Zapis znakov in uporaba korpusov

Tomaž Erjavec
Korpusno jezikoslovje
UNG
2008/2009

9. 1. 2009

Pregled predavanja

1. Zapis znakov v računalnikih
2. Primeri študij uporabe korpusov
3. Diskusija seminarskih nalog

Kodiranje znakov

- Digitalni računalniki shranjujejo podatke kot (binarne) številke
- Ne obstaja vnaprej dana povezava med temi številkami in znaki (abecede)
- Če ni konvencij za preslikavo ali jih je preveč → kaos
- Standardi in pol-standardi:
  ASCII, ISO 8859, (Windows, Mac), Unicode

Osnovni pojmi I.

znak *(character)*

- abstrakten pojem (An „A“ is something like a Platonic entity: it is the idea of an „A“ and not the „A“ itself)
- sam po sebi znak nima preslikave v številko ali določene grafične podobe
- ponavadi je opisno definiran, npr. „grška črka mala alfa“, grafična podoba pa podana samo kot vodilo, „α“

Osnovni pojmi II.

- ***nabor znakov*** *(character set)*
  - množica znakov
  - vsakemu znaku pripisana njegova koda
- ***koda znaka*** *(character code)*
  - 1-1 relacija med znakom iz nekega nabora znakov in številko, npr. A = 26, B = 27, ...
  - Pozor!
    Kode znakov se dostikrat zapisujejo šestnajstiškem sistemu:
    0 → 0, 1 → 1, 2 → 2, ... 9 → 9, 10 → A, 11 → B, ...,
    15 → F, 16 → 10, 17 → 11, ...,
    254 → FE, 255 → FF, 266 → 100

Primer: nabor znakov ASCII

Below is the standard ASCII characters.

Dec	Char	Dec	Char	Dec	Char	Dec	Char	Dec	Char	Dec	Char
33	!	49	1	65	A	81	Q	97	a	113	q
34	"	50	2	66	B	82	R	98	b	114	r
35	#	51	3	67	C	83	S	99	c	115	s
36	$	52	4	68	D	84	T	100	d	116	t
37	%	53	5	69	E	85	U	101	e	117	u
38	&	54	6	70	F	86	V	102	f	118	v
39	'	55	7	71	G	87	W	103	g	119	w
40	(	56	8	72	H	88	X	104	h	120	x
41	)	57	9	73	I	89	Y	105	i	121	y
42	*	58	:	74	J	90	Z	106	j	122	z
43	+	59	;	75	K	91	[	107	k	123	{
44	,	60	<	76	L	92	\	108	l	124	|
45	-	61	=	77	M	93	]	109	m	125	}
46	.	62	>	78	N	94	^	110	n	126	~
47	/	63	?	79	O	95	_	111	o	127	_
48	0	64	@	80	P	96	`	112	p		

npr.

v naboru znakov ASCII

ima znak mali latinični a

kodo znaka 97

Osnovni pojmi III.

- ***pismenka*** *(glyph)*
  - grafična predstavitev znaka
  - enemu znaku lahko ustreza več kot ena pismenka
    npr. znak "veliki latinični A" ↔ pismenke 𝐀, 𝒜, 𝔸
  - redko tudi eni pismenki ustreza več znakov
    npr. pismenka P ↔ znaki "veliki latinični P", "veliki cirilični R", "veliki grški Ro")
- ***font***
  - nabor pismenk (za nek nabor znakov):
    𝐀, 𝐁, 𝐂, Č, 𝐃, ...
  - včasih font ne pokriva celotnega nabora znakov!

