Standardi za zapis korpusov

Tomaž Erjavec
Korpusno jezikoslovje
UNG
2008/2009

12. 12. 2008

Pregled predavanja

1. čemu standardi
2. XML in TEI
3. oblikoslovno označevanje

Kaj so standardi?

- SSKJ: za posamezno državo obvezen enotni predpis za mere, kakovost izdelkov // kar določa, kakšno sme, mora kaj biti
- konsenzualno sprejeti predpisi, ki so javni in vsebujejo jasne definicije
- glavni namen je poenotiti industrijsko prakso na posameznih področjih z namenom, da se olajša izmenljivost

Zgodovina

- XVIII stoletje: v Franciji ima vsaka regija (vas) svoje merske enote, poleg tega pa ima npr. njiva lahko drugačno "mero" kot vinograd
- proces definicije enotne merske enote iz katere bi bilo mogoče izpeljati vse ostale → meroslovje
- meter: ena destmilioninka dolžine poldnevnika skozi Paris, od severnega tečaja do ekvatorja
- pomembnost standardizacije naraste z industrijsko revolucijo: vijaki, elektrika, gradbeništvo…
- sedaj standardi tudi za informacijske tehnologije in npr. organizacijo podjetij (ISO 9000)
- podjetja, ki preverjajo upoštevanje standardov (akreditacija, npr. SIQ)

Standardizacijska telesa

izdajajo standarde po natančnem postopku, kjer sodelujejo predstavniki vključenih držav:

- nacionalni standardi: DIN, ANSI, JUS
- **SIST**: Slovenski inštitut za standardizacijo
- mednarodni standardi: IEC, ISO
- ISO International Organization for Standardization, Geneva (1947)
- ISO razdeljen na tehnične odbore (Technical Committee, TC) s člani iz sodelujočih držav, ki nato sprejemajo standarde iz svojega področja

Standardi in jeziki

- vsak standard mora vsebovati terminološke definicije
- standardi ISO se lahko tudi prevajajo
- ISO TC 37: Technical Committee on **Terminology and other language and content resources** (SC4: Language Resources Management)
- Novo telo za standarde, vezane na splet: W3C: The World Wide Web Consortium
- začetek delovanja W3C: HTML

Zakaj standardi za digitalni zapis podatkov

Zapis računalniških podatkov je bil tipično vezan na določen program, npr. urejevalnik.

Problemi:

- *trajnost:* z bliskovitim napredkom tehnologije programi zelo hitro zastarajo, podatki pa postanejo neberljivi
- *izmenljivost*: podatki so vezani na konkretno računalniško platformo
- *uporabnost:* podatki so težko uporabni za nov namen
- *razumljivost:* podatki so razumljivi samo programu oz. njegovim razvijalcem (ni javnih in stabilnih specifikacij)
- *preverljivost:* ne vemo, ali so podatki zapisani v skladu s specifikacijami ali ne

Jezikovni podatki

- urejevalniki besedil: preohlapen zapis, preveč usmerjen v izgled
- podatkovne baze: preveč omejen zapis, ne dopušča mešanja vsebine (besedila) in strukture (oznak)
- ISO 8879 SGML (Standard Generalised Markup Language), 1986
- določa jezik za predstavitev dokumentov nad katerimi bodo delovali programi za procesiranje besedil

SGML

Zagotoviti način zapisa, ki je:

- prenosljiv med računalniškimi platformami
- odporen na tehnološke spremembe
- omogoča uporabo dokumentov v različne namene
- omogoča avtomatsko preverljivost, ali je nek dokument zapisan v skladu s standardom

Problemi s SGML

- standard je zelo kompleksen
- orodja za uporabo “akademska” ali zelo draga
- konverzija dokumentov v SGML je bila velika naložba
- potreba po standardu predvsem v velikih podjetjih (dokumentacija)
- zato je bila uporaba SGML razmeroma omejena

