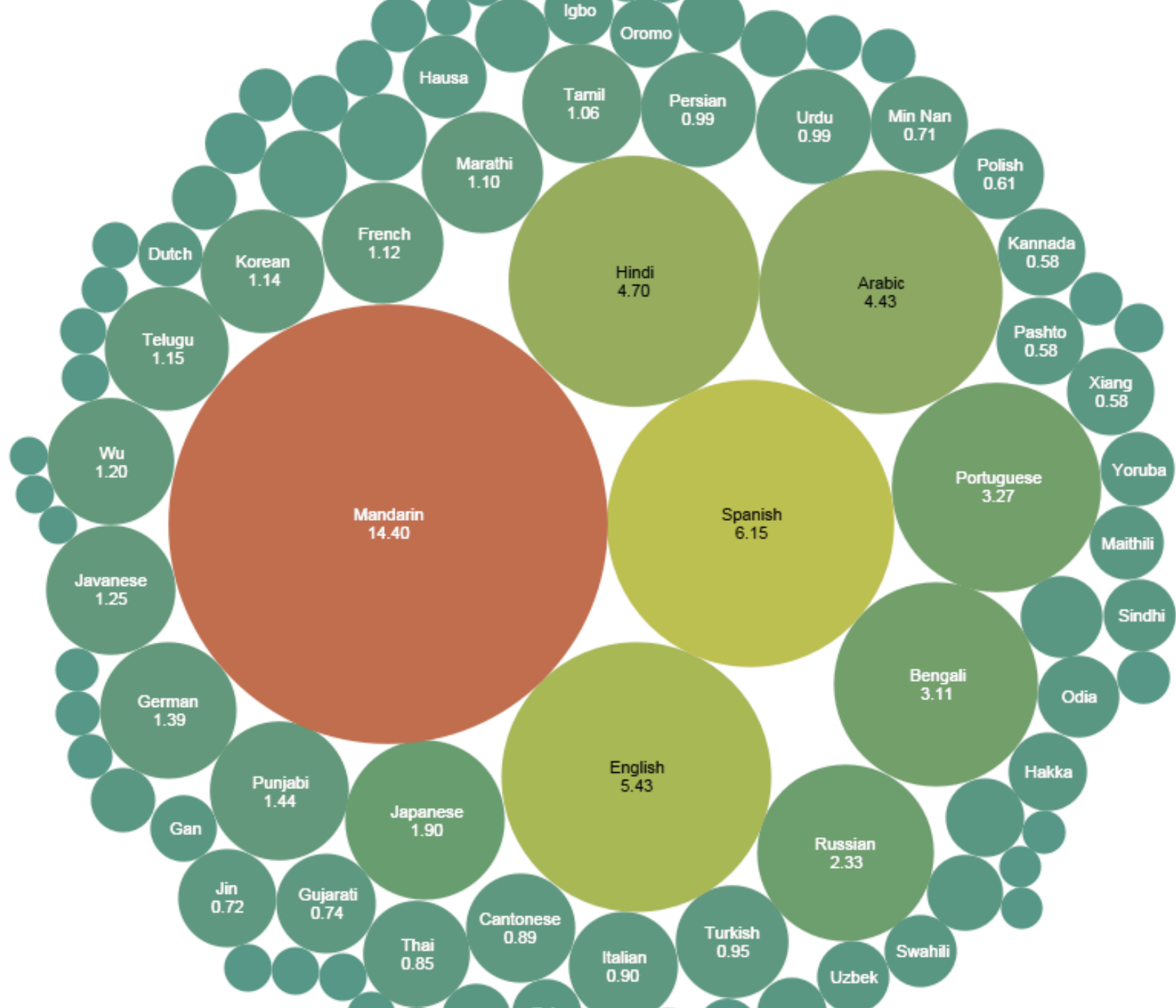
A nyelvekről



Nyelvek százmillió anyanyelvi beszélővel

(forrás: Ethnologue: Languages of the World, 2019)

	Nyelv	L1 beszélő	L1	L2 beszélő	L2	Összesen
1	Angol	379 millió	3	753 millió	1	1132 millió
2	Mandarin	918 millió	1	199 millió	5	1117 millió
3	Hindi	341 millió	4	274 millió	2	615 millió
4	Spanyol	460 millió	2	74 millió	10	534 millió
5	Francia	77 millió	14	203 millió	4	280 millió
6	Arab	-	-	274 millió	3	274 millió
7	Bengáli	228 millió	5	37 millió	14	265 millió
8	Orosz	154 millió	7	104 millió	7	258 millió
9	Portugál	221 millió	6	13 millió	16	234 millió
10	Indonéz	43 millió	29	155 millió	6	198 millió





A magyar nyelv „helye”

(forrás: Ethnologue: Languages of the World)

...

- 57. Cebuano (Fülöp-szk)
- 58. Sinhala (Sri Lanka)
- 59. Rangpuri (Bangladesh)
- 60. Északkelet-thai (Thaiföld)
- 61. Zhuang (Kína)
- 62. Malgas (Madagaszkár)
- 63. Nepáli (Nepál)
- 64. Szomáli (Szomália)
- 65. Khmer (Kambodzsa)
- 66. Madura (Indonézia)
- 67. Görög (Görögország)
- 68. Csittagóniai (Bangladesh)
- 69. Haryanvi (India)
- 70. Magahi (India)
- 71. Dekkan (India)
- 72. Magyar

...