Nekateri nabori znakov

- ASCII - najstarejši, vsebuje samo črke ameriške abecede + ločila, številke
- Družina naborov znakov ISO 8879
- Družina naborov Windows
- Unicode

ASCII

- American Standard Code for Information Interchange (1950')
- 7-bitni zapis znakov: kode znakov 0-127
- 0-31 - kontrolni znaki + znaki za formatiranje: Esc, Line Feed, tabulator, presledek,...
- 32-126 – ločila in posebni znaki, številke, velike in male angleške črke :
  ! " # $ % & ' ( ) * + , - . / 0 1 2 3 4 5 6 7 8 9 : ;
  < = > ? @ 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 [\ ] ^ _ ` 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 { | } ~

Družina naborov znakov ISO 8859

- potreba po dodatnih znakih za nacionalne (evropske) pisave:
  - v 80's se pojavljajo novi nabori znakov
  - obsegajo ASCII kot podmnožico
- International Standard Organisation izda kodne nabore za posamezne skupine (evropskih) jezikov: družina standardov **ISO 8859**
- ISO 8859-1 (ISO Latin 1) – zahodnoevropski jeziki

¡ ¢ £ ¤ ¥ ¦ § ¨ © ª « ¬ - ® ¯ ° ± ² ³ ´ µ ¶ · ¸ ¹ º » ¼ ½ ¾ ¿ À Á Â Ã Ä Å Æ Ç È É Ê Ë Ì Í Î Ï Ð Ñ Ò Ó Ô Õ Ö × Ø Ù Ú Û Ü Ý Þ ß à á â ã ä å æ ç è é ê ë ì í î ï ð ñ ò ó ô õ ö ÷ ø ù ú û ü ý þ ÿ

Nabori znakov za ne-zahodno evropske jezike

- za slovenščino in ostale srednje- in vzhodno evropske (latinične) jezike - anarhija:
  - ISO 8859-2 (ISO Latin 2)
  - Windows CP1250 (grrr!)
  - lastni „standardi“: IBM, Apple, …

ISO 8859-2 (zgornja polovica)

	–0	–1	–2	–3	–4	–5	–6	–7	–8	–9	–A	–B	–C	–D	–E	–F
A–	NBSP 00A0 160	Ą 0104 161	˘ 02D8 162	Ł 0141 163	¤ 00A4 164	Ľ 013D 165	Ś 015A 166	§ 00A7 167	¨ 00A8 168	Š 0160 169	Ş 015E 170	Ť 0164 171	Ź 0179 172	SHY 00AD 173	Ž 017D 174	Ż 017B 175
B–	° 00B0 176	ą 0105 177	˛ 02DB 178	ł 0142 179	´ 00B4 180	ľ 013E 181	ś 015B 182	ˇ 02C7 183	¸ 00B8 184	š 0161 185	ş 015F 186	ť 0165 187	ź 017A 188	˝ 02DD 189	ž 017E 190	ż 017C 191
C–	Ŕ 0154 192	Á 00C1 193	Â 00C2 194	Ă 0102 195	Ä 00C4 196	Ĺ 0139 197	Ć 0106 198	Ç 00C7 199	Č 010C 200	É 00C9 201	Ę 0118 202	Ë 00CB 203	Ě 011A 204	Í 00CD 205	Î 00CE 206	Ď 010E 207
D–	Đ 0110 208	Ń 0143 209	Ň 0147 210	Ó 00D3 211	Ô 00D4 212	Ő 0150 213	Ö 00D6 214	× 00D7 215	Ř 0158 216	Ů 016E 217	Ú 00DA 218	Ű 0170 219	Ü 00DC 220	Ý 00DD 221	Ţ 0162 222	ß 00DF 223
E–	ŕ 0155 224	á 00E1 225	â 00E2 226	ă 0103 227	ä 00E4 228	ĺ 013A 229	ć 0107 230	ç 00E7 231	č 010D 232	é 00E9 233	ę 0119 234	ë 00EB 235	ě 011B 236	í 00ED 237	î 00EE 238	ď 010F 239
F–	đ 0111 240	ń 0144 241	ň 0148 242	ó 00F3 243	ô 00F4 244	ő 0151 245	ö 00F6 246	÷ 00F7 247	ř 0159 248	ů 016F 249	ú 00FA 250	ű 0171 251	ü 00FC 252	ý 00FD 253	ţ 0163 254	˙ 02D9 255

8-bitni nabori (ISO 8859, Windows)

- prednosti:
  - lahko zapišemo znake posameznih nacionalnih abeced (slovenščina)
- slabosti:
  - v istem kodnem naboru ne moremo zapisati večjezičnih besedil
  - zmeda zaradi večih kodnih naborov, ki pokrivajo iste jezike
  - ni pokritja npr. za vzhodno-azijske jezike ali bolj zahtevne znake: ločila, matematični simboli, naglasna znamenja, …
  - datoteka ne vsebuje podatka v katerem kodnem naboru je vsebina:
    © Global publishing ~ Ž Global publishing

Unikod I.