Svetovni splet

- HTML (1989) je zapisan v SGML
- vendar pa s SGML skladen HTML uporablja zelo malo internetnih strani..
- HTML je tudi premalo ekspresiven za zapis poljubnih spletnih podatkov
- potreba po novem standardu za zapis mrežnih podatkov, ki naj bi imel vse prednosti SGML, brez njegovih slabosti
- → eXtended Markup Language, XML (1998)

XML sedaj

- XML postal izredno popularen, in postaja univerzalni način zapisa (jezikovnih) podatkov
- veliko pridruženih standardov
- veliko število prosto dostopnih orodij za procesiranje XML
- veliko programov že podpira izvoz/uvoz podatkov v XML

Extended Markup Language XML

- XML je definicija od platforme neodvisnih method za hranjenje in procesiranje besedil v elektronski obliki
- XML je “metajezik” – jezik za opis drugih jezikov, v katerem lahko definiramo svoje lastne jezike za označevanje različnih zvrsti besedil
- XML je projekt konzorcija W3C, zato je specifikacija XML odprta in nima lastnika
- XML je podmnožica SGML

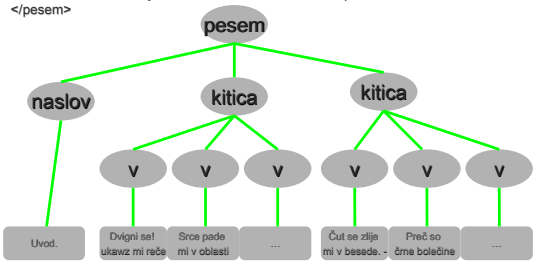
Dokument XML

```
<pesem>
  <naslov>Uvod.</naslov>
  <kitica>
    <v>Dvigni se! ukawz mi reče.</v>
    <v>Srce pade mi v oblasti</v>
    <v>Silne, prej neznane strasti,</v>
    <v>Ki ko živi ogenj peče.</v>
  </kitica>
  <kitica>
    <v>Čut se zlije mi v besede. -</v>
    <v>Preč so črne bolečine,</v>
    <v>Strast občutkov divjih mine,</v>
    <v>Jasen mir se v prsi vsede.</v>
  </kitica>
</pesem>
```

- **dokument = besedilo + oznake**
- **element = začetna oznaka + vsebina + končna oznaka**
- **element vsebuje besedilo ali elemente ali oboje (ali nič)**

Hierarhične strukture

<pesem><naslov>Uvod.</naslov> <kitica> <v>Dvigni se ukawz mi reče.</v> <v>Srce pade mi v oblasti</v> <v>Silne, prej neznane strasti,</v> <v>Ki ko živi ogenj peče.</v> </kitica> <kitica> <v>Čut se zlije mi v besede. -</v> <v>Preč so črne bolečine,</v> <v>Strast občutkov divjih mine,</v> <v>Jasen mir se v prsi vsede.</v> </kitica> </pesem>



Prazne oznake

- oznake z vsebino:
  <oznaka> … </oznaka>
- prazne oznake nimajo vsebine:
  <oznaka/>
- uporabljajo se za označevanje “točk” v dokumentu, npr. prelomi strani
- v resnici
  <oznaka/> = <oznaka></oznaka>

Atributi XML

- elementom XML lahko pripišemo lastnosti
- lastnosti zapišemo v začetne oznake kot pare
  atribut = "vrednost"
- vrednost mora biti v enakih enojnih ali dvojnih narekovajih: " ali '

Npr.

