Slikovni priročnik za uporabo spletnega iskalnika po korpusu govorjene slovenščine

gos

KAJ JE KORPUS GOVORJENE SLOVENŠČINE GOS?

1. Elektronsko urejena zbirka različnih tipov govora v slovenskem jeziku

Korpus GOS vsebuje okrog 110 ur govora v slovenskem jeziku (oz. 1 mio. besed) v najrazličnejših situacijah, kjer vsak dan uporabljamo slovenski jezik: radio, televizija, šole in fakultete, zasebni pogovori, nezasebni pogovori na različnih sestankih, ob kupovanju stvari, opravljanju raznih storitev, svetovanju, posredovanju informacij...

TIP DISKURZA	KANAL	GOVORNI DOGODEK
JAVNI INFORMATIVNO-IZOBRAŽEVALNI	televizija	novinarski prispevek
		moderirani pogovor
	radio	moderirani program
		moderirani pogovor
	osebni stik	osnovnošolska učna ura
		srednješolska učna ura
		tečaj
		fakultetno predavanje
		javno predavanje
JAVNI RAZVEDRILNI	televizija	moderirani pogovor
		moderirana oddaja
		resničnostni šov
		športni prenos
	radio	moderirani program
		moderirani pogovor
NEJAVNI NEZASEBNI	osebni stik	formalni delovni sestanki
		neformalni delovni sestanki
		konzultacija na fakulteti
		storitev
	telefon	storitev
NEJAVNI ZASEBNI	osebni stik	pogovor v družini
		pogovor med prijatelji/znanci
	telefon	pogovor v družini
		pogovor med prijatelji/znanci

Ob tem so skladno pokrite vse regije, kjer se govori slovenski jezik, tudi v slovenskem zamejstvu v Italiji, Avstriji in Madžarski, ter vsi sloji prebivalstva glede na demografske značilnosti: spol, starost, izobrazba, prvi jezik.

Natančnejša sestava korpusa GOS je opisana pod jezičkom O korpusu na spletni strani www.korpus-gos.net.

2. Vsebuje posnetke in transkripcije pogovorov

Korpus GOS je nastal tako, da so bile najprej z dovoljenjem govorcev posnete različne vsakdanje (po)govorne situacije.

Na podlagi posnetkov so bile narejene do besede natančne transkripcije v t. i. pogovornem zapisu, ki upošteva značilnosti govorjenega jezika, kot so redukcije in druge značilnosti izgovorjave po slovenskih regijah.

A: po se pa le uč
B: nnn ja sej se bom
A: z() ja
B: ja dva v pondelk po pa v sredo pa v petek

V tretjem koraku je bil pogovornemu zapisu dodan t. i. standardizirani zapis govora, ki pripiše vsaki besedi v pogovornem zapisu standardizirano obliko te besede.

A: pol se pa le uči
B: nnn ja saj se bom
A: z() ja
B: ja dva v ponedeljek pol pa v sredo pa v petek

Originalni posnetki so ostali vključeni v korpus in povezani s transkripcijami, vendar so zaradi varovanja govorcev ustrezno anonimizirani, slišimo pa lahko samo posamezne odseke.

3. Vsebuje osnovne jezikoslovne oznake

Potem ko so bile izdelane vse transkripcije posnetkov, so bile k standardiziranemu zapisu avtomatsko dodane osnovne jezikoslovne oznake, ki so običajno vključene v jezikovne korpuse: podatki o osnovni obliki (tj. lema) in o oblikoslovnih lastnostih besede (tj. spol, sklon, število, oseba...). Zaradi avtomatskega dodajanja te informacije niso stoodstotno natančne.

```
<valtem ident="Sometn">
<desc>
besedna_vrsta=samostalnik
vrsta=občno_ime
spol=moški
število=ednina
sklon=tožilnik
živost=ne</desc>
```

4. Vsebuje osnovne podatke o posneti situaciji in govorcih

Poleg zapisa govora vsebujejo transkripcije tudi osnovne podatke o posnetkih: tip govora (javni, zasebni...), kanal snemanja (radio, TV, telefon, osebni stik), regijo in leto snemanja, število govorcev, kratek opis govornega dogodka...

Nejavni zasebni diskurz

osebni stik
pogovor v družini
regija: KP
12.07.2003 15:00
8 udeležencev
vir: terenski posnetek
Opis: Družinsko kosilo in druženje s sorodniki; splošni pogovor, obujanje spominov itd.