Unikod (Unicode oz. ISO 10646)

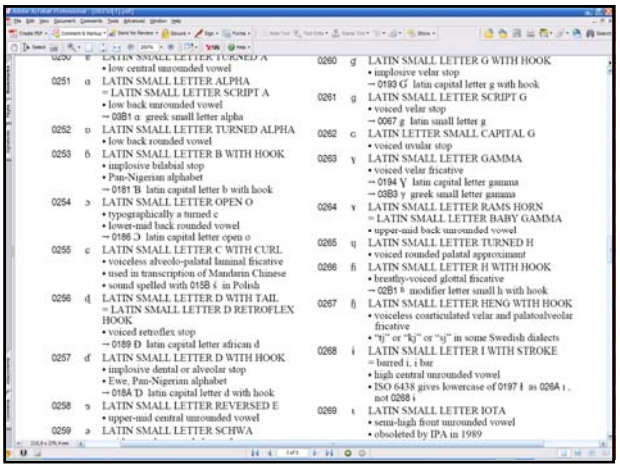
1991 – Unicode Consortium: http://www.unicode.org/

- definira univerzalni nabor znakov
- vsebuje 30 svetovnih abeced, ki pokrivajo več sto jezikov, definiranih približno 40.000 znakov
- …arabščina, sanskrt, kitajščina, japonščina, korejščina,…
- tudi zgodovinske pisave, ločila, matematični simboli, naglasna znamenja,..
- Unikod razdeli znake v „bloke“
  npr. Basic Latin, Latin-1 Supplement, Latin Extended-A, Latin Extended-B, IPA Extensions, Combining Diacritical Marks, Greek, Cyrillic, …

Slovenske črke

Latin Extended-A

Position	Decimal	Name	Appearance
0x0100	256	LATIN CAPITAL LETTER A WITH MACRON	Ā
0x0101	257	LATIN SMALL LETTER A WITH MACRON	ā
0x0102	258	LATIN CAPITAL LETTER A WITH BREVE	Ă
0x0103	259	LATIN SMALL LETTER A WITH BREVE	ă
0x0104	260	LATIN CAPITAL LETTER A WITH OGONEK	Ą
0x0105	261	LATIN SMALL LETTER A WITH OGONEK	ą
0x0106	262	LATIN CAPITAL LETTER C WITH ACUTE	Ć
0x0107	263	LATIN SMALL LETTER C WITH ACUTE	ć
0x0108	264	LATIN CAPITAL LETTER C WITH CIRCUMFLEX	Ĉ
0x0109	265	LATIN SMALL LETTER C WITH CIRCUMFLEX	ĉ
0x010A	266	LATIN CAPITAL LETTER C WITH DOT ABOVE	Ċ
0x010B	267	LATIN SMALL LETTER C WITH DOT ABOVE	ċ
0x010C	268	LATIN CAPITAL LETTER C WITH CARON	Č
0x010D	269	LATIN SMALL LETTER C WITH CARON	č



Unikod II.

- ISO 8859, Windows-CP,... so uporabljali samo 256 kod (znakov) in torej potrebovali samo en zlog (byte) za zapis posameznega znaka
- V Unikodu je kod znakov bistveno več
  1 znak ≠ 1 zlog, kaj sedaj?
- za Unikod obstaja več zapisov znakov:
  - UTF-32
    - 1 znak – 4 zlogi
  - UTF-16
    - 1 znak – 2 zloga (za „osnovne" znake)
  - **UTF-8**
    - spremenljiva velikost: 1-6 zlogov na znak
    - če je znak v ASCII, potem v 1 zlogu (kompatibilnost)
    - znaki slovenske abecede se kodirajo v dveh zlogih

