Kratek pregled jezikoslovnih zadreg pri oblikoslovnem označevanju slovenščine

Simon Krek,
Odsek za tehnologije znanja,
Institut Jožef Stefan

Koper, 4. 4. 2008

O čem govorimo?

- segmentacija
- tokenizacija
- lematizacija
- oblikoslovno označevanje (ang. tagging)
- skladenjsko razčlenjevanje (ang. parsing)
- semantično razdvoumljanje
- označevanje oz. razdvoumljanje sonanašanja
- prepoznavanje lastnih imen (NER)

Slovenske tabele oznak

- Multext(-East) – JOS
  - na podlagi:
    - EAGLES (Expert Advisory Group on Language Engineering Standards), ISLE (International Standard for Language Engineering)
  - IJS: enojezični (ORW, VAYNA, GORE, DSI), dvojezični (SVEZ, ELAN, TRANS)
    - http://nl2.ijs.si
  - FIDA, FidaPLUS, KoRP itd.
  - LC-Star (FERI, UM)
  - Presis, BesAna, Ases itd. (Amebis d.o.o.)

korpus Beseda (ZRC SAZU)

Spremembe MTE – JOS

- od oblikoskladenjskega izhodišča bliže oblikoslovnemu
  - oblikoslovna informacija mora biti vsebovana že v sami obliki (prim. negacija)
  - skladenjske informacije označujemo na višji, načeloma neodvisni ravni (prim. sest. gl. časi)
- prilagoditev tradicionalnega slovničnega modela za potrebe avtomatskega označevanja velike količine besedil
- racionalizacija dolžine oznak (prej zaradi združljivosti z drugimi jeziki)

MTE = JOS

- 1 samostalnik S
- 2 glagol G
- 3 pridevnik P
- 4 prislov R
- 5 zaimek Z
- 6 števnik K
- 7 predlog D
- 8 veznik V
- 9 členek L
- 10 medmet M
- 11 okrajšava O
- 12 neuvrščeno N

MTE=JOS

- pregibne + odprte (lematizacija)
  - samostalnik, glagol, pridevnik, števnik
- pregibne + zaprte (lematizacija)
  - zaimek
- nepregibne + odprte
  - prislov, okrajšava?, medmet?
- nepregibne + zaprte
  - veznik, predlog, členek
- ostane
  - neuvrščeno

Primer “glagol” MTE

0	besedna vrsta	glagol
1	vrsta	polnopomenski naklonski vezni
2	glagolska oblika	povednik velelnik pogojnik nedoločnik deležnik namenilnik
3	čas	sedanjik prihodnjik nesedanjik
4	oseba	prva druga tretja
5	število	ednina množina dvojina
6	spol	moški ženski srednji
7	način	tvornik trpni deležnik
8	nikalnost	nezanikani zanikani
14	vid	nedovršni dovršni

Primer “glagol” JOS

0	besedna vrsta	glagol
1	vrsta	glavni pomožni
2	vid	dovršni nedovršni dvovidski
3	glagolska oblika	nedoločnik namenilnik deležnik sedanjik prihodnjik pogojnik velelnik
4	oseba	prva druga tretja
5	število	ednina množina dvojina
6	spol	moški ženski srednji
7	nikalnost	nezanikani zanikani

glagol – vrsta

- MTE:
  - polnopomenski // naklonski // vezni
- JOS:
  - glavni // pomožni
- naklonskost ni izražena samo glagolsko
- če je razlog skladenjski, manjkajo drugi (npr. fazni)
- JOS: izločen samo pomožnik “biti”

glagol – oblika

- MTE:
  - povednik // velelnik // pogojnik // nedoločnik // deležnik //namenilnik
- JOS:
  - nedoločnik // namenilnik // deležnik // sedanjik // prihodnjik // pogojnik // velelnik
- kategorija “čas” se v celoti prestavlja v “obliko”
- pri kategoriji deležnik ostane le opisni deležnik za tvorbo časov
- ostale deležniške oblike v celoti gredo pod pridevnik (vrsta: deležniški)

glagol – vid

- MTE:
  - dovršni // nedovršni
- JOS:
  - dovršni // nedovršni // dvovidski
- vid je “leksikonska” lastnost
- semantično razdvoumljanje je izvzeto iz oblikoslovne ravni označevanja
- prim. stiskati, vpiti, zamikati, -irati