Pri vseh zasebnih posnetkih, v veliki meri tudi pri nezasebnih in deloma pri nekaterih javnih (zlasti predavanja ipd.) posnetkih so dodani podatki o govorcih: spol, starost, izobrazba, regionalna pripadnost in prvi jezik.

Govorec

spol: ženski
starost: 35 do 59
izobrazba: višja šola
regija: KP, GO
prvi jezik: slovenščina

5. Je zakodiran v standardu XML in v skladu s priporočili TEI

Tekstovni del korpusa je bil v zadnjem koraku zapisan v računalniškem standardu XML in v skladu s priporočili TEI (Text Encoding Iniciative) ter je v tej obliki na voljo zahtevnejšim uporabnikom. Ti si ga lahko snamejo s spletnih strani www.korpus-gos.net. Opis sheme XML je vključen v paket za prenos korpusa.

```
<?xml version="1.0" encoding="UTF-8" ?>
- <TEI xmlns="http://www.tei-c.org/ns/1.0"
xmlns:rng="http://relaxng.org/ns/structure/1.0"
xmlns:gos="http://nl.ijs.si/ssj/gos" xml:lang="en">
- <teiHeader>
- <fileDesc>
- <titleStmt>
<title xml:lang="en">The TEI Schema for GOS
speech corpus of Slovene</title>
<title xml:lang="sl">Shema TEI za govorni korpus
GOS</title>
<author>Tomaž Erjavec,
tomaz.erjavec@ijs.si</author>
</titleStmt>
```

6. Je prosto dostopen prek spletnega iskalnika na www.korpus-gos.net

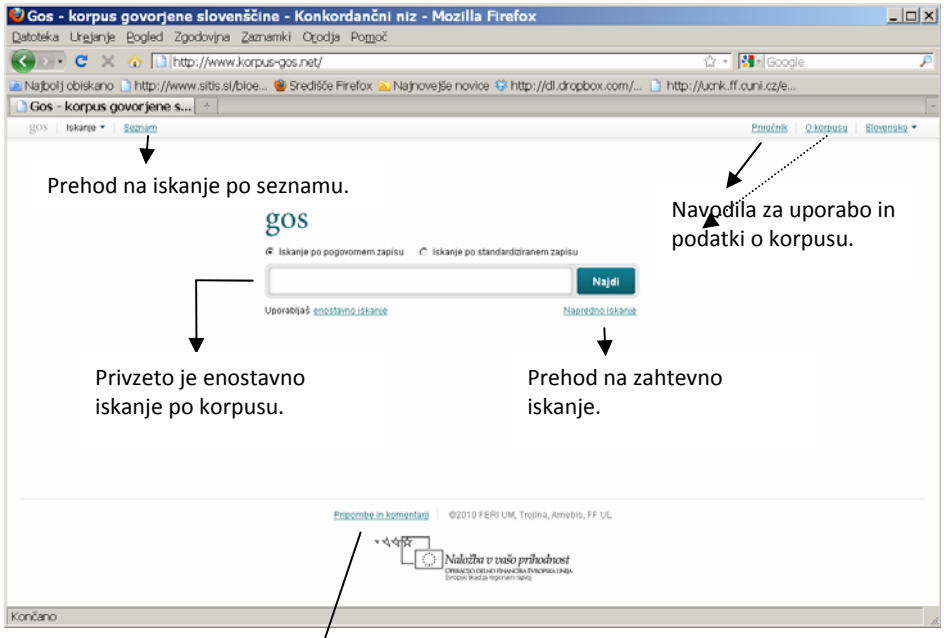
Po korpusu GOS lahko uporabniki iščejo s pomočjo izdelanega spletnega iskalnika, t. i. konkordančnika, ki je integriran na spletnih straneh www.korpus-gos.net. Konkordančnik GOS omogoča nastavljanje iskalnih pogojev po različnih ključih in zahtevah uporabnikov ter pri tem izkorišča podatke, vključene v transkripcije. V nadaljevanju vas vodnik po spletnem konkordančniku GOS vodi skozi različne funkcije iskanja po korpusu.