Unikod in naglasna znamenja

- naglasna znamenja obstajajo kot znaki brez širine (combining diacritical marks)
- npr.: a + ˆ + ¨ = ậ
- vendar problemi pri prikazu kompleksnih kombinacij, npr.
  a + ˆ + ˚ = â
  - specializirani fonti to rešujejo bolje

Nazaj v ASCII

ASCII je včasih še vedno edini varen zapis:

- problemi pri vnosu ali izpisu znakov
- problemi pri prenosu podatkov (elektronska pošta)

Prekodiranje v ASCII:

- elektronska pošta - standard MIME
- SGML (HTML) in XML - entitete za znake, s kodnimi mesti iz Unikoda
  npr. š = Š = š

Določanje nabora znakov

- **HTML:**

```
<HTML>
  <HEAD>
    <TITLE>Recept za ribano kašo</TITLE>
    <META http-equiv="Content-Type"
          content="text/html; charset=ISO-8859-2">
  </HEAD>
  <BODY>
    …
```

- **XML:**

```
<?xml version="1.0" encoding="utf-8"?>
<recept>
  <naslov>Recept za ribano kašo</naslov>
  …
```

- Nekateri dovoljeni nabori znakov:
  - utf-8, iso-8859-2, us-ascii

Vaje iz zapisa znakov: Word

Pri tej vaji uporabimo Word, da spoznamo razlike med kodnimi nabori.
Besedilo:

Mačka, miška in žolna so šli na izlet v Črnomelj, nato prav počasi v Šujico, na koncu so pa pristali v Žužemberku, kjer so srečali čmrlja. »Kako dolgo smo hodili!« je mu je potožila miška. Mačka pa mu je rekla "Čuj čmrlj, koliko samoglasnikov je pravzaprav v tvojem imenu?"

- Odpri novo datoteko v Wordu, in gornje besedilo prilepi vanjo. Shrani jo kot besedilo (.txt), v kodnem naboru ISO-8859-2 (»Central European ISO«). Za katere znake javi Word, da bodo nepravilno shranjeni? Zakaj?
- Datoteko zapremo, nato .txt ponovno odpri v urejevalniku Word; katere kodne nabore (od teh, ki smo jih omenili v predavanju) ponudi Word? Kaj se zgodi, če besedilo odpremo v privzeti kodni tabeli za Windows? Datoteko zapremo, ne da bi jo shranili.
- Datoteko odpremo in ponovno shranimo kot .txt, tokrat v kodnem naboru UTF-8 (»UTF«) in jo zapremo.
- Besedilo ponovno odpremo. Kako izgleda besedilo, če ga odpremo v kodnem naboru ISO-8859-2? Zapremo, ne da bi shranili.
- Besedilo pravilno odpremo (torej v UTF-8), nato pa shranimo v UTF-16 (»Unicode«), datoteko spet zapremo, in ponovno odpremo. Kaj se zgodi, če jo odpremo s katerim od 8-bitnih kodnih naborov? Zakaj?

Vaje iz zapisa znakov: Unikod

Učimo se poiskati želeni znak v urejevalniku Word (vstavi simbol) in na spletnih straneh Unikoda, na http://www.unicode.org/charts/ .

Fonetično želimo napisati »čmrlj«, ta ko da je med »mr« in »rl« znak za polglasnik, namesto »č« pa znak iz mednarodne fonetične abecede IPA. Ker je prvi polglasnik naglašen, mu dodajte še ostrivec, torej:

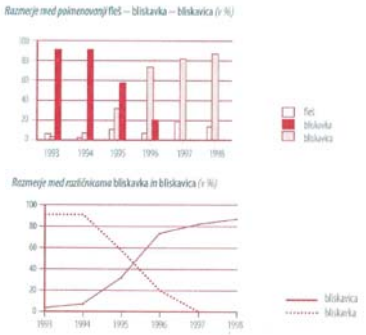
ʧmə́rəlj

1. Najdi potrebne znake med »Vstavi znak« v urejevalniku Word. Pozor: v okencu levo zgoraj izberi font, ki podpira te fonetične znake, verjetno Arial Unicode MS. Kako so ponujeni znaki znaki urejeni? Kako se znaka imenujeta?
2. Znaka nato poišči še preko spletne strani Unicode. Kateri sta njuni kodni mesti?
3. V katerem kodnem naboru lahko shranimo to datoteko?

