REPUBLIKA SLOVENIJA
MINISTRSTVO ZA ŠOLSTVO IN ŠPORT

Naložba v vašo prihodnost
OPERACIJO DELNO FINANCIRA EVROPSKA UNIJA
Evropski socialni sklad

SPORAZUMEVANJE V SLOVENSKEM JEZIKU – GOVORNI KORPUS

Vrsta dokumenta: navodila za standardizacijo zapisa govora

Datum: 20. julij 2010
Verzija: koncna-01

KAZALO

1 CILJ

Standardizacija zapisa govora omogoča avtomatsko iskanje po besednih oblikah z enako oblikoslovno in semantično vlogo, a različnimi glasovnimi podobami, ter avtomatsko označevanje besedila (tokenizacija, lematizacija, oblikoslovno označevanje, skladenjsko označevanje).

2 JEZIKOSLOVNI VIDIK STANDARDIZACIJE ZAPISA

2.1 NAČELA STANDARDIZACIJE ZAPISA

Pri pretvorbi pogovornega v standardizirani zapis odpravimo glasoslovne premene, ki so prisotne pri posamezni besedni obliki, ob upoštevanju pogostosti rabe. Ciljna oblika je knjižna različica istega leksema. Na drugih jezikovnih ravneh besed ne spreminjamo. Za ločevanje, kdaj gre za glasovno premeno in kdaj ne, se oblikujejo primeri dobre prakse, popisani v sekciji 2.3 tega dokumenta. Če določenega leksema ni v knjižni normi, ga ohranimo v obliki, ki se pojavlja v govoru.

Podrobna razčlemba
Standardizirani zapis govora mora ohraniti vse jezikovne lastnosti gradiva, ki so pomembne za nadaljnje stopnje označevanja:

- tokenizacijo
- lematizacijo
- oblikoslovno označevanje
- skladenjsko označevanje

To pomeni:

a) standardizirani zapis mora upoštevati potencialno lemo vsake posamezne besedne oblike in slediti ideji, da kar najbolj olajša avtomatsko lematizacijo korpusa
b) standardizirani zapis ne sme spremeniti leksema
c) standardizirani zapis ne sme spremeniti formalnih oblikoslovnih lastnosti (besedna vrsta, spol, sklon, število, oseba, tip...) besede
d) standardizirani zapis ne sme spremeniti skladenjskih lastnosti besedila (stavčni členi, struktura besednih zvez...)
e) standardizirani zapis poenoti različne, predvsem glasoslovne variacije določene besedne oblike tako, da jih lahko v nadaljevanju vodimo pod eno enoto, ki naj sovpada s knjižno enoto, če taka enota obstaja v knjižnem jeziku

2.3 VIRI ZA POMOČ PRI STANDARDIZACIJI ZAPISA

Slovarski: SSKJ, besedišče, pravopisni slovar 2001, Veliki slovar tujk

Korpusni: Fida+, Nova beseda, korpus JOS

Jezikovnotehnološki: specifikacije JOS, specifikacije leksikona SSJ

Internetni: Google

2.3 PRIMERI DOBRE PRAKSE

2.3.1 GLASOSLOVNA RAVEN

Spremenimo:

a) redukcije
bl > bolj, kešnmu > kakšnemu, kolk > koliko, pršl > prišlo, tolk > toliko, tle > tule...

b) premene
a uš > a boš, bliži > bliže, druzga > drugega, fčeraj > včeraj, gnar > denar, de na bi > da ne bi, dej > daj, jejli > jedli, jeskat > iskat/i, kukr > kakor, lohka/lah > lahko, mejčken > majčken, mojmo brato > mojemu bratu, nej blu > ni bilo, neki > nekaj, rečmo > recimo, spomnem > spomnim, tazga > takega, tok > tako/toliko, udspredi > odspredaj, štengca -> štengica, zadi > zadaj, zdele > zdajle, zred > zaradi, zmiri > zmeraj...

2.3.2 LEKSEMSKA RAVEN

Pogovorne besede, ki bi jim težko določili povsem ustrezno knjižno različico, ohranjamo. Pri odločitvah glede zapisa se opiramo na pisne korpuse in druge vire.