VODNIK PO SPLETNEM KONKORDANČNIKU GOS

Osnovno okno

V osnovnem oknu imamo na izbiro:

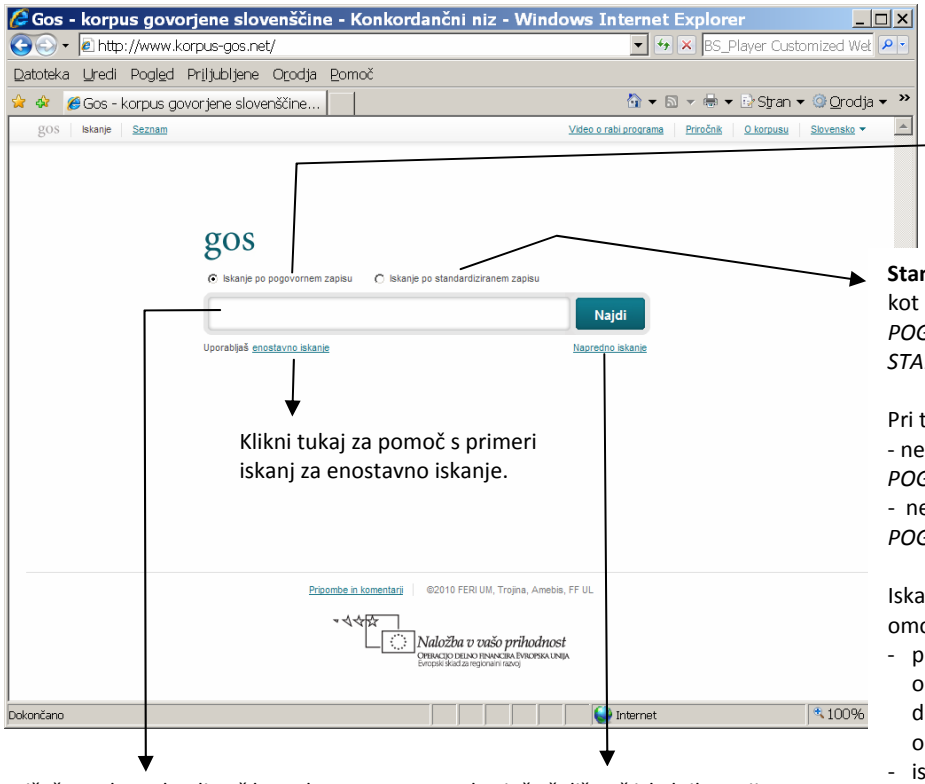
- privzeto nastavljeno enostavno iskanje s pomočjo iskalnega okna, ki vrne zadetke v kontekstu izjave (t. i. konkordance)
- prehod na napredno iskanje, ki prav tako vrne zadetke v kontekstu izjave (t. i. konkordance)
- prehod na iskanje po seznamu (zavihek Seznam), ki vrne zadetke v obliki seznama besed in oblik s podatki o pogostosti rabe
- prehod na priročnik, ki ponuja pisni opis iskalnih možnosti na spletni strani, video o rabi konkordančnika in prenos tega priročnika
- prehod na dodatne informacije o korpusu GOS



Enostavno iskanje

Osnovno iskalno okno ponuja privzeto **enostavno iskanje** po korpusu, ki omogoča iskanje na treh nivojih: (1) po pogovornem zapisu v korpusu, (2a) po avtomatsko določeni osnovni standardizirani obliki besed ali (2b) po standardiziranem zapisu v korpusu. Zadetki so vedno prikazani v pogovornem zapisu.

1. iskanje besede ali več besed, ločenih s presledkom, po zapisu govora, ki je imenovan **pogovorni zapis** (npr. *b reku da*) in pri katerem so zadetki vedno enaki iskanemu nizu (*b reku da*)
2. iskanje besede ali več besed, ločenih s presledkom, po zapisu govora, ki je imenovan **standardizirani zapis** in je razdeljen v dva načina iskanja:
   a. brez narekovajev: odkljukamo iskanje po standardiziranem zapisu, iskane besede pa lahko vpišemo v osnovni (*biti reči da*) ali v katerikoli drugi obliki (*bi rekel da*), zadetki so vedno, kot bi vpisali osnovno obliko (npr. *b reku da, je rekla da, sn reko da*...); osnovna oblika (lema) običajno sovpada s slovarsko obliko (npr. *biti, reči, da*)
   b. z narekovaji: odkljukamo iskanje po standardiziranem zapisu, besede vpišemo v obliki, kot jih pišemo (npr. *bi rekel mami*), in jih damo med narekovaje (*»bi rekel da«*); zadetki se prikažejo v pogovornem zapisu samo za iskane oblike (npr. *b reku da, bi reko da, bi rekel da*...)