<prelom stran= "11" />
<razdelek številka= "5.1" zvrst="podpoglavje"> …
<recept vir='http://nl.ijs.si/recepti/kaj="pizza"'> …

Primer: oznake v korpusu

```
<s id="Osl.1.2.2.1">
  <w lemma="biti" ana="Vcps-sma">Bil</w>
  <w lemma="biti" ana="Vcip3s--n">je</w>
  <w lemma="jasen" ana="Afpmsnn">jasen</w>
  <c>,</c>
  <w lemma="mrzel" ana="Afpmsnn">mrzel</w>
  <w lemma="aprilski" ana="Aopmsn">aprilski</w>
  <w lemma="dan" ana="Ncmsn">dan</w>
  <w lemma="in" ana="Ccs">in</w>
  <w lemma="ura" ana="Ncfpn">ure</w>
  <w lemma="biti" ana="Vcip3p--n">so</w>
  <w lemma="biti" ana="Vmps-pfa">bile</w>
  <w lemma="trinajst" ana="Mcnpnl">trinajst</w>
  <c>.</c>
</s>
```

Primer: zapis slovarja

```
- <entry id="jaslo.4509">
  - <form type="hw">
      <orth type="roma">shuurisuru</orth>
      <orth type="kana">しゅうりする</orth>
      <orth type="kanji">修理する</orth>
    </form>
  - <gramGrp>
      <pos>Vs</pos>
      <subc>trans.</subc>
    </gramGrp>
  - <trans>
      <tr>popraviti</tr>
    </trans>
  - <eg>
      <q>ラジオがこわれたので修理した。 </q>
      <tr>Ker se je radio pokvaril, sem ga popravil.</tr>
    </eg>
  - <eg>
      <q>そろそろ屋根（やね）を修理してもらわなければならない。 </q>
      <tr>Počasi bomo morali dati popraviti streho.</tr>
    </eg>
  - <xr type="lesson" n="L1.7">
      <xref>1. letnik, lekcija 7</xref>
    </xr>
    <usg type="level">0</usg>
    <note type="admin" resp="TER">2005-07-11 Add Romaji</note>
    <note type="admin" resp="TER">2005-07-10 Add levels</note>
    <note type="admin" resp="KHS">2003-03-12 L1 (642)</note>
    <note type="admin" resp="VOJ">2005-02-22 V (342)</note>
    <note type="admin" resp="ISE">2005-02-28 Merge</note>
  </entry>
```

Entitete (delci)

- dokument XML lahko vsebuje tudi delce, ki se ob procesiranju nadomestijo z nečim drugim
- sklic na entiteto se začne z znakom “in” in konča s podpičjem: &…;
- predefinirane entitete za posebne znake:
  < = < > = >
  & = &
  ' = ' " = "
- 1 < 2 (formula) →
  <formula>1 < 2</formula>
- Procter & Gamble (podjetje) →
  <podjetje>Procter & Gamble</podjetje>

Dobro napisani dokumenti XML

- dokument se začne s prologom XML:
  <?xml version="1.0"?>
- oznake in entitete so zapisane pravilno
- vsaki začetni oznaki ustreza končna oznaka (<ime> ≠ <Ime> ≠ <IME>)
- oznake so pravilno gnezdene
  Narobe: <a>………
- dokument ima en sam vrhnji element

→ dobro napisan (well-formed) dokument XML

Kaj vse je narobe?

```
<?xml version="1.0"!>
<mesto pomembno geo="13°43"59'N 45°59"55'W ">
   <ime>Nova Gorica,
   <prebivalcev>prebivalcev = 36.155</Prebivalcev>
   <znamenitosti>Politehnika<zanmenitosti>
</mesto>
<mesto geo="13°44"3'N 45°59"58'W ">
   <ime>Postojna,
   <prebivalcev> prebivalcev < 300.000</prebivalcev>
   <znamenitosti>Postojnska jama</Znamenitosti>
</mesto>
```

Definicije tipov dokumentov

- DTD poda formalno gramatiko elementov za določen tip dokumentov
- določi kateri elementi so dovoljeni, kateri obvezni, in v kakšnih medsebojnih razmerjih lahko nastopajo
- določi dovoljene in obvezne atribute elementov in določi tip njihovih vrednosti
- DTD naj bi vseboval tudi dokumentacijo, ki pove kaj elementi *pomenijo*