Primer: označevanje samostalnikov

- glagolnike označujemo kot samostalnike
- posamostaljene pridevnike označujemo kot pridevnike, ne kot samostalnike
- živost označujemo samo pri samostalnikih moškega spola, kadar se pojavljajo v ednini v tožilniku

Lastna - občna

- kot občni samostalnik označujemo tudi citatno pisane samostalnike iz drugih jezikov (*slasher*)
- kadar je v besedilu samostalnik, ki ga prepoznamo (iz oblike, konteksta, splošnega znanja) za stvarno ime, ga označimo kot lastno (lema z veliko začetnico) samo, če nima prekrivne leme z občnim samostalnikom. Kadar je stvarno ime prekrivno z občnim samostalnikom, ga označimo za občni samostalnik, lema v tem primeru z malo: *Nastale so klasike Rozmarijin otrok, Izganjalec duhov in slasher Noč čarovnic.*
- kot občna označujemo kratična imena tipa WAP, GPRS, USB

Lastna - občna

- kot lastna imena označujemo kratična imena:
  - geografskih oz. političnih enot: EU, RS, ZR itd.
  - političnih strank: CDU, SD, SDS itd.
  - podjetij in organizacij (tudi njihovih delov): BBC, BTC, DURS, DZ, RTV, SFOR, OŠ itd.
  - kratice zakonov: ZDR, ZJU, ZSPJS itd.
  - kot lastnih imen ne označujemo delov kratičnih poimenovanj blagovnih znamk ali imen programskih orodij: M30X, pocketPC, palmOS, LX, DNKA, (F 650) CS

Pridevnik

- vrsta: splošni, svojilni, deležniški
- stopnja: nedoločeno, primernik, presežnik
- spol, sklon, število
- določnost: določni nedoločni

- kje so vrstni/kakovostni?

Pridevnik

- Deležniški so tisti, ki imajo bližnjo povezavo s svojim glagolom, kar se tiče vezljivosti, kolokabilnosti in podobnih odnosov. Primer odločanja (prvih 10 kolokatorjev eno mesto desno po korpusu FidaPLUS):
- OPRIJETI brez OPRIJET
- OPRIJET brez OPRIJETI
- 1 vzdevek hlače
- 2 ime majica
- 3 sloves kavbojke
- 4 se oblačilo
- 5 naziv majčka
- 6 telo obleka

Enoznačnost označevanja?

- primeri, kjer ni mogoče določiti neke lastnosti zato, ker:
  - referent besedne oblike ne obstaja
  - iz okolice ni mogoče ugotoviti, za katero lastnosti gre
  - oblikoslovno pa je možnih več lastnosti

- spol: a. srednji spol v primerih, ko gre v osnovi za pridevniške pojavnice, prim. *Iskreno ozračje groze, ki resnično krepi občutek neznanega in skrivnostnega*

Enoznačnost označevanja?

- moški spol v primerih, kjer gre za posamostaljene pridevnike, prim. *Poudarek bo veljal razvoju gospodarstva, da bo mladim omogočena zaposlitev.*
- množina, če ni mogoče zanesljivo ugotoviti, ali gre za dvojino ali množino, prim. *Ni razloga, da bi sodišče nadaljevalo postopek, za katerega nobena od strank ne izkaže pravega interesa.*
- če gre za pravo pomensko dvoumnost, pripišemo tisto vrednost, ki se nam po najboljši presoji zdi verjetnejša, prim. *Če ribe ne pokažejo zanimanja za muho ali blestivko, nam za asa v rokavu ostaneta rakec ali črv.*

Iztočnice za debato?

- standardizacija
  - da ali ne?
  - stopenjskost?
  - združljivost
- prosta dostopnost
  - odprta koda
  - programski paket (OS?)
- dokumentiranost