Pogovorni zapis sledi principu, »zapiši, kot slišiš« (*tud, neki, mislm, prov, navm, najraj b vidu*...). Pri iskanju po njem iščemo vedno samo točno tako obliko iskane besede, kot jo vpišemo.

Standardizirani zapis sledi principu, »zapiši, kot pišemo«:

POG.:	*kok*	*boš*	*traku*	*zgubu*
STAN.:	*koliko*	*boš*	*traku*	*izgubil*

Pri tem pa:
- ne spreminja besed kot nosilcev pomena
  POG.: furež *STAN.: furež*, ne *koline*
- ne spreminja skladenjskih lastnosti besed:
  POG.: sva šle *STAN.: sva šle,* ne *sva šli*

Iskanje po standardiziranem zapisu omogoča:
- privzeto je iskanje po standardizirani osnovni obliki besede (če vpišemo *rekel*, dobimo vse zadetke, kjer je osnovna oblika besede *reči*)
- iskanje po standardiziranem zapisu, če uporabimo narekovaje (če vpišemo *»rekel«*, dobimo vse zadetke, kjer je standardizirani zapis *rekel*)

Vpišeš eno besedo ali več besed, ločenih s presledkom, brez narekovajev ali med narekovaji.

Izberi, če želiš več iskalnih opcij.

Za dodatne informacije o enostavnem iskanju glej zavihek Priročnik na spletni strani www.korpus-gos.net.

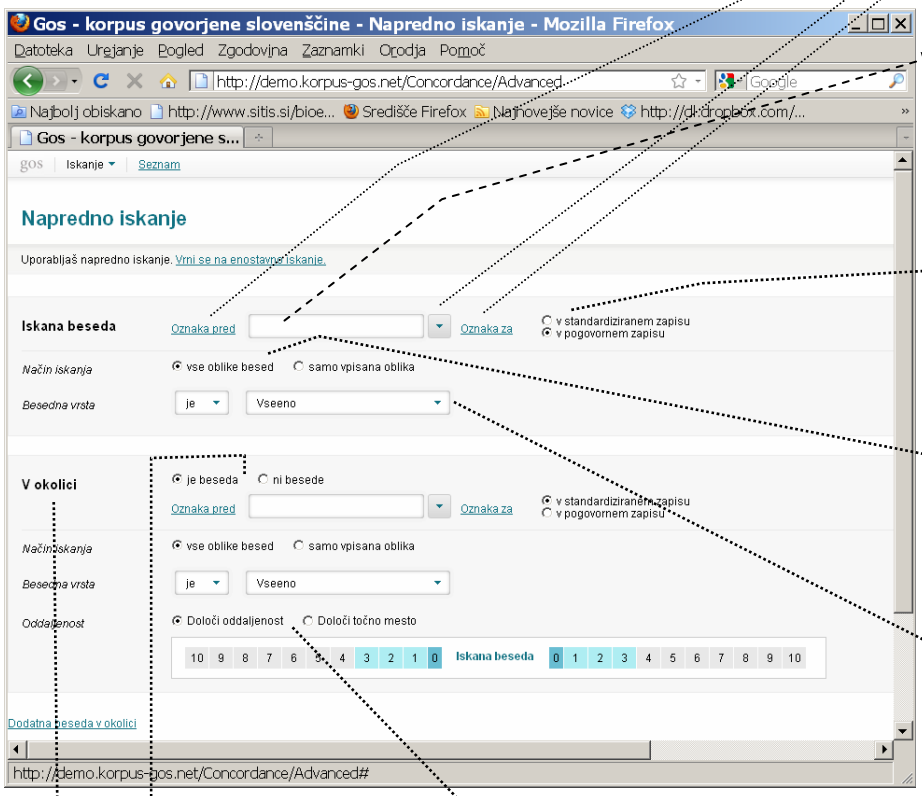
Napredno iskanje

Napredno iskanje poleg iskalnih opcij, ki jih omogoča že enostavno iskanje, to je:

- izbira, ali iščemo po pogovornem zapisu
- ali po standardiziranem zapisu (**način iskanja: samo vpisana oblika**)
- oz. osnovni obliki standardiziranega zapisa (**način iskanja: vse oblike besed**)

omogoča še:

- dodajanje posebnih oznak kot dodatnega iskalnega pogoja (kliknemo **oznaka pred** oz. **oznaka za** ali puščico za iskalnim okencem)
- določanje besedne vrste in oblikoslovnih lastnosti besede kot dodatnega iskalnega pogoja
- določanje okolice besede



Izbereš posebne oznake v transkripciji kot dodatne iskalne pogoje.