A nyelvek területi megoszlása

(forrás: Ethnologue: Languages of the World)

Földrész	Élő nyelvek		Beszélők		ny -kénti átlag
	száma	%	száma	%	
Afrika	2 146	30,2	789 138 977	12,7	367 726
Amerika	1 060	14,9	51 109 910	0,8	48 217
Ázsia	2 304	32,4	3 742 996 641	60,0	1 624 565
Európa	284	4,0	1 646 624 761	26,4	5 797 975
Óceánia	1 311	18,5	6 551 278	0,1	4 997
Összesen	7 105	100.0	6 236 421 567	100.0	877 751

A nyelvek anyanyelvi beszélők szerinti megoszlása

(forrás: Ethnologue: Languages of the World)

Populáció	Élő nyelvek			Beszélők		
	száma	%	Σ %	száma	%	Σ %
100 000 000 to 999 999 999	8	0,1	0 ,%	2 308 548 848	40,20	40,20%
10 000 000 to 99 999 999	82	1,2	1,3%	2 480 078 977	39 42	79,62%
1 000 000 to 9 999 999	304	4,3	5,5%	951 916 458	14,55	94,18%
100 000 to 999 999	895	13,0	18,6%	283 116 716	4 75	98,84%
10 000 to 99 999	1 824	26,4	45,0%	60 780 797	1 01	99,86%
1 000 to 9 999	2 014	29,2	74,1%	7 773 810	0 13	99,99%
100 to 999	1 038	15,0	89,2%	461 250	0 00	99,99%
10 to 99	339	4,9	94,1%	12 560	0 00	99,99%
1 to 9	133	1,9	96,0%	521	0 00	100,00%
Ismeretlen	277	4,0	100,0%			
Összesen	6 909	100,0		5 959 511 717	100,0	



A nyelvi gépi feldolgozásának szintjei

A szó

- ❑ Definíciója nincs
- ❑ **Fonológiai szó**
(pl. *ház+am* fonológiai szó,
de az összetett szavak nem ilyenek)
- ❑ **Lexikai szó (lexéma)**
(pl. a *számítógépezés* szó nem lexéma,
mert levezethető alkotóelemei jelentéséből)
- ❑ **Morfológiai szó**
(pl. *láttam volna*)
- ❑ **Szintaktikai szó**
(pl. *szép+ség+e+im*,
de az igekötős ige nem ilyen,
mert a szintaxis megváltoztatja az alakját)

Még nehezebben azonosítható nyelvi egységek

- ❑ **Morféma:** a nyelv legkisebb jelentéssel bíró alkotórésze; nincs természetes határa
- ❑ **Szószerkezet, frázis:**
nincs természetes határuk; egyrészt lexémák, másrészt mondat alatti egységek
- ❑ **Megnyilvánulás, mondat:**
írásban egyszerűbb (nem nehéz a formális definíció)
- ❑ **Szöveg:**
mondatok sora, melyben az összefüggések túlmutatnak a mondatnyelvtanokon
- ❑ Megkülönböztetendő: **type** és **token**



Hang, fonéma, betű, karakter

Fonéma

❑ Szembenállás

Ha két beszédhang egyikének a helyére behelyettesítjük a másikat, és más szótagi egységet kapunk eredményül, a két beszédhangot két különböző fonémához soroljuk:

géz, kéz, méz, néz, réz -> /g/, /k/, /m/, /n/, /r/

❑ Kiegészítő eloszlás

Ha két beszédhang mindig csak egymást kölcsönösen kizáró környezetekben fordul elő, feltételesen ugyanazon fonémához sorolhatjuk őket:

pl. *hó : doh* -> ún. allofónok

Fonéma (2)

❑ Fonetikai hasonlóság

Az előző két ismerv önmagában nem zárja ki a téves fonémaazonosítás lehetőségét,
pl. a veláris nazális nemcsak a dentális nazálissal, hanem más hangokkal is a kiegészítő eloszlás viszonyában van, mégsem soroljuk sem pl. a /h/-hoz vagy a /j/-hez, hanem csakis az /n/-hez

❑ Szabad váltakozás

Ha két beszédhang ugyanazon környezetben egymással helyettesíthető anélkül, hogy a szóban forgó szótári elem azonossága veszélyben forogna, ugyanazon fonéma szabad változataival állunk szemben

Fonetikus ábécé

- ☐ Az IPA a latin (angol) ábécén alapszik
- ☐ A legtöbb latin betűs nyelv ugyanúgy ejt többet is: [b], [d], [f], [h], [k], [l], [m], [n], [p], [t], [v] stb.
- ☐ Vannak közöttük olyanok, melyek kiejtése a legtöbb latin betűvel író nyelvben eltér: [j], [r], [y] stb.
- ☐ Vannak olyan jelek, melyeket a latin betűk átalakításával hoztak létre, vagy írott formájukat használják: [ɛ], [ɜ] stb.
- ☐ Bizonyos elemeket a görög ábécéből vettek: [β], [ɣ], [ɛ], [θ], [χ], [v] stb.
- ☐ A magyar nyelv 14 magánhangzó- és 24 mássalhangzó-fonémája az IPA rendszerben is megjeleníthető

IPA-karakterkészletek

- ❑ SIL: Summer Institute of Linguistics
- ❑ Charis SIL: teljes IPA, tonális mellékjelek és sok más kevésbé általános fonetikai jel is
- ❑ Doulos SIL (a Charis SIL, csak bold és italics nélkül)
- ❑ SIL93: a korábbi SIL IPA93 betűtípusok
(Manuscript és Sophia) Unicode-ban újrakódolva

Az „extra kódolás” igénye

- ❑ Nem elég a latin ábécé
- ❑ Diakritikus jelek: „kidíszítjük” a meglevőket
- ❑ Multigráfok: kombináljuk a meglevőket
- ❑ Pl. amit a magyar „s” betűvel jelöl:
 - s magyar
 - sh angol, albán, cigány
 - ch francia, portugál
 - sch német
 - si gael, indonéz
 - sci olasz
 - skj svéd, norvég
 - š cseh, szlovák, horvát, szlovén, lett, litván, észt
 - sz/ś lengyel
 - ş román, török
 - Ŝ eszperantó



Európai ékezetes betűk (1)

Nyugat

à á â ã ä å æ ã è é ê ë ì í î ï ò ó ô

õ ö ø œ ù ú û ü û w ý ÿ ŷ

ç ħ ñ ŕ ş ŧ đ þ ß

Európai ékezetes betűk (2)

Kelet

áâäāąéëēėęěíîïıóôõö
ōóúûüūůűųý
çćĉčċďđġĝġġĥĥĵķĺłł
ńņñŕŕŕŝŝŝšt'žžž

Diakritikumok multigráfos kódolásai

❑ e-mail:

arvizturo tukorfurogep
(fokabel???)

❑ távirati:

aarviiztuerooe tuekoerfuuroogeep

❑ repülő ékezetes:

a'rví'ztu"ro" tu:ko:rfu'ro'ge'p

❑ számkódos:

a1rví1ztu3ro3 tu2ko2rfu1ro1ge1p

❑ TeX:

\'arv\'ízt\Hur\Ho t\'uk\'orf\'ur\'og\'ep

❑ HTML:

árvíztűrő

Karakter, karakterkód, karakterkészlet

- ☐ **Karakter:** absztrakt objektum, a számítógépen tárolt szövegek legkisebb, tovább nem bontható egysége, Egy karakter egy betű, szám, írásjel, speciális jel, illetve egyetlen egyszerűbb szövegformázó parancs ábrázolására alkalmas
- ☐ **Karakterkód:** a karakternek mint absztrakt objektumnak megfeleltetett számkód, amely számítógépen közvetlenül ábrázolható
- ☐ **Karakterkészlet:** karaktereket és számkódjaikat összerendelő táblázat, Meghatározza, hogy milyen karakterek használhatók, és azt is, hogy az egyes karakterkódokat miképpen kell értelmezni. A karakterkészlet megjeleníthető karaktereket és speciális műveleteket előíró vezérlőkaraktereket tartalmaz
- ☐ **Kódlap:** több karakterkészletet is kezelő számítógéprendszerben egy saját számkódjával azonosított karakterkészlet
- ☐ **Betűkészlet:** olyan táblázat, amely adott karakterkészlet(ek) megjeleníthető karaktereihez karakterképeket rendel, amelyek a számítógép képernyőjén vagy nyomtatásban jelennek meg. Az alkalmazott betűkészlet a szöveg számítógépbeli értelmezésére – például a szöveget nyelvi szempontból feldolgozó programokra – nincs befolyással, csak annak megjelenítésekor használatos

Kódlapok

ASCII kódtábla:

ASCII-Zeichensatz										
+	0	1	2	3	4	5	6	7	8	9
30				!	"	#	\$	%	&	'
40	(	)	*	+	,	-	.	/	0	1
50	2	3	4	5	6	7	8	9	:	;
60	<	=	>	?	@	A	B	C	D	E
70	F	G	H	I	J	K	L	M	N	O
80	P	Q	R	S	T	U	V	W	X	Y
90	Z	[	\	]	^	_	`	a	b	c
100	d	e	f	g	h	i	j	k	l	m
110	n	o	p	q	r	s	t	u	v	w
120	x	y	z	{		}	~			

Kódlapok (2)

Az ISO 8859-2 (Latin-2) kódtábla:

- ☐ magyar
- ☐ német
- ☐ horvát
- ☐ lengyel
- ☐ román
- ☐ szlovák
- ☐ szlovén
- ☐ szerb
- ☐ cseh
- ☐ albán

iso-8859-2										
+	0	1	2	3	4	5	6	7	8	9
160		À	Á	Â	Ã	Ä	Å	Æ	Ç	Š
170	Ð	Ė	Ž	-	Ž	Ž	°	ª	¸	ł
180	ˆ	İ	Š	ˆ	ˆ	š	š	ť	ž	˜
190	ž	ž	Ř	Á	Â	Ĥ	Ä	Ĺ	Ć	Ç
200	Č	É	Ê	Ë	Ě	Í	Î	Ď	Đ	Ñ
210	Ň	Ó	Ô	Õ	Ö	×	Ř	Ů	Ú	Ů
220	Ü	Ý	Ť	ß	ř	ă	â	ă	ä	ı
230	ć	ç	č	é	ę	ë	ě	í	î	ď
240	đ	ñ	ň	ó	ô	õ	ö	÷	ř	ů
250	ú	ü	ü	ý	ı	'				

A Unicode és a UCS

- ❑ UCS: Universal Character Set
- ❑ Unicode szabvány: 16 biten tárolt síkokra osztja a – szabvány legutóbbi változatában rögzített – kb. 100 ezer karaktert
- ❑ Minden beszélt, holt és egyéb nyelv (pl. Braille) karakterei (betűk, számok, írásjelek stb.)
- ❑ Minden egyes karakternek egyedi száma és neve van:
 - U+0041 Latin capital letter A
 - U+0151 Latin small letter o with double acute
- ❑ Eredetileg 2 bájton (16 biten) tárolta az adatokat (Basic Multilingual Plane) , de ez nem volt elég, ezért 4 bájtusra (32 bitesre) bővítették

A Unicode és a UCS

- ❑ UTF-8 (Unicode Transformation Format): változó hosszúságú kódolással (1–6 bájt) képezi le a Unicode karaktertáblát, pl. 1 bájton tárolt kódjai az ASCII-nak felelnek meg, így a latin betűs UTF-8 kódolású szövegek a régi ASCII környezetben is olvashatóak maradnak
- ❑ UTF-16: 1 112 064 karaktert kódol
- ❑ UCS-2 (elavult), 2 bájton (UTF-16 részhalmaza)
- ❑ UCS-4: 4 bájton (31-bit) = UTF-32, ISO 10646

A Unicode és a UCS

- ❑ 1980-as évek: igény egy egységes, minden karaktert leíró kódtáblára
 - **ISO 10646** – az ISO szervezet szabványa
 - **Unicode** – amerikai cégek konzorciuma
- ❑ 1991: egyesítik a két projekt eredményeit, de megmarad a két szervezet
- ❑ Teljesen kompatibilisek azóta is!
- ❑ Különbségek is vannak: az ISO 10646 csak karakterkészlet, a Unicode-ban viszont léteznek további kiegészítések, pl. egyeztetési szabályok

Egészen pontosan...

- ☐ ISO/IEC 10646-1:1993 $\approx$ Unicode 1.1
- ☐ ISO/IEC 10646-1:2000 $\approx$ Unicode 3.0
- ☐ ISO/IEC 10646-2:2001 $\approx$ Unicode 3.2
- ☐ ISO/IEC 10646:2003 $\approx$ Unicode 4.0
- ☐ ISO/IEC 10646:2003 és az 1. kiegészítés $\approx$ Unicode 4.1
- ☐ ISO/IEC 10646:2003 és
az 1. kiegészítés,
a 2. kiegészítés, és részben
a 3. kiegészítés $\approx$ Unicode 5.0
- ☐ ...
- ☐ Unicode 12.0 (2019. június)
- ☐ Unicode Character Database:
<http://www.unicode.org/Public/UNIDATA/>
- ☐ Továbbá: <http://www.decodeunicode.org/> (137 374 karakter)



Ábécék és rendezési elvek

Az ábécék „eleje”

Α Α Α χ | U T ɔ

1 2 3 4 5 6 7 8

अ अ म अ व् अ ए ए ञ ञ ॐ

9 10 11 12 13 14 15 16 17 18 19

あ ア ヲ 卜 阿

20 21 22 23 24

1. latin, 2. görög, 3. cirill, 4. héber, 5. arab, 6. örmény, 7. régi grúz, 8. mai grúz, 9. dévanagári, 10. bengáli, 11. gurmukhi, 12. gudzsaráti, 13. orijá, 14. tamil, 15. telugu, 16. kannada, 17. malajalám, 18. thai, 19. lao, 20. hiragana, 21. katakana, 22. bopomofo, 23. koreai, 24. kínai

Ábécék és rendezés

- A kiejtés szerinti írás (?)
- Az ábécébe rendezés elvei
- Átjárás ábécék között: transzliterálási problémák
- A könyvtári rendezés

Az olyan sajátos célú munkák (lexikonok, enciklopédiák, atlaszok és térképek névmutatói stb.), amelyekben magyar és idegen nyelvű szóanyag erősen keveredik egymással, rendszerint az úgynevezett általános latin betűs ábécét követik: *a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z*. Ebben a rendszerben mind a magyar, mind az idegen többjegyű betűknek minden egyes eleme külön, önálló egységnek számít, és a besorolás nincs tekintettel sem a magyar ékezetekre, sem az idegen betűk mellékjeleire. Az ilyenfajta munkák szavainak betűrendbe sorolása bonyolult feladat, itt például a több szóból álló egységeket vagy az egész kapcsolatot, vagy csak az első elem alapján szokás besorolni, ahogyan a könnyebben kezelhetőség érdeke vagy egyéb szempont kívánja. A követendő eljárást a könyvtárügy és a szakirodalmi tájékoztatás (dokumentáció) területén szabvány (MSZ 3401) írja elő, illetőleg belső szakmai útmutatók szabályozzák.

A magyar akadémiai rendezés (1)

A különböző betűvel kezdődő szavakat az első betűk ábécébeli helye szerint állítjuk rendbe, illetőleg keressük meg

Példa:

*acél, cukor, csók, gép, hideg, kettes, olasz,
öröm, Nagy, nyúl, remény, sokáig, szabad,
Tamás, vásárol*

A magyar akadémiai rendezés (2)

Az egyjegyű betűt teljesen elkülönítjük az azonos írásjeggyel kezdődő, de külön mássalhangzót jelölő kétjegyű (ill, háromjegyű) betűtől: mindig az egyjegyű betű van előbb

Példa:

*cudar, cukor, cuppant, csalit, csata, Csepel
dac, domb, duzzog, dzsem, dzsungel, Zoltán, zongora,
zúdul, zsalu, zseni*

A magyar akadémiai rendezés (3)

A többjegyű betűk kettőzött változatait sohasem az egyszerűsített alakok szerint soroljuk be a betűrendbe, hanem a megkettőzött betűt mindig két külön betűre bontjuk, s így soroljuk a szót a megfelelő helyre ($ccs = cs + cs$; $ggy = gy + gy$; $ddzs = dzs + dzs$)

Példa:

*kas, Kassák, kastély, kaszinó, kassza, kaszt
mennek, menü, menza, Menyhért, mennyi,
nagy, naggyá, nagygyakorlat, naggyal, nagyít*

A magyar akadémiai rendezés (4)

A magánhangzók rövid és hosszú változatát jelölő betűk (*a* – *á*, *e* – *é*, *i* – *í*, *o* – *ó*, *ö* – *ő*, *u* – *ú*, *ü* – *ű*)

a kialakult szokás szerint mind a szavak elején, mind pedig a szavak belsejében azonos értékűnek számítanak a betűrendbe sorolás szempontjából.

A magánhangzó hosszú változatát tartalmazó szó tehát meg is előzheti a rövid változatút

Példa:

ír, Irak, Irán, írandó, iránt, író, iroda, irónia

A magyar akadémiai rendezés (5)

A rövid magánhangzós szó kerül viszont előbbre olyankor, ha két szó betűsora csak az azonos magánhangzók hosszúsága tekintetében különbözik

Példa:

*égbolt, Eger, egér, éger, égés, egész,
keres, kérés, koros, kóros, kórós,
szel, szél, szelel, szeles, szelés, széles, szelet*

A magyar akadémiai rendezés (6)

A különírt elemekből álló szókapcsolatokat és az egybeírt vagy kötőjellel kapcsolt összetételeket minden tekintetben olyan szabályok szerint soroljuk betűrendbe, mint az egyszerű szavakat, A szóhatárokat tehát nem vesszük figyelembe. Ugyanez a szabály érvényes a közzavak közé besorolt tulajdonnevekre is

Példa:

*Kis részben, kissé, Kiss Ernő, kis sorozat,
kissorozat-gyártás, kis számban, kistányér, kis virág,
tiszafa, Tiszahát, Tisza Kálmán, Tisza menti, Tiszántúl, Tisza-
part, tiszavirág, tiszt*

A magyar akadémiai rendezés (7)

A betűrendbe soroláskor a szokásostól némiképp eltérően kezeljük a régies írású magyar családnevekben, valamint az idegen szavakban és tulajdonnevekben előforduló régi magyar, illetőleg idegen betűket

Példa:

Czibere, Czuczor, Cházár, Császár, Cselényi

Régies és idegen hangjelölés

- ☐ *aa* = á: *Gaal* [e : gál]
ch = cs: *Madách* [e : madács]
ee = é: *Veér* [e : vér]
cz = c: *Czuczor* [e : cucor]
eő = ö: *Eötvös* [e : ötvös]
s = zs: *Jósika* [e : józsika]
ew = ö: *Thewrewk* [e : török]
th = t: *Csáth* [e : csát]
oó = ó: *Soós* [e : sós]
ts = cs: *Takáts* [e : takács]
y = i: *Kölcsey* [e : kölcsei] stb
w = v: *Wesselényi* [e : veselényi] stb
- ☐ angol: *Greenwich* [e : grínics], *joule* [e : dzsúl]
- ☐ cseh: *Dvořák* [e : dvorzsák], *Škoda* [e : skoda]
- ☐ francia: *Eugène* [e : özsen], *Nîmes* [e : nim]
- ☐ német: *Schäfer* [e : séfer], *Werther* [e : verter]
- ☐ olasz: *Bologna* [e : bolonya], *quattrocento* [e : kvatrocsentó]
- ☐ portugál: *Guimarães* [e : gimarajs], *piranha* [e : piránya]
- ☐ román: *Tîrgoviște* [e : tirgoviste], *piața* [e : piaca]
- ☐ szerbhorvát: *Đurić* [e : gyúrity], *Živogošće* [e : zsvogostye]



Továbbá...

1-szer 1 az 1

II. János Pál

10 kérdés a személyi jövedelemadóról

11. Távhő-konferencia

XIX. század

88 kérdés az AIDS-ről

8000 germanizmus

ABALKIN, L. I.

Aczél György

ACÉLOK



Nyelvstatisztikai alapok

Betűstatisztikák

- ❑ Hangstatisztikák, betűstatisztikák
- ❑ Titkosírás-megfejtések
- ❑ Morse-ábécé, írógép-billentyűzet
- ❑ A nyelvek tipizálhatók n-gráfjaik alapján (pl. mássalhangzó-torlódással kezdődő szavaink régen egyáltalán nem voltak, és ma is csak bizonyosak vannak)
- ❑ Információ (Shannon): kisebb valószínűségű üzenetnek az információtartalma nagyobb, és független események együttes információtartalma az események információtartalmának összege

Egy kis információelméleti háttér

- ❑ Shannon (1949): tfh egy kommunikációs folyamatban veszünk részt, melynek csatornáján az X halmaz jeleiből összetevődő véges sorozatok, üzenetek áramlanak
- ❑ Ha elég sok üzenet áll rendelkezésre, akkor mérni tudjuk azt a $p(x)$ valószínűséget, hogy adott $x \in X$ elem milyen gyakran fordul elő várhatóan egy üzenetben
- ❑ Az x jel információtartalma: $I(x) := -\log_2[p(x)]$
- ❑ Az $\underline{X}$ üzenet információtartalma is:
ha $\underline{X} = (x_1, x_2, \dots, x_j) \in X^j$ jelek sorozata, akkor
$$H(\underline{X}) = H(x_1, x_2, \dots, x_j) = I(x_1) + I(x_2) + \dots + I(x_j) =$$
$$= -\log_2 p(x_1) - \log_2 p(x_2) - \dots - \log_2 p(x_j) =$$
$$= -\log_2 [p(x_1)p(x_2)\dots p(x_j)]$$
- ❑ Shannon (1951): Prediction and Entropy of Printed English

n-gramok

- ❑ n-gram: egymás után következő n db karakter egy szövegben
- ❑ Az n-gramok lehetnek szavak is, de nem feltétlenül azok
- ❑ Az n-gramok használatának előnye, hogy nyelvfüggetlen és hibatűrő megoldások megvalósítására igen alkalmas
- ❑ Stochastic Language Models (N-Gram) Specification:
<http://www.w3.org/TR/ngram-spec/>
- ❑ Microsoft Web N-gram Services:
<http://research.microsoft.com/en-us/collaboration/focus/cs/web-ngram.aspx>
- ❑ Google N-gram Viewer: <http://books.google.com/ngrams>

Magyar betűstatisztika (1-gramok)

	Gyakoriság	Információ bitben
A	9,35	3,43
Á	3,72	4,77
B	1,72	5,87
C	0,60	7,40
D	1,71	5,90
E	9,71	3,37
É	3,87	4,71
F	0,88	6,87
G	3,55	4,83
H	1,23	6,37
I	4,39	4,53
J	1,21	6,39
K	5,35	4,24
L	6,30	4,00
M	3,92	4,69

	Gyakoriság	Információ bitben
N	5,47	4,21
O	4,47	4,50
Ö	2,14	5,57
P	1,04	6,61
Q	0,00	?
R	4,22	4,58
S	6,57	3,94
T	7,87	3,68
U	1,29	6,30
Ü	0,93	6,77
V	1,81	5,81
W	0,00	?
X	0,01	13,33
Y	2,21	5,52
Z	4,46	4,50

n-gramok a gyakorlatban

Küpfmüller (1954) szöveggenerálása

1-, 2-, 3- és 4-gramokkal:

1. EME GKNEET ERS TITBL VTZEN
2. AUSZ KEINU WONDINGLIN DUFRN
3. PLANZEUNDGES PHIN INE UNDEN
4. ICH FOLGEMASZIG BIS STEHEN DISPONIN

n-gramok a magyarban

1. nórmígenytnaeőettesy sdrúk üze neée issgnis k, őentetnsmrddtca.
2. Bégiz érij, ercszépót, ekalfönembenyett kmáb os alőrre.
3. Ezenlapoltike lehátszökmég hog, aztradon két. Korcsant ván.
4. Hárone – Az a léhez, meg, akarózsák hogy öregész egény látán, járnagyon kérdeket egy emberet. Tordárd sem tan.
5. Senki adsaadott, hogy egy ki tornak, a kapta a mielőtte szor ismerni most vele aótsap a eipővel. - Ó, biztosítás úr csak társallatság vele. Hallj ott.
6. Történik. Telt év óta meglátta bevallott kissé lehajtott és mindig, légez erligyúneáenl eltűnt.

Megfigyelhető: a 4, után itt nem javul igazán a minőség, Miért?

Mert kicsi tanítókorpusz esetén (esetünkben 8 Mbájt) a négyesnél nagyobb csoportok eloszlásának megbízhatósága már nem elegendő

Zipf-törvény

- ❑ G. K. Zipf az 1930-as években vette észre: ha egy szövegben összeszámoljuk az egyes szavak előfordulási gyakoriságát, majd sorba rendezzük őket gyakoriságuk alapján, és ábrázoljuk a gyakoriságot rang szerint (azaz annak a függvényében, hogy az illető szó hányadik a sorban), akkor (közelítéssel) egy hatványfüggvény szerint lecsengő görbét kapunk -1-hez közeli kitevővel
- ❑ Tehát a lista i -edik elemének gyakorisága: $R_i \sim \frac{1}{i^\alpha}$
(α a korpuszra jellemző, 1 körüli konstans)
- ❑ Következmény: ha a leggyakoribb szavakat tartjuk csak meg, a korpusz nagy része megmarad
- ❑ A jelenség az emberi nyelvek univerzális tulajdonságának bizonyult: pl. hasonló viselkedést kapunk, ha nem a szavakat, hanem az összefüggő szöveg n -gramjait vizsgáljuk

Heaps-törvény

- ❑ Heaps (1978): egy N szóból álló korpusz esetén a különböző szavak száma:

$$V \approx K \cdot N^{\beta}$$

K : a szövegtől függő konstans ($10 \leq K \leq 100$)

β : a szövegtől függő konstans

(angolnál: $0,4 \leq \beta \leq 0,6$ - magyarnál: $0,6 \leq \beta \leq 0,7$)

- ❑ Következmény: egy korpuszban a különböző szavak száma új dokumentumok hozzáadásával folyamatosan növekszik (pl. az elírásoknak, a tulajdonneveknek és az új szavaknak köszönhetően), de ez a növekedés szublineáris, azaz az új szavak nem lesznek gyakori szavak

További statisztikai nyelvtörvények

- ☐ Mandelbrot: összefüggés van a szavak hosszúsága és rangja között is, ui. a rövidebb szavak minden nyelvben sokkal nagyobb gyakorisággal fordulnak elő, mint a hosszabbak, méghozzá a rangjuk a hosszal fordítottan arányos
- ☐ Észrevétel: ha a történelmi fejlődés során egy szó gyakorisága megnő, rendszerint meg is rövidül
- ☐ Következmény: nyelveink legősibb rétegét (pl. a testrészek neveit) szinte kivétel nélkül egytagú, két-három fonémából álló szavak alkotják
- ☐ „*Matematikai igazság, hogy kinek két szótagú neve van, fél annyi kortessel választhatik meg, mint az, kinél négy szótagot kell kiáltani*”

(Eötvös: A falu jegyzője)

Egy érdekességek

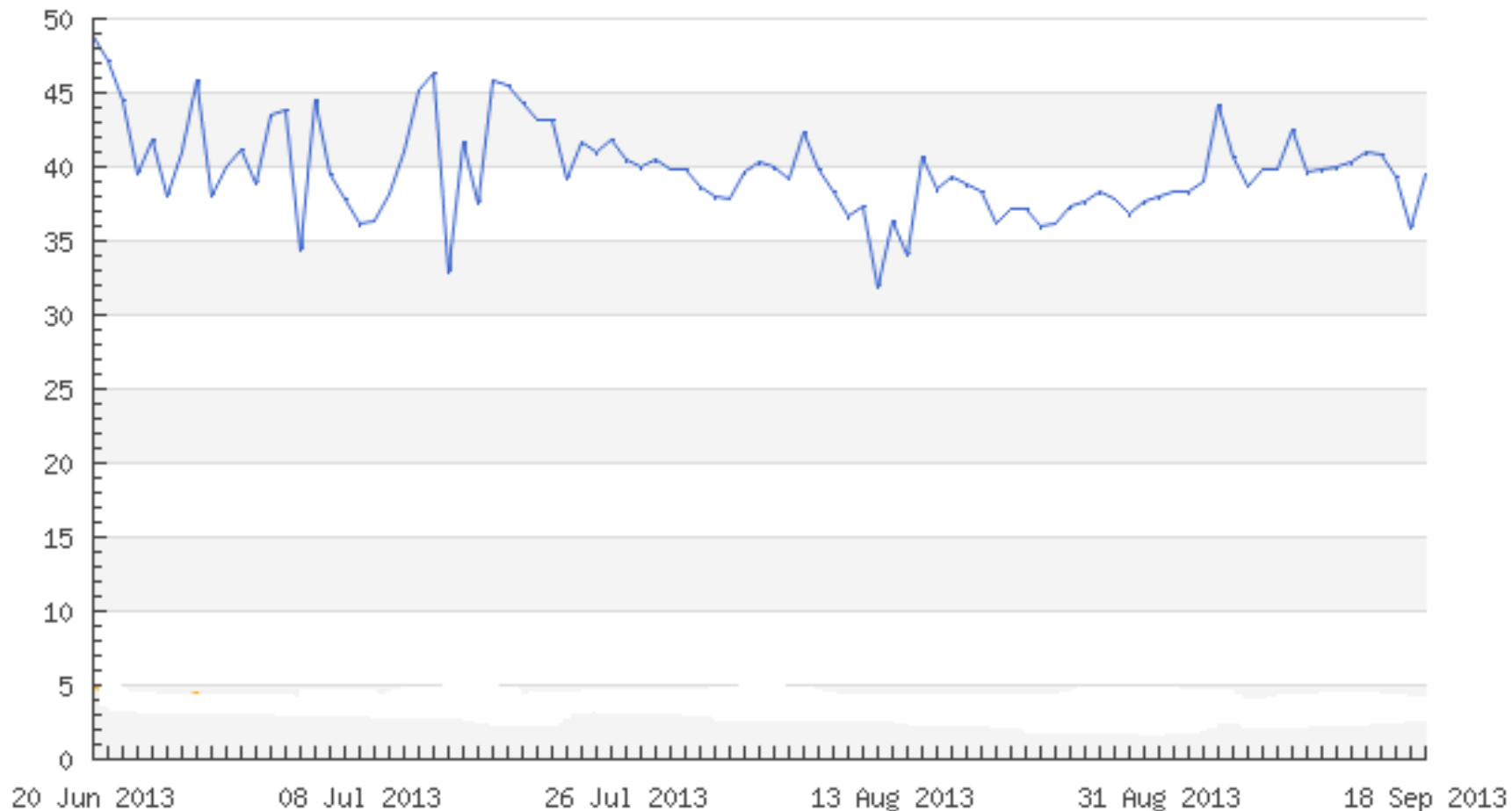
Egy angol nyelvű kutatói szerzők szerint nem számít meglepetésnek az a tény, hogy az első és az utolsó betűk a helyükön állnak. A többi betű lehet teljesen összevisszaság, mégis probléma nélkül olvasható a szöveg. Ennek oka, hogy nem olvasunk el minden betűt egyenként, hanem a szót olvasunk a maga egészében.



Szövegek és nyelvek a weben



Az indexelt web mérete (milliárd weboldalban)

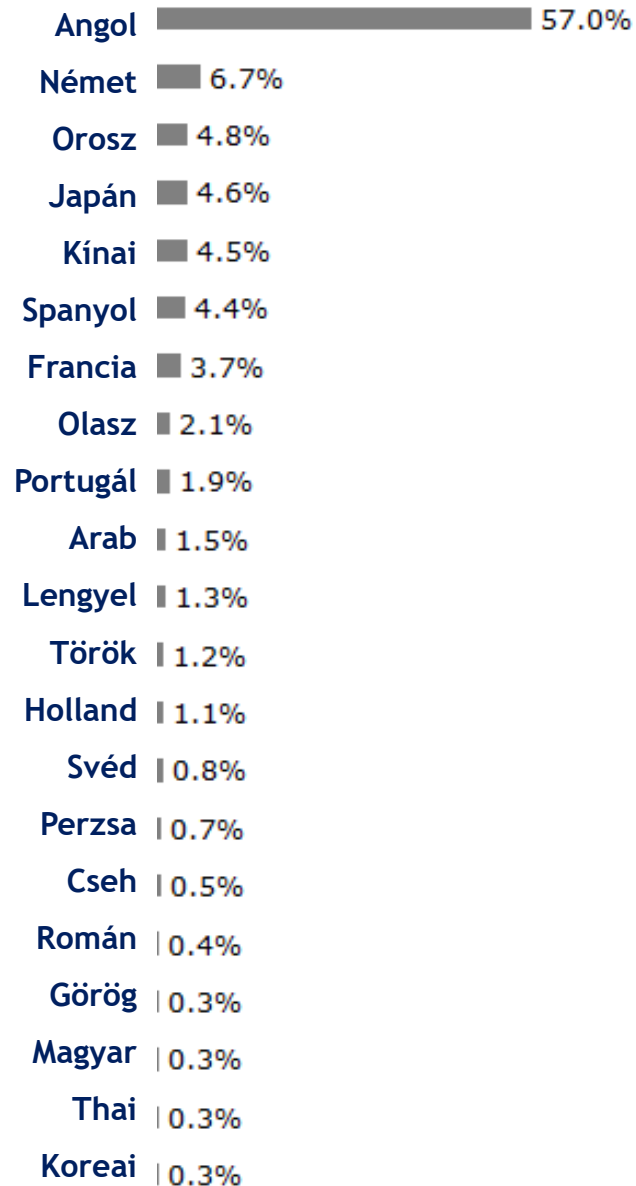


A web és a nyelvek

Nyelv	Internet- használók	Internet- penetráció	Növekedés (2000-)	Internet- használók a világ %-ában	Népesség
Angol	565 004 126	43,4 %	301,4 %	26,8 %	1 302 275 670
Kínai	509 965 013	37,2 %	1478,7 %	24,2 %	1 372 226 042
Spanyol	164 968 742	39,0 %	807,4 %	7,8 %	423 085 806
Japán	99 182 000	78,4 %	110,7 %	4,7 %	126 475 664
Portugál	82 586 600	32,5 %	990,1 %	3,9 %	253 947 594
Német	75 422 674	79,5 %	174,1 %	3,6 %	94 842 656
Arab	65 365 400	18,8 %	2501,2 %	3,3 %	347 002 991
Francia	59 779 525	17,2 %	398,2 %	3,0 %	347 932 305
Orosz	59 700 000	42,8 %	1825,8 %	3,0 %	139 390 205
Koreai	39 440 000	55,2 %	107,1 %	2,0 %	71 393 343
Σ	1 615 957 333	36,4 %	421,2 %	82,2 %	4 442 056 069
A többi	350 557 483	14,6 %	588,5 %	17,8 %	2 403 553 891
Összesen	2 099 926 965	30,3 %	481,7 %	100,0 %	6 930 055 154



A weblapok nyelvei



Következmények

1. Az internetes tartalom több mint a fele angol
2. Az internethasználók háromnegyedének az anyanyelve nem angol
3. $1+2 \rightarrow$ A nem-angol anyanyelvűek világában a web angol tartalmának egyre nagyobb része jelenik meg
4. $3 \rightarrow$ Aki a webet olvassa, az növekvő számban nem-angol anyanyelvű; aki pedig a webes tartalmat közzéteszi, az csökkenő részben angol anyanyelvű
5. $4 \rightarrow$ A nem-anyanyelvűek nagy része nem ismeri fel a nyelvi hibákat és pontatlanságokat

Hosszú távú következmény:

mi lesz a statisztikai alapú nyelvi programokkal?