Korpusno jezikoslovje
2008/2009

Univerza v Novi Gorici
Fakulteta za humanistiko

2. kolokvij 16/1/09, 11:30 - 12:30

1. naloga (3 točke)

Naštejte in opišite vsaj tri načine, kako se lahko jezikoslovno označuje korpuse in vsaj tri ravni jezikoslovnega označevanja.

Odgovor:
Označuje se jih lahko ročno ali pa strojno z ročno napisanimi pravili oz. z avtomatsko naučenimi pravili. Pri ročnem označevanju jezikoslovec s pomočjo primernega urejevalnika označuje korpus, pri čemer je potrebna natančna definicija nabora dovoljenih kategorij oz. relacij. Strojno označevanje z ročno napisanimi pravili je klasični način pisanja programov za jezikoslovno analizo, kjer je potrebno pisati formalna pravila, omejena glede na izbrano teorijo, formalizem in implementacijo. Pri avtomatskem označevanju s strojno naučenimi pravili se program nauči modela jezika na osnovi ročno označenih podatkov (korpusa).

Označujemo lahko npr. besede z oblikoslovnimi oznakami, z lemami, ali pa povedi s skladenjskimi odnosi. Primer prvih dveh so oznake v korpusu FidaPLUS, npr. označitev besede »človeka« z oblikoslovno oznako »Somer« in lemo »človek«, primer skladenjskega označevanje pa drevesnica SDT.

2. naloga (2 točki)

V čem je razlika med zapisi HTML in XML?

Odgovor:
HTMLima vnaprej določen nabor oznak, ki so opisujejo usmerjene predvsem videz okumenta, oznake lahko izpuščamo, strani v HTML pa dostikrat niso pravilno napisane. Pri XML oznake definiramo sam, tipično opisujejo pomen dokumenta, vse oznake morajo biti prisotne, dokumenti pa morajo biti dobro napisani.

3. naloga (3 točke)

Iz niza »Janez se uči XML & SGML.« naredi dobro napisan dokument XML, pri čemer uporabite element <poved> za označitev povedi, <b> za besedo in <l> za ločilo. Besede naj imajo atribut »lema« z ustrezno vrednostjo.
Za dodatni 2 točki napišite še DTD za dokument.

Odgovor:

XML:

```
<?xml version="1.0" encoding="utf-8"?>
<poved>
<b lema="Janez">Janez</b>
<b lema="se">se</b>
<b lema="učiti">uči</b>
<b lema="XML">XML</b>
<l>&amp;</l>
<b lema="SGML">SGML</b>
<l>.</l>
</poved>
```

DTD:

```
<!ELEMENT poved (b | l)+>
<!ELEMENT b (#PCDATA)>
<!ELEMENT l (#PCDATA)>
<!ATTLIST b
          lema CDATA #REQUIRED>
```

4. naloga (2 točki)

V Prešernovi pesmi:

> Al prav se piše kaša ali kaſa,
> se šola novočŕkarjov srdita
> z ljudmi prepira starega kopita;

se pojavita dva, za sodobno slovenščino nestandardna znaka. Napišite kako se imenujeta v unikodu in kateri kodi imata.

Odgovor:

- ſ je LATIN SMALL LETTER LONG S s kodo 017F (Unikod blok »Latin Extended-A«)
- ŕ je LATIN SMALL LETTER R WITH ACUTE s kodo 0155 (Unikod blok »Latin Extended-A«)
- ŕ lahko tudi napišemo kot kombinacijo navadnega »r« (LATIN SMALL R, koda 072) z znakom COMBINING ACUTE ACCENT, koda 0301 (blok Combining Diacritical Marks)