Ohranimo:

a) izposojenke
bek, čuješ, fak, fajrala, ferker, ful, gruntali, hambrt, kafič, kao, kuhla, može, ni mus, ornk, pašeš, plata, pošlihtaš, rajsar, ratati, singl, spedenan, štima, šparati, valjda, ziher, žijaš...

b) besedilne aktualizatorje, zaimke, pridevnike, členke...
- a (a se čva midva umakniti [ime]?)
- anche (zdaj bojo oni ki so kriminalci imeli anche svoj urad ne)
- bem (bem vsake toliko zadržijo kaj zase ne)
*- en, ena, eno (*nosilec besedilne nedoločnosti: *eno štorijo ti bom povedal o enem profesorju fizike)*
*- hal (*avstrijska Koroška: *tisto pa to sem lepo naredila hal?)*
- jel (ko že zdaj tudi nisi tako grozno mlad jel)
- ma (saj pol vidim ma če jih nimam gor pa nankar gospoda h oltarju ne vidim)
- nankar (je pa rekla da je bilo tako fajn da ne ve kako bo upala vsem sošolcem nankar povedati)
- ta (ki ustreza funkciji določnega člena pred pridevniško besedo in je za vse tri spole enak, prav tako je enak v vseh sklonih*: ta mlada, ta stara, ta rdeč avto)*

- te (MB: *kaj si te tam delala?; kako jo te to ožemaš; kaj sem te hotela zdaj reči; koliko sta te dala za to?; kaj te vem),* **vendar spremenimo:**

ko ima pogovorna oblika enako oblikoskladenjsko in pragmatično vlogo kot knjižna: *te > potem (ja v to v to sploh ne dvomim ne ker potem bi se po mojem skoz kregali; kaj pa še potem ima [ime] sploh?)*

c) pogovorne besede, ki jim knjižni izvor sicer lahko določimo, vendar knjižna različica tako rekoč ni v rabi

tlele...

Spremenimo:

a) če je raba neenotna in oblika besede variira po regijah ali govorcih, določimo enotno krovno obliko

- jest, jz, jst > jaz
- kva, kej, ka, kogà > kaj (in koga si kupu? > in kaj si kupil?)
- pol, pole, pouli, puole -> pol
- lej, glej > glej
- ta, toti, teti > ta (sicer so nori ti kolegi ker pol so šli na Damjana Murkota; ti si zdaj ekspert za te Čehinje pa te Poljakinje pa to)
- prislov *ene (ob ene sedmih,* prisotna tudi pregibna oblika: *imam v enih treh vrečkah > ima v ene treh vrečkah)*
*- un, gun, uen, oni > oni (*kazalec zunajbesedilne predmetnosti*: oni je šel pa kar za njo; pol smo pa kupili ono ta drugo kljuko)*
- anke, anka > anche
- nenka, nenkar, nankar > nankar
- al, ali > ali (kaj bi pa ti izbral to ali to)
*- ta, tada > tedaj (*avstrijska Koroška*: tedaj pa ta luftbalon poštrihajva)*
*- oba, aba > aber (*avstrijska Koroška*: ja aber pol pa morava spet po travi)*

2.3.3 OBLIKOSLOVNA RAVEN

Na oblikoslovni ravni pogovorne oblike besed spremenimo, pri čemer pazimo, da ohranimo prvotno oblikoskladenjsko in pragmatično vlogo besed.

Spremenimo:

a) pogovorne oblike glagolov

- narest, naret > naredit(i)
- rečt, pečt, vržt > vreč(i)
- najdit, najdt > najt(i)
- najdli > našli

b) kratki nedoločnik

- se niso pripravljeni pogovarjat > pogovarjati
- ne bi želel reč nič drugega kot to > reči

- bi se pa upal trdet in mogoče malo korigirat eee je pa definitivno mimo doba špekulacij in pač potegnt določene poteze določene ukrepe > trditi, korigirati, potegniti

c) regionalne različice besednih oblik

- boma > bova (*pa saj bo mislim boš videla da bomo nekaj si bova našle midve bova itak milijonarke pol ko bova velike*)

Ohranimo:

a) če knjižna oblika obstaja, vendar z drugačno oblikoskladenjsko vlogo

*- čem, češ, čmo > čem (*lema 'hočem': *kako čmo temu reči a je to šlamparija; a čmo iti v kino),* **vendar spremenimo:**

če je oblikoskladenjska vloga enaka tisti, ki jo ima normativna različica: *češ > hočeš (a češ čokolado > a hočeš čokolado)*

b) daljšanje osnove (s '*t'...*)

Markota...

