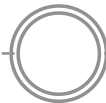




A nyelvtechnológia alapjai



1. Gyakorlat

Karakterkódolás, Unicode



- több mint 7000 nyelv
- sokféle írás:
 - képírás (piktográfia)
 - fogalomírás (ideografikus)
 - szó- és szótagírás (logografikus és fonetikus írás)
 - betűírás (alfabetikus írás)
- az írás iránya különböző lehet
- azonos hangokat különféleképpen írunk
- azonos írásjeleket különféleképpen ejtünk ki

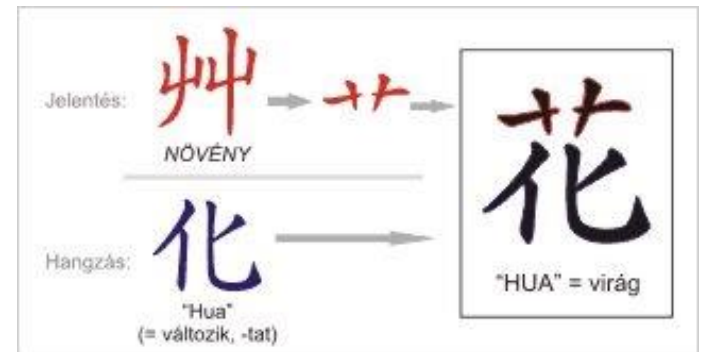
A világ írásai

- Képirás (piktográfia)
 - Barlangrajz
- Fogalomírás



A kínai szójelek fejlődése

hegy					山
Nap					日
Hold					月
eső					雨
felhő					雲



- Szótagírás

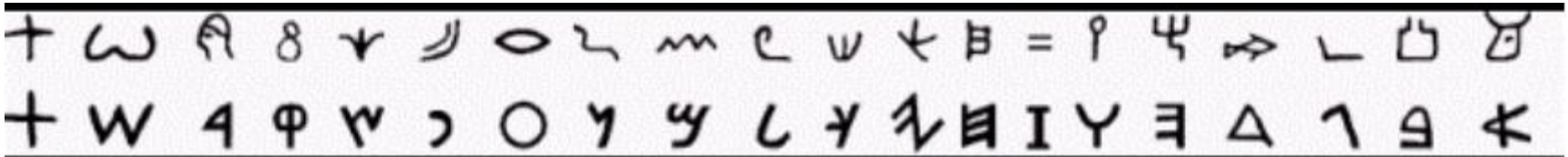
-



- követ = kő + magot szóró alak
 - silány = sí + lány

- Betűírás

- a protosínai hieroglifák és a belőlük származó föníciai ábécé (akrofónia); → görög, etruszk → latin:



A világ írásai - az 'A' betű

ΑΑΑх|ԱႱ ᵃ

1 2 3 4 5 6 7 8

অঅਅਅඒஅఅఅഅഃ ൃ

9 10 11 12 13 14 15 16 17 18 19

あアㇰ 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

- görög Ἐν ἀρχῇ ἐποίησεν ὁ θεὸς τὸν οὐρανὸν καὶ τὴν γῆν.
ἡ δὲ γῆ ἦν ἀόρατος καὶ ἀκατασκεύαστος, καὶ σκότος ἐπάνω τῆς ἀβύσσου, καὶ πνεῦμα θεοῦ ἐπεφέρετο ἐπάνω τοῦ ὕδατος.
καὶ εἶπεν ὁ θεός Γενηθήτω φῶς.
καὶ ἐγένετο φῶς.
καὶ εἶδεν ὁ θεὸς τὸ φῶς ὅτι καλόν. καὶ διεχώρισεν ὁ θεὸς ἀνὰ μέσον τοῦ φωτὸς καὶ ἀνὰ μέσον τοῦ σκότους.

- örmény

Հայր մեր որ յերկինս էս, սարք
էղիցի անան քո, էկեցե ար-
քայութիլն քո. էղիցին կամք քո
որսլես յերկինս և յերկրի. Չի-
աց մեր Հանապազորդ տուր-
մեզ այսալր. և թող մեզ զպա-
րտիա մեր որպես և մեք թո-
ղովք և էրոց պարտապանաց.
և մի տանիր զմեզիփորձով.
Թիլմ, այլ փրկես զմեզ ի չարե.
ամեն.