Študije iz korpusnega jezikoslovja

- Knjiga
  Vojko Gorjanc: *Uvod v korpusno jezikoslovje*.
  Domžale: Izolit, 2005.
  163 str.
- prva knjiga o korpusnem jezikoslovju pri nas
- Predstavitev korpusov in korpusnega jezikoslovja
- Četrti del:
  Korpusni opisi slovenskega jezika
  (uporaba korpusa Fida)
- *Dva primera:*
  - *spreminjanje jezika - terminologija*
  - *zajem semantičnih povezav med besedam*

Spreminjanje jezika



Semantične relacije

- enak, nasproten pomen:
  - sinonimi (neodvisnost ↔ samostojnost)
  - anotonimi (lahek ↔ težek)
- nad, podpomen:
  - hipernimi (ptič → vrabec)
  - hiponimi (vrabec → ptič)
- del in celota:
  - holonimi (avto → vrata)
  - meronimi (vrata→ avto)
- pomembno za izdelavo tezavrov in aplikacije v jezikovnih tehnologijah

Uporaba vzorcev za odkrivanje semantičnih relacij

- Iskalni pogoj je niz, ki v besedilu vzpostavlja semantično relacijo, ki nas zanima

Pomenski označevalci glede na uspešnost zajetja pod-/nadpomenskosti

	abs. pog./pog. označ. podpomenskosti		uspešnost	
označevalec	A	B	A	B
je vrsta	90/19	13/4	0,21	0,31
je * vrsta	156/25	20/0	0,15	0
prištevamo med	12/9	5/5	0,75	1
sodi med	182/45	79/39	0,24	0,49
spada med	168/34	65/30	0,20	0,46
sodi v družino	13/4	0/0	0,30	0
uvrščamo med	60/22	7/3	0,36	0,43

Sopomenke: *imenovan tudi*

Del konkordančnega niza iskalnega pogoja imenovan tudi/imenujemo tudi

opisan nepoveden način odkriti dvižkov oksid,	**imenovan tudi**	smejalni plin, zaradi katerega postane človek omotičen
Vitamin B1,	**imenovan tudi**	tiamin, je verjetno najbolj znan med šestimi vitamini
Vitamin B2,	**imenovan tudi**	riboflavin, je prazaprav deležen najmanj pozornosti
Stopnjo dovzetna do kode	**imenujemo tudi**	dozneg procedure.
rumenkastorjave marage. Ta samotarski kuščar,	**imenovan tudi**	Železglavi leguan, je v peteklivosti
pisanost morskega živ|enja. To skupino zvetrnic	**imenujemo tudi**	ehinoderme.
Je kdaj slišal(-a), da Zemljo	**imenujemo tudi**	modri planet?
Zato spletne strani	**imenujemo tudi**	HTML dokument. V osnovi je HTML dokument
Večplastno osebnost	**imenujemo tudi**	razcepljena zvestnost: to je izraz, s katerim
karte meril 1 : 10 000 in 1 : 5 000	**imenujemo tudi**	detajlne geološke karte, karte v še večjih merilih
Oblikovanje hitrih elektronov	**imenujemo tudi**	sevanje žarkov ß, ves pojav pa
Snovi v tretjem agregatnem stanju	**imenujemo tudi**	trdnine. Tudi pr ijih nas zanima, kako se
tla bogata predvsem z železovimi spojinami, ph	**imenujemo tudi**	železnata tla, ferisol.