Vpišeš vedno samo eno besedo. Če iščeš dve besedi, klikni **beseda v okolici,** če iščeš več besed, pa zatem še **dodatna beseda v okolici**.

Izberi, ali želiš iskati po pogovornem ali standardiziranem zapisu govora. Glej razlago pod Enostavno iskanje.

Izberi, ali želiš iskati vse slovnične oblike iskane besede (tj. iskanje na nivoju osnovnih oblik besed) ali samo vpisano slovnično obliko (tj. iskanje na nivoju standardiziranega zapisa).

Po želji določi besedno vrsto iskane besede in pod **podrobnosti** oblikoslovne lastnosti iskane besede.

Izberi, če želiš iskati več besed naenkrat.

Določi, ali druga iskana beseda je ali ni v okolici prve.

Po želji določi, koliko je druga iskana besede od prve oddaljena levo ali desno v besedilu (**določi oddaljenost**) oz. koliko točno je oddaljena od prve iskane besede (**določi točno mesto**).

Za dodatne informacije o naprednem iskanju glej zavihek Priročnik na spletni strani www.korpus-gos.net.

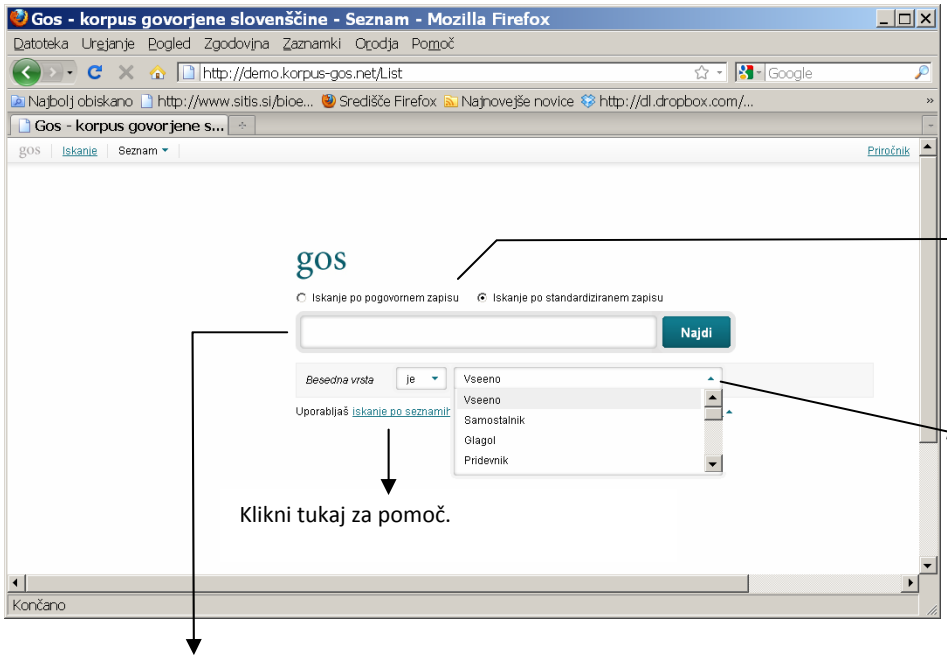
Seznam

Iskanje po seznamu (zavihek Seznam) se razlikuje od enostavnega in naprednega iskanja pod zavihkom Iskanje po tem, da ne prikaže zadetkov v obliki vrstic s kontekstom, kjer se je iskani niz pojavil, ampak v obliki seznama s podatki o oblikah in pogostosti pojavitve. Poleg iskalnih opcij, ki jih omogoča že napredno iskanje, to je:

- izbira, ali iščemo po pogovornem zapisu
- ali po standardiziranem zapisu (med narekovaji ali brez)
- ter določanje besedne vrste in oblikoslovnih lastnostih iskane besede

omogoča še:

- iskanje z nadomestnima znakoma:
  - znak * nadomešča več znakov
  - znak ? nadomešča en znak



Izberi, ali želiš iskati po pogovornem ali standardiziranem zapisu govora. Glej razlago pod Enostavno iskanje.

Po želji določi besedno vrsto iskane besede in pod **podrobnosti** oblikoslovne lastnosti iskane besede.

Vpišeš samo eno besedo. Uporabiš lahko znaka:

- * nadomešča več znakov
- ? nadomešča en znak

Brez narekovajev: iščeš vse oblike iskane besede.
Med narekovaji: iščeš samo vpisano obliko besede.

