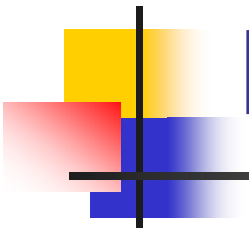
Azbuka, reč, jezik

- Jezik- sredstvo za komunikaciju
- Prihvatljiv za sve učesnike
- Prirodni jezik: ekspresivan, nejednoznačan, neprecizan
- Veštački jezik: u matematici, hemiji, saobraćaju
- Programski jezici: nivoi apstrakcije računara i čoveka



Programski jezik

- Prirodan i jednostavan / precizan: jezički procesori
- ANSI: “Programski jezik je jezik koji se koristi za pripremanje računarskih programa”
- Klasifikacija
- Backus, 1957: FORTRAN

Azbuka, niska

- Sintaksa
- Semantika
- Azbuka: konačni skup znakova (simbola, slova)
- Simbol: nedeljiva jedinica jezika
- Niska: nizanjem simbola
- Niska ω – dužina $|\omega|$: broj simbola
- Prazna niska: ε
- α – slovo, $\alpha^i = \alpha\alpha...\alpha$
- $\alpha^0 = \varepsilon$, $\alpha^1 = \alpha$, $\alpha^2 = \alpha\alpha$, itd.

Azbuka, niska: formalizam

- Azbuka A
- 1. ε je niska nad A
- 2. ω niska nad A i α – slovo iz A -> ωα niska nad A
- 3. niska: pravilima 1, 2.
- Operacije
  - Konkatenacija (dopisivanje)
  - Inverzija (obrtanje)...
- $A = \{\alpha_1\ \alpha_2, \ldots, \alpha_n\}$
- $A^* = \{ \varepsilon , \alpha_1, \alpha_2, \ldots, \alpha_n , \alpha_1\alpha_1, \alpha_1\alpha_2, \ldots, \alpha_1\alpha_n , \alpha_2\alpha_1, \alpha_2\alpha_2, \ldots, \alpha_1\alpha_1\alpha_1, \ldots \}$
- $A^+ = A^* \setminus \varepsilon$

Karakterski skup

- Tekstuelna komunikacija sa računarom
- Spoljašnja azbuka, npr. {A,B,...,Z,1,2,..., 9,?,…}
- Kodiranje: kombinacije 0,1 fiksne dužine – kôd fiksne dužine nad {0,1}
- Svaki računar: svoj skup karaktera
- Spoljašnji oblik i unutrašnja reprezentacija: standardna kodna shema

Kodne sheme

- Kodna reč
  - dužine 7 (8) bita (128 (256) kodnih reči)
  - Dužine 16 bita (65536 kodnih reči)
  - Kodna reč fiksne dužine: *karakter*
  - 1 znak → bajt ✡ 1 bajt → 1 znak (sa ili bez grafičkog lika)
- 1983. g. ISO (International Standard Organization): 7-bitni kôd
- Nacionalna američka verzija: ANSI (American Standards Institute): 1968.g. American Standard Code for Information Interchange: ASCII kôd.

ASCII kôd

- Npr. A – 1000001 (=65)
- B – 1000010 (=66)
- 0 – 0110000 (=48)
- 9 – 0111001 (=57)
- Struktura ASCII kôda:
  - Kodovi 0-31, kôd 127 – kontrolni karakteri – bez grafičkog lika, npr. CR (13), LF (10), itd.
  - Uređenje prema unutrašnjiom kodovima
  - A-Z (65-90) – abecedni poredak
  - a-z (97-122)
  - 0-9 (48-57) rastući brojčani poredak
  - Kôd(velSlovo) = kôd(maloSlovo)-32 (2^5)
- ISO 7-bitni kôd: prostor za druge nacionalne verzije
- Slobodno korišćenje karaktera 64, 91-94, 96, 123-126

YU-ASCII

- Jugoslovenski zavod za standarde 1986.g.
- Narušen abecedni poredak

velikoSlovo	kôd	maloSlovo	kôd
Ž(@)	64	ž(‘)	96
Š([ )	91	š({)	123
Dj(\)	92	đ(|)	124
Ć(])	93	ć(})	125
Č(^)	94	č(~)	126

Proširenja 7-bitnog kôda

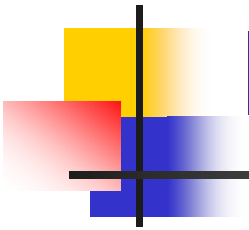
- 7-bitni kôd: svega 95 pozicija - nedovoljno
- 8-bitno proširenje
  - Donja i gornja kodna stranica
  - Donja – ponovljen ASCII: nepostavljen najviši bit
  - Gornja: postavljen najviši bit; bitovi 128-19 (32) nemaju grafički lik
  - 95+96=191 kôd
  - ISO 8859
    - 8859-1 – Latin-1 – većina zapadnoevropskih jezika
    - 8859-2 – Latin2 – većna srednjeevropskih i slovenskih jezika sa latiničnim pismom - i srpska latinica
    - 8859-3 – Latin-3 – esperanto, turski,...
    - 8859-15