2.3.4 SKLADENJSKA RAVEN

Ohranimo:

Na skladenjski ravni pogovorne prvine ohranjamo, ker ne vplivajo na lematizacijo. Primeri:

- *sva šle; ste šla; ste izjavil; bova počasi morale zaključiti; ko bova velika bova milijonarke...*
- *te dva problema; vso drevje se suši; ta dvigalo je pokvarjen...*
- *s čem vse; s svojimi otroci; od kje vse so prišli; odvisno od kje gledamo; moram živeti v temu delu...*

- Enrique Iglesias nam je polepšal ponedeljkov dopoldne...
- to si mi že zadnjič načel pa nisi nič dokončal to debato; potem pa so zavodi ugotovili da to ne morejo financirati ne...
- bi mogli it; on more biti pripravljen in bo mogla hoditi na ta drug oddelek; moreš preživeti kaj češ; poseglo se je samovoljno brez vprašati ljudi; mogel bi prekiniti službo...
- to so vse male naklade; zdaj imam pa čisto mali avto...
- sem bil skoz večji; mater kakšno gužvo imata skozi...
- kaj ji je že ime; noben si ne bo dovolil; gre za eden interes farmacevtske industrije po razvoju in profitu; če bomo pošiljali v tujino civiliste se je treba zavedati da bo potrebno ustanoviti tisto kar ta država nima; to se nisem jaz zmislil...

3 TEHNIČNA NAVODILA ZA STANDARDIZACIJO ZAPISA

1. Če eno besedo v prvotni transkripciji pretvorimo v dve ali več besed v standardiziranem zapisu, zapišemo te besede v standardiziranem zapisu stično z znakom »+« (plus)
   navm -> ne+bom
   dount kraj - > don'+t cry

2. Če po dve ali več besed v prvotni transkripciji pretvorimo v eno besedo v standardiziranem zapisu, zapišemo to besedo v standardiziranem zapisu z znakom »_« (podčrtaj) stično
   v pričo-> v_pričo
   Mak Donalc -> Mc_Donald+'s

3. Onomatopeje, medmete, besedne fragmente in druge glasove, za katere v knjižnem jeziku ni standardnega zapisa, pustimo zapisane tako, kot so bili zapisani v prvotni transkripciji.

4. Lapsuse v izgovorjavi, če so nedvoumni, odpravimo.
   indidualnih -> individualnih

5. Zloženke pišemo enako kot v prvotni transkripciji, samo skupaj ali narazen, brez vezajev.

6. Kratice zapišemo z velikimi črkami.
   ce -> C
   veveve -> WWW
   es i -> S_I
   ha pe -> H_P
   s o n c h e k -> S_ O_ N_ C_ H_E_K
   kagebe -> KGB
   ertees -> RTS

   Če so že izpričane v korpusih in drugod v besedilih, imajo prednost izpisane besede.
   piar proti *PR*

   Spletne naslove pišemo po načelu:
   WWW pika Maribor T_O_U_R_I_S_M pika S_I ali
   WWW pika Kompas pika SI (ali *S_I*, če je na prvem nivoju *es i*).
   Spletne naslove, ki vsebujejo večdelna ali osebna imena, pišemo po načelu
   WWW {Radio Aktual} pika SI
   WWW [ime] [priimek] pika com

7. Tuja lastna imena zapišemo po pravopisni normi.
   {Kos Porta} -> {Cost Porta}
   {Šarm el Šejk} -> {Sharm El Sheikh}
   Atene -> Atene
   {vadi Mudžep Kinks Vej} -> {Wadi Mudžep Kings Vej}

8. Citatne občne besede zapišemo po pravopisni normi, ali citatno ali v poslovenjenem zapisu, pri čemer se odločimo za tisto obliko, ki je v virih (korpusi, internet...) bolj pogosta. Vezajev ne pišemo, ampak se odločimo za zapis ali skupaj ali narazen.
   granč scena **->** *grunge* scena
   ofišl suporterja -> official supporterja
   rentakar -> rentacar
   sori -> sori
   tu mač -> tu mač
   pab -> pub

9. Ločil ne spreminjamo in ne dodajamo (tudi ne pik in vejic).

10. Začetki izjav ostanejo z malo začetnico.

11. Za vprašljive, negotove primere knjižnega zapisa dodamo poseben znak '*' (zvezdica).