Enostaven DTD

Dokument XML:

```
<mesto>
    <ime>Nova Gorica</ime>
    <prebivalcev>36.155</prebivalcev>
    <znamenitosti>Politehnika</znamenitosti>
</mesto>
```

DTD:

```
<!ELEMENT mesto (ime, prebivalcev, znamenitosti)>
<!ELEMENT ime      (#PCDATA)>
<!ELEMENT prebivalcev    (#PCDATA)>
<!ELEMENT znamenitosti  (#PCDATA)>
```

Bolj kompliciran DTD

```
<!ELEMENT antologija (pesem+)>
<!ELEMENT pesem (naslov?, kitica+)>
<!ELEMENT naslov (#PCDATA) >
<!ELEMENT kitica (v+) >
<!ELEMENT line (#PCDATA) >
```

```
<antologija>
 <pesem>
  <naslov>Uvod.</naslov>
  <kitica>
   <v>Dvigni se! ukawz mi reče.</v>
   <v>Srce pade mi v oblasti</v>
  </kitica>
  <kitica>
   <v>Čut se zlije mi v besede. -</v>
   <v>Preč so črne bolečine,</v>
  </kitica>
 </pesem>
</antologija>
```

```
<antologija>
 <pesem>
  <kitica>
   <v>Dvigni se! ukawz mi reče.</v>
  </kitica>
 </pesem>
 <pesem>
  <kitica>
   <v>Čut se zlije mi v besede. -</v>
   <v>Preč so črne bolečine,</v>
  </kitica>
 </pesem>
</antologija>
```

Operatorji v DTDjih

- združevanje: (in)
- sledi: ,
- ali: |
- ponavljanje: ? (0 ali 1x), * (0, 1, …), + (1, 2, ...)

```
<!ELEMENT pesem
    (naslov?, ( (v+ ) | (refren?, (kitica, refren?)+)))
>
<!ELEMENT ljudje (moški | ženska)+ >
<!ELEMENT odstavek (#PCDATA | hi | b)* >
```

Atributi

V DTD:

```
ime atributa;  tip atributa;        status atributa
<!ATTLIST tabela
    tip         CDATA               #IMPLIED        dovoljen atribut
    id          ID                  #REQUIRED obvezen atribut
    status      ( osnutek |
                  popravljeno |
                  končno )          "osnutek“       privzeta vredn.
>
```

V dokumentu XML:

```
<tabela id="tab.12" tip="sumarna" status="popravljeno">
```

Pravilen (valid) dokument XML

- dokument vsebuje ali se sklicuje na DTD
- dokument je dobro zapisan in skladen s podanim DTD-jem

```
<!DOCTYPE mesto SYSTEM “http://mesta.net/mesto.dtd”>
<mesto>
    <ime>Nova Gorica</ime>
    <prebivalcev>36.155</prebivalcev>
    <znamenitosti>Politehnika</znamenitosti>
</mesto>
```

Preverjanje XML

- dobro napisanost in pravilnost dokumentov XML preverjamo z razčlenjevalnikom XML
- obstaja veliko število razčlenjevalnikov, tudi vgrajenih v aplikacije, npr. internet brskalnike
- Firefox, Internet Explorer, Word: prikažejo strukturo dokumentov XML
- za urejanje je najbolje imeti specializiran urejevalnik XML

Razlike med HTML in XML

HTML	XML
vnaprej določen nabor oznak	oznake definiramo sami
oznake usmerjene v videz dokumenta	oznake opisujejo pomen dokumenta
oznake lahko izpuščamo	vse oznake morajo biti prisotne
strani dostikrat niso pravilno napisane	dokumenti morajo biti dobro napisani

Pridruženi standardi

- sheme XML: bolj kompleksni DTD
- XSLT: pretvorba XML v XML, HTML ali navadno besedilo
- Xlink, Xpointer: povezovanje dokumentov XML
- XQuery: iskanje po dokumentih XML
- …

Ampak kaj z dokumentov v XML dejansko počnemo??