Proširenja 7-bitnog kôda

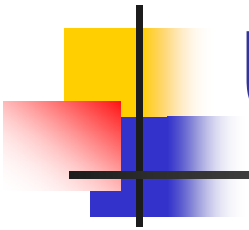
- Naša slova u Latin-2:

velikoSlovo	kôd	maloSlovo	kôd
Ž	174	ž	190
Š	175	š	191
Dj	208	đ	240
Ć	198	ć	230
Č	200	č	232

Industrijski standardi

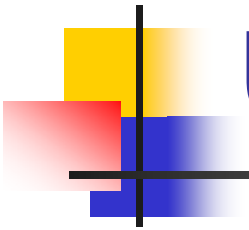


- EBCDIC (Extended Binary Coded Decimal Interchange Code), IBM
- Windows kodne strane: 8-bitna proširenja ASCII koda: svih 128 pozicija gornje kodne strane, ukupno 223 (i €, œ)
  - CP1252 (WinLatin1)
  - CP1250 (WinLatin2)
  - CP1251 (WinCyrillic)...



Unicode

- Xerox Parc, Apple, 1989.g. novi sistem kodiranja - Unicode, cilj:
  - Univerzalan (UNIversal) – sve savremene jezike sa pismom
  - Jedinstven (UNIque), bez dupliranja karaktera – kodiraju se pisma a ne jezici
  - Uniforman (UNIform) - svaki karakter istim brojem bitova: 16
  - 1991: Unicode 1.1
- ISO: standard višebajtovskog kodiranja
  - Universal Multiple-Octet Coded Character Set 4 (ISO 10646)
  - 4-bajtovsko kodiranje (CJK pisma)
  - 1990.g. radna verzija

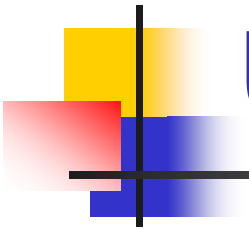


Unicode

- 1993.g. – koordinacija: ISO 10646 kompatibilan sa Unicode-om
- Repertoar Unicode-a: 65536 pozicija
  - Prvih 8192 pozicije: za standardne alfabete
    - Prvih 256: identične ISO 8859-1
  - Sledećih 4096 pozicija specijalni karakteri (0x2000 – 0x3000)
  - Sledećih 4096 pozicija za CJK simbole (0x3000 – 0x4000)
  - CJK ideografsko pismo

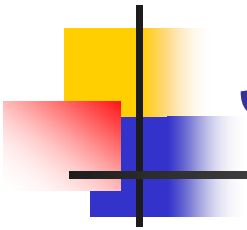
Unicode – naša latinična slova

- 262, 263 – Ć, ć
- 268, 269 – Č, č
- 272, 273 – Đ, đ
- 352, 353 – Š, š
- 381, 382 – Ž, ž



Unicode

- http://www.unicode.org/charts/
- Zadatak: pronaći na Internetu tekstove o Unicode-u i kodne sheme



Jezik

- *Jezik L* nad azbukom *A*: $L \nearrow A^*$
- Beskonačno mnogo jezika nad *A*
- $\omega \in L$ reč jezika *L*
- Pravila koja razlikuju reči od ne-reči; *sintaksa jezika*
- *Semantika* : pravila značenja
- Programski jezik: precizna sintaksa konstante, identifikatora, izraza, iskaza, funkcije, programa, itd.
- Programi – jezički procesori
- Zadavanje sintakse:
  - Nabrajanje
  - Formalna gramatika

Formalna gramatika: primer

- Azbuka: {0,1,2,3,4,5,6,7,8,9,+,-}
- Jezik celih brojeva
- +1205, 1205, -1205
- niske 12+05, 1205+ nisu iz jezika
- Oznake: B – broj; b - neoznačen ceo broj; c – cifra
- Pravila:
  - 1. B je b, +b ili –b;
  - 2. b je c ili b na koji je dopisana c
  - 3. c je simbl iz skupa {0,1,2,3,4,5,6,7,8,9}
- Pomoćni simboli: B, b, c

Formalna gramatika: primer (nast.)

- Zapis pravila:
  - B → b | +b | -b
  - b → c | bc
  - c → 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9
- Početni simbol: B
- Primer izvođenja:
  - B✡b ✡ bc ✡ bcc ✡ bccc ✡ cccc ✡ 1ccc ✡ 12cc ✡ 123c ✡ 1234
  - Moguća i druga izvođenja
  - Izvođenje *n*-tocifrenog broja

Formalna gramatika

- (N,T,P,S)
- N: nezavršni (neterminalni) simboli
- T (A) – završni (terminalni) simboli
- P – pravila
- S□N – početni (startni) simbol
- Primer
  - N = {B,b,c}
  - T = {0,1,2,3,4,5,6,7,8,9,+,-}
  - P = { B → b, B → +b, B → -b, b → c, b → bc, c → 0, c → 1, c → 2, c → 3, c → 4, c → 5, c → 6, c → 7, c → 8, c → 9 }
  - S = B
- Gramatička forma: niska koja se izvodi iz S (npr. B, bc, bcc, bccc, cccc, 1ccc, 12cc, 123c, 1234)
- završna niska: niska koja se izvodi iz S i pripada T*, npr. 1234
- Jezik generisan gramatikom: skup završnih niski