“znan kot”

Del konkordančnega niza iskalnega pogoja znan kot

[illegible] z visokoenergijskimi rentgenskimi žarki, je proces	**znan kot**	rentgenoluminiscenca
svetlobo odda [illegible] galaksij. Zaradi pojava, ki je danes	**znan kot**	rdeči premik, je jasno iz spremembe barvne
kot jih je imel slavni člen lambda (tudi	**znan kot**	kozmološka konstanta), ki ga je
eksplozije pred približno 17 milijardami let. Ta dogodek je	**znan kot**	"Veliki bum" (tudi "Veliki pok")
sestavljen iz zvezd, ki tvorijo asterizem,	**znan kot**	Jobova krsta. Dve od teh zvezd imata čudni
v observatoriju Maunt Palomar, ki je pozneje postal	**znan kot**	Halov teleskop. Za postavitev tega teleskopa
arabski astronom je bil Al Batani, v Evropi	**znan kot**	Albategnius. Deloval je v zadnji četrtini
je delo Hertzsprunga prikazal grafično. Graf, sedaj	**znan kot**	Hertzsprung-Russllov diagram, prikazuje zvezo
antibiotike, kot npr. fatalni tip streptokoka A,	**znan kot**	smrtonosna bakterija, za katero je značilno, da je
nad njim in tudi povečane. Pojav je	**znan kot**	fata morgana in je pogosto opazovan

Samostalnik (Samostalnik)

Del prečiščenega konkordančnega niza iskalnega pogoja Sam (Sam) v podkorpusu naravoslovne vede (Cobiss)

Enocelični plazmodiji razgrajajo rdeča krvna	**telesca (eritrocite)**	in ob tem povzročajo silne napade
proces strjevanja tedaj, ko vmešamo v smolo sredstvo za	**strjevanje (trdilec).**	
vrste lesa, papir, kovine, steklo, acetatna	**celuloza (celuloid),**	akrilno steklo (PMMA) in nekatere
v vodik in kisik. Vodik se nabira na negativni	**elektrodi (katodi),**	kisik pa na pozitivni
negativni elektrodi (katodi), kisik pa na pozitivni	**elektrodi (anodi).**	Pri gorilni celici je postopek
lastnosti dimnih zaves temeljijo na optičnih pojavih	**disperzije (razprševanja)**	in absorpcije (vsrkanja) svetlobe
zvišamo, olje začne zgorevati in pri tem nastane oglen	**prah (čad),**	ki lebdi v zraku kot dimni oblak črne barve
ptiče bogov in kraljev, ki se v času	**ženitve (spomladi)**	v resnici prelevijo v pravljična bitja.
osrednjega dela ali telesa nevrona, več krajših, vejastih	**izrastkov (dendritov)**	in le enega dolgega izrastka (aksona).
dneh na zemeljski ekvator (polutnik) ter na oba	**pola (tečaja),**	severnega in južnega. Če naprej
je pri svojih operacijah uporabljal karbolno	**kislino (fenol),**	da je preprečil zastrupitve. Kasneje so
Znanstveniki menijo, da ta 35-centimetrska figura prikazuje	**vrača (šamana),**	ki vstopa v drugo stanje.
del motorja, pač pa le kot sestavni del motornega	**bloka (ohišja).**	Zaradi čim lažjega
tega ima sodobna kopija kar 8-krat večji delovni	**pomnilnik (RAM)**	in 4-krat večji trajni pomnilnik (ROM)
pomnilnik (RAM) in 4-krat večji trajni	**pomnilnik (ROM)**	od originala. Sicer pa je vse sestavne dele
rastlinske dednostne molekule dezoksiribonukleinske	**kisline (DNA).**	Poskus je trajal nekaj mesecev
kemijski postopek, kako iz slanice pridobivati natrijev	**hidroksid (lug),**	ki je za izdelavo mila neprimerno boljši
lastnosti sta hitro učinkovanje in visoka stopnja	**strupenosti (toksičnosti);**	so brez barve, vonja in okusa.
postopoma sušijo, potem ko so jih prepojili s polietilen	**glikolom (PEG),**	v vodi topljivo polimerno smolo, katere
sestava je odvisna od matične kamnine, odnašanja	**prsti (erozije)**	in živih bitij, ki sodelujejo pri nastajanju

Meronimi

- katere vzorce bi uporabili za iskanje meronimov - holonimov?

III. Korpusi v leksikografiji: terminološki sovarček

Vaja

1. specializiran korpus s področja primerjamo z referenčnim korpusom, da dobimo ključne besede
2. ključnim besedam pogledamo kolokacije, da dobimo večbesedne termine (iztočnice)
3. iztočnicam preko konkordanc najdemo podpomene (če so) in poiščemo dobre primere uporabe