- ..karkoli; osnovna ideja je, da so dokumenti XML neodvisni od aplikacije in prenosljivi
- kar pa pomeni, da jih bo za dejansko uporabo treba prej verjetno še predelati
- nato pa lahko npr. isti dokument uporabimo:
  - za tisk
  - kot del podatkovne baze za iskanje
  - za predstavitev na medmrežju
  - …

Text Encoding Initiative

- iniciativa za zapis besedil TEI (Text Encoding Initiative) je bila ustanovljena leta 1987
- namen: standardiziracija zapisa besedil, ki bi se uporabljala pretežno v znanstvene namene
- razlog: zmanjšati razdrobljenost obstoječih načinov digitalnega zapisa, poenostaviti računalniško obdelavo in spodbuditi razširjanje in izmenjevanje elektronskih besedil
- TEI je kot osnovo vzel SGML, verzija TEI P4 (2002) pa je izražena v XML
- trenutna verzija je P5

Razširjenost TEI

- TEI je postal de-facto standard za izdelavo znanstvenih digitalnih izdaj, korpusov in, do neke mere, slovarjev
- TEI uporablja okoli 100 projektov, ki pokrivajo prek 30 jezikov
- BNC, MULTEXT-East, SVEZ-IJS, SDT, jaSlo, …

Priporočila TEI

Priporočila TEI so sestavljena iz

- priročnika (tiskan pribl. 1200 strani)
- modulov (naborov oznak ~DTDjev)

prosto dostopno na straneh konzorcija TEI : priporočila, orodja, učbeniki, seznam projektov, …

Zaključek

Spoznali smo:

- standarde, malo zgodovine, kaj so, zakaj so dobri in kdo jih objavlja
- standarde vezane na zapis podatkov, predvsem XML
- kaj je dobro napisan dokument XML
- kaj so DTD-ji in pravilni dokumenti XML
- in še malo o TEI

Standardizacija oblikoslovnih oznak

- Oblikoslovne oznake so namenjene označevanju oblikoslovnih lastnosti posameznih besed
- za slovenski jezik sta najbolj poznana dva nabora:
  - (Nova)Beseda
  - Fida(PLUS)

Oznake ZRC SAZU

- uporabljene v oblikoslovno označenem korpusu ZRC SAZU, v programu EVA in spletnem vmesniku, c.f. http://bos.zrc-sazu.si/oblikoslovno_oznacevanje.html
- oznake osnovane na slovenski slovnici
- vsaki oblikoslovni lastnosti pripada ena ali več črk ali številk:
  - **Sže2**- samostalnik ženskega spola, ednina, rodilnik
  - **A** - prislov
  - **Gcp** - glagol, tretja oseba, množina
  - **Pme1i** - pridevnik, moški spol, ednina, imenovalnik, določna oblika
- nabor vseh oznak ni podan

Oznake FidaPLUS

- izhajajo iz projektov EU:
  - EAGLES: Expert Advisory Group on Langage Engineering Standards (1993-1996)
  - MULTEXT: Multilingual Text Corpora and Tools (1993-1996)
  - MULTEXT-East: MULTEXT for Central and Eastern European Languages (1995-1997)
    - Slovenija: IJS, Amebis
  - MULTEXT-East V2 (2002), V3 (2004)

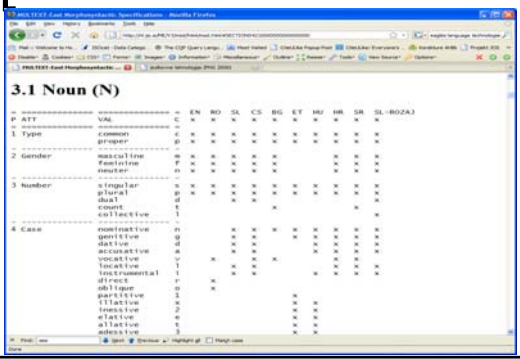
MULTEXT-East

- prosto dostopni večjezični oblikoslovni viri za namene jezikovnih tehnologij:
- oblikoslovne specifikacije
- oblikoslovni leksikon
- oblikoslovno označen, standardno zapisan, vzporedni korpus: “1984”
- c.f. http://nl.ijs.si/ME/

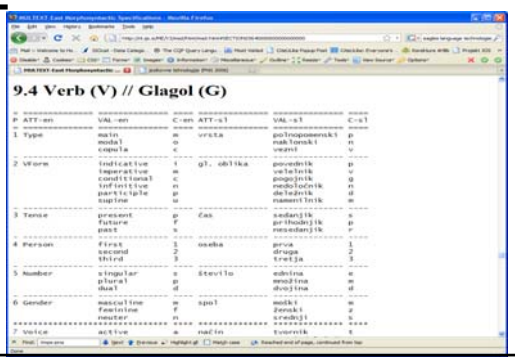
Oblikoslovne specifikacije

- definirajo besedne vrste, njihove oblikoslovne lastnosti, in vrednosti
- in za katere jezike so dovoljene
- določijo preslikavo v oblikoslovne oznake npr. `Category=Noun, Type=common, Gendar=masculine, Number=singular, Case=nominative` ↔ `Ncmsn`
- preslikava je avtomatično izvedljiva
- če nek atribut ni definiran, se uporabi pomišljaj:
  `Vmip1s-n` (čakam)
  `Vmps-pna` (čakala)

Primer večjezične tabele za samostalnik (kaj je Npfdv?)



Tabele za slovenščino: dvojezičnost



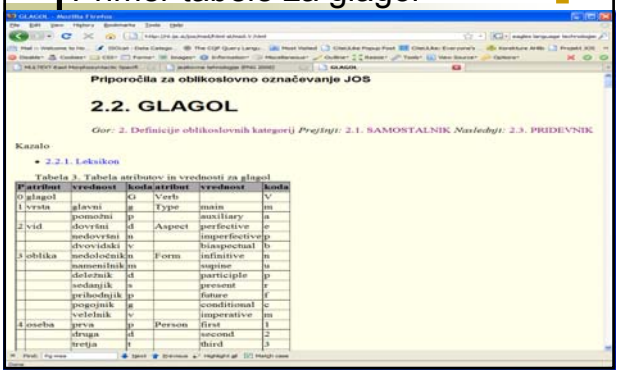
Fida(PLUS)

- Fida prevzela (slovenske) oznake MULTEXT-East
  Pg-msa----a-y = Zc-met----p-d
- skoraj enake oznake nato uporabljene tudi za FidaPLUS
- namesto "-" se uporablja "x":
  vsega `Zc-met----p-d`
  `Zcxmetxxxxpxd`

JOS

- projekt "Jezikoslovno označevanje slovenščine", http://nl.ijs.si/jos/
- izdelava novih oblikoslovnih specifikacij in ročno označenih korpus
- specifikacije sedaj napisane v XML-TEI
- spremembe pri vrstnem redi lastnosti, imenih lastnosti in katere oznake so pripisane katerim besednim oblikam

Primer tabele za glagol



Priporočila za oblikoslovno označevanje JOS

2.2. GLAGOL

Gor.: 2. Definicije oblikoslovnih kategorij *Prejšnji:* 2.1. SAMOSTALNIK *Naslednji:* 2.3. PRIDEVNIK

Kazalo

- 2.2.1. Leksikon

Tabela 3. Tabela atributov in vrednosti za glagol

P	atribut	vrednost	koda	atribut	vrednost	koda
0	glagol		G	Verb		V
1	vrsta	glavni	g	Type	main	m
		pomožni	p		auxiliary	a
2	vid	dovršni	d	Aspect	perfective	e
		nedovršni	n		imperfective	p
		dvovidski	v		biaspectual	b
3	oblika	nedoločnik	n	Form	infinitive	n
		namenilnik	m		supine	u
		deležnik	d		participle	p
		sedanjik	s		present	r
		prihodnjik	p		future	f
		pogojnik	g		conditional	c
		velelnik	v		imperative	m
4	oseba	prva	p	Person	first	1
		druga	d		second	2
		tretja	t		third	3

Primer nabora oznak

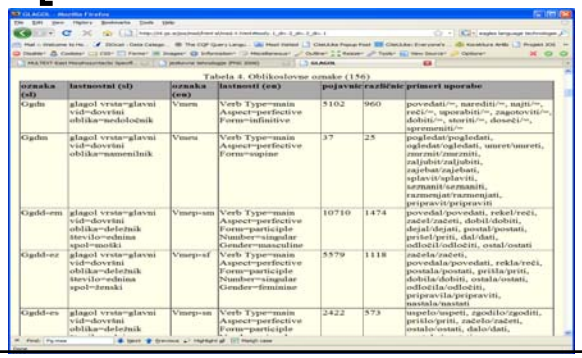


Tabela 4. Oblikoslovne oznake (156)

oznaka (sl)	lastnostni (sl)	oznaka (en)	lastnosti (en)	pojavnic	različnic	primeri uporabe
Ggdn	glagol vrsta=glavni vid=dovršni oblika=nedoločnik	Vmern	Verb Type=main Aspect=perfective Form=infinitive	5102	960	povedati/=, narediti/=, najti/=, reči/=, uporabiti/=, zagotoviti/=, dobiti/=, storiti/=, doseči/=, spremeniti/=
Ggdm	glagol vrsta=glavni vid=dovršni oblika=namenilnik	Vmeu	Verb Type=main Aspect=perfective Form=supine	37	25	pogledat/pogledati, ogledat/ogledati, umret/umreti, zmrzmit/zmrzniti, zaljubit/zaljubiti, zajebat/zajebati, splavit/splaviti, sezmanit/sezmasiti, razmenjat/razmenjati, pripravit/pripraviti
Ggdd-em	glagol vrsta=glavni vid=dovršni oblika=deležnik število=ednina spol=moški	Vmep-sm	Verb Type=main Aspect=perfective Form=participle Number=singular Gender=masculine	10710	1474	povedal/povedati, rekel/reči, začel/začeti, dobil/dobiti, dejal/dejati, postal/postati, prišel/priti, dal/dati, odločil/odločiti, ostal/ostati
Ggdd-ez	glagol vrsta=glavni vid=dovršni oblika=deležnik število=ednina spol=ženski	Vmep-sf	Verb Type=main Aspect=perfective Form=participle Number=singular Gender=feminine	5579	1118	zazela/začeti, povedala/povedati, rekla/reči, postala/postati, prišla/priti, dobila/dobiti, ostala/ostati, odločila/odločiti, pripravila/pripraviti, nastala/nastati
Ggdd-es	glagol vrsta=glavni vid=dovršni oblika=deležnik	Vmep-sn	Verb Type=main Aspect=perfective Form=participle	2422	573	uspelo/uspeti, zgodilo/zgoditi, prišlo/priti, začelo/začeti, ostalo/ostati, dalo/dati,

JOS oblikoslovne oznake

- število oblikokoslovnih oznak: 1902
- vendar zelo neenakomerno razporejene:
  v korpusu s 100.000 besed je samo 1064 oznak

ime	koda	ime	koda	št. atributov
samostalnik	S	Noun	N	5
glagol	G	Verb	V	7
pridevnik	P	Adjective	A	6
prislov	R	Adverb	R	2
zaimek	Z	Pronoun	P	8
števnik	K	Numeral	M	6
predlog	D	Preposition	S	1
veznik	V	Conjunction	C	1
členek	L	Particle	Q	0
medmet	M	Interjection	I	0
okrajšava	O	Abbreviation	Y	0
neuvrščeno	N	Residual	X	1