- grúz

მამაო ჩუჴნო, რომელი ხარ ცათა
შინა, წმიდა იყავნ სახელი შენი.
მოვედინ სუფევა შენი, იყავნ ნება
შენი ვითარცა ცათა სინა, ეგრეცა
ქუჴჟანასა ზედა. პური ჩუჴნი
არსობისა მომეც ჩუჴნ დღეს, და
მომიტევენ ჩუჴნ თანა-ნადებნი
ჩუჴნნი ვითარცა ჩუჴნ მიუტევებთ
თანა-მდებთა მათ ჩუჴნთა, და ნუ
შემიჟანებ ჩუჴნ განსაცდელსა,

- héber בְּרֵאשִׁית בָּרָא אֱלֹהִים אֶת הַשָּׁמַיִם
וְאֶת הָאָרֶץ:
וְהָאָרֶץ הָיְתָה תֵּהוֹ וּבְהוּ
וַחֲשָׁךְ עַל־פְּנֵי תְהוֹם וְרוּחַ
אֱלֹהִים מְרַחֶפֶת עַל־פְּנֵי הַמַּיִם:
וַיֹּאמֶר אֱלֹהִים יְהִי אוֹר וַיְהִי־אוֹר:
וַיֵּרָא אֱלֹהִים אֶת־הָאוֹר כִּי־טוֹב
וַיַּבְדֵּל אֱלֹהִים בֵּין הָאוֹר
וּבֵין הַחֹשֶׁךְ:

- arab

ايونا الذي في السموات لتقدس
اسمك لتات ملكوتك لتكن مشيتك
كما في الماء، وعلى الارض ضبنا
كفافنا اعطنا اليوم واغف لنا
خطايانا كما نغف خن لمن اخطا،
اينا ولا تدخلنا في التحباين لكن
نحبنا من الشرى امين



A világ írásai

- kínai

聖母經

申爾福。瑪利亞。滿被聖寵者。
主與爾偕焉。女中爾爲讚美。
爾胎子耶穌。併爲讚美。天
主聖母瑪利亞。爲我等罪人。
今祈天主。及我等死候。亞孟。

- a latin, és ami történt vele:

s	magyar, vietnami
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
ê	eszperantó



- Kelet

áâäāąéëēėęěíîïıóôõö
ōóúûüūůűųý
çćĉčċďđġğĝĥħĵķĺļł
ńņñŕŗřśŝșšţţž

- a nyugat-európai nyelvek ékezetes betűi:

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 ð þ ß

- ASCII kódtábla
- (7 bit = 127 karakter)

	0	1	2	3	4	5	6	7
0	NUL	DLE	SP	0	@	P	`	p
1	SOH	DC1	!	1	A	Q	a	q
2	STX	DC2	"	2	B	R	b	r
3	ETX	DC3	#	3	C	S	c	s
4	EOT	DC4	\$	4	D	T	d	t
5	ENQ	NAK	%	5	E	U	e	u
6	ACK	SYN	&	6	F	V	f	v
7	BEL	ETB	'	7	G	W	g	w
8	BS	CAN	(	8	H	X	h	x
9	HT	EM	)	9	I	Y	i	y
A	LF	SUB	*	:	J	Z	j	z
B	VT	ESC	+	;	K	[	k	{
C	FF	FS	,	<	L	\	l	
D	CR	GS	-	=	M	]	m	}
E	SO	RS	.	>	N	^	n	~
F	SI	US	/	?	O	_	o	DEL

A magyar betűk kódolásai ASCII karakterekkel

- ekezet nélkül:
 - arvizturo tukorfurogep (fokabel???)
- távirati:
 - aarviiztuerooe tuekoerfuuroogeep
- repülő ékezetek:
 - a'rviztu''ro'' tu:ko:rfu'ro'ge'p
- Prószéky-kód:
 - a1rvi1ztu3ro3 tu2ko2rfu1ro1ge1p
- (La)TeX:
 - \ 'arv\ 'izt\ h{u}r\ h{o} t\ ''uk\ ''orf\ 'ur\ 'og\ 'ep
- SGML (HTML):
 - ´rvíztűrő

- ISO8859-2 (Latin-2)

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

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	DLE	SP	0	@	P	`	p			NBSP	°	Ř	Đ	í	đ
1	SOH	DC1	!	1	A	Q	a	q			À	à	Á	Á	á	ñ
2	STX	DC2	"	2	B	R	b	r			˘	˘	Â	Ñ	â	ñ
3	ETX	DC3	#	3	C	S	c	s			Ł	ł	Ă	Ó	ă	ó
4	EOT	DC4	\$	4	D	T	d	t			¸	¸	Ä	Ô	ä	ô
5	ENQ	NAK	%	5	E	U	e	u			Ł	ł	Í	Õ	í	õ
6	ACK	SYN	&	6	F	V	f	v			Š	š	Ć	Ö	ć	ö
7	BEL	ETB	'	7	G	W	g	w			Š	˘	Ç	×	ç	÷
8	BS	CAN	(	8	H	X	h	x			˘	˘	Č	Ř	č	ř
9	HT	EM	)	9	I	Y	i	y			Š	š	É	Ů	é	ů
A	LF	SUB	*	:	J	Z	j	z			Š	š	Ę	Ú	ę	ú
B	VT	ESC	+	;	K	[	k	{			Ť	ť	Ě	Ů	ě	ů
C	FF	FS	,	<	L	\	l				Ž	ž	Ě	Ü	ě	ü
D	CR	GS	-	=	M	]	m	}			-	"	Í	Ý	í	ý
E	SO	RS	.	>	N	^	n	~			Ž	ž	Î	Ť	î	ť
F	SI	US	/	?	O	_	o	DEL			Ž	ž	Ď	ß	d'	'

• ISO8859-5

- bolgár
- belorusz
- macedón
- orosz
- szerb
- ukrán

		0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	DLE	SP	0	@	P	`	p				NBSP	А	Р	а	р	№
1	SOH	DC1	!	1	А	Q	a	q				Ё	Б	С	б	с	ё
2	STX	DC2	"	2	В	Р	b	r				Ъ	В	Т	в	т	ђ
3	ETX	DC3	#	3	С	Ѕ	c	ѕ				Ѓ	Г	У	г	у	ѓ
4	EOT	DC4	\$	4	Д	Т	d	t				Є	Д	Ф	д	ф	є
5	ENQ	NAK	%	5	Е	U	e	u				Ѕ	Е	Х	e	х	ѕ
6	ACK	SYN	&	6	Ф	В	f	v				І	Ж	Ц	ж	ц	і
7	BEL	ETB	'	7	Г	W	g	w				Ї	З	Ч	з	ч	ї
8	BS	CAN	(	8	Н	Х	h	x				Ј	И	Ш	и	ш	ј
9	HT	EM	)	9	І	Y	i	y				Љ	Й	Щ	й	щ	љ
A	LF	SUB	*	:	Ј	З	j	z				Њ	К	Ъ	к	ъ	њ
B	VT	ESC	+	;	К	[	k	{				Ћ	Л	Ы	л	ы	ћ
C	FF	FS	,	<	Л	\	l					Ќ	М	Ь	м	ь	ќ
D	CR	GS	-	=	М	]	m	}				-	Н	Э	н	э	§
E	SO	RS	.	>	Н	^	n	~				Ў	О	Ю	о	ю	ў
F	SI	US	/	?	О	_	o	DEL				Џ	П	Я	п	я	џ



- 1980-as évek: igény egy egységes, minden karaktert leíró kódtáblára
 - ISO 10646 szabvány:
 - Universal Character Set (UCS)
 - Unicode – amerikai cégek konzorciuma
- 1991 – egyesítik a két projekt eredményeit.
- minden beszélt, holt és egyéb nyelv (pl. Braille) karakterei (betűk, számok, írásjelek, stb.)
- minden egyes karakternek egyedi száma és neve van

A Unicode és a UCS

- eredetileg 16 biten (2 byte-on) tárolta az adatokat, de ez nem volt elég, ezért 4 byte-osra bővítették
 - $2^{16} = 65\,536$ Basic Multilingual Plane (BMP)
 - $2^{32} = 4\,294\,967\,296$
- Az első 255 karakter megegyezik a Latin-1 (ISO8859-1) karakterekkel
- nemcsak szám, hanem név is:
 - U+0041 Latin capital letter A
 - U+0151 Latin small letter o with double acute
- Unicode Code search
 - <http://www.fileformat.info/info/unicode/char/search.htm>
- UCS-2: 2 bájtton
- UCS-4: 4 bájtton

- UTF – Unicode Transformation Format
 - UTF-8:

U+0000 - U+007F	0xxxxxxx
U+0080 - U+07FF	110xxxxx 10xxxxxx
U+0800 - U+FFFF	1110xxxx 10xxxxxx 10xxxxxx
U+10000 - U+1FFFFF	11110xxx 10xxxxxx 10xxxxxx 10xxxxxx
U+200000 - U+3FFFFFFF	111110xx 10xxxxxx 10xxxxxx 10xxxxxx 10xxxxxx
U+4000000 - U+7FFFFFFF	1111110x 10xxxxxx 10xxxxxx 10xxxxxx 10xxxxxx 10xxxxxx

- sokféle nyelvi-területi beállítási lehetőség
- nyelvi locale:
 - kis/nagybetű konverzió
 - ABC rendezés
 - kódlap
 - dátum formátum
 - pénznem

English vs. Turkish Case Mappings

Language	Letter	Lowercase Map	Uppercase Map
English	i	ı	I
Turkish	dotted i	İ	İ
Turkish	dotless ı	ı	I

- cong, Alma, égbolt, chang, cukor, egér, pácsó, Álom, éger, csoki, alma, Eger

- Láng Attila : Íráskalauz
<http://www.mek.iif.hu/porta/szint/tarsad/konyvtar/tortenet/iras/html/>
- Naszódi Mátyás (BME-SZIT)
- Unicode:
<http://www.unicode.org>
- Kálmán-Trón: Bevezetés a nyelvtudományba, Tinta, 2005



Köszönöm a figyelmet



ध्यान देने के लिए आपका धन्यवाद
आप अपना ध्यान के लिए धन्यवाद
थैंक यू फॉर योर अटेंशन

Thank you for your attention
Thank you for your attention
Thank you for the attention

Merci de votre attention
Je vous remercie pour votre attention
Merci pour l'attention

感谢您的关注
谢谢你的关注
谢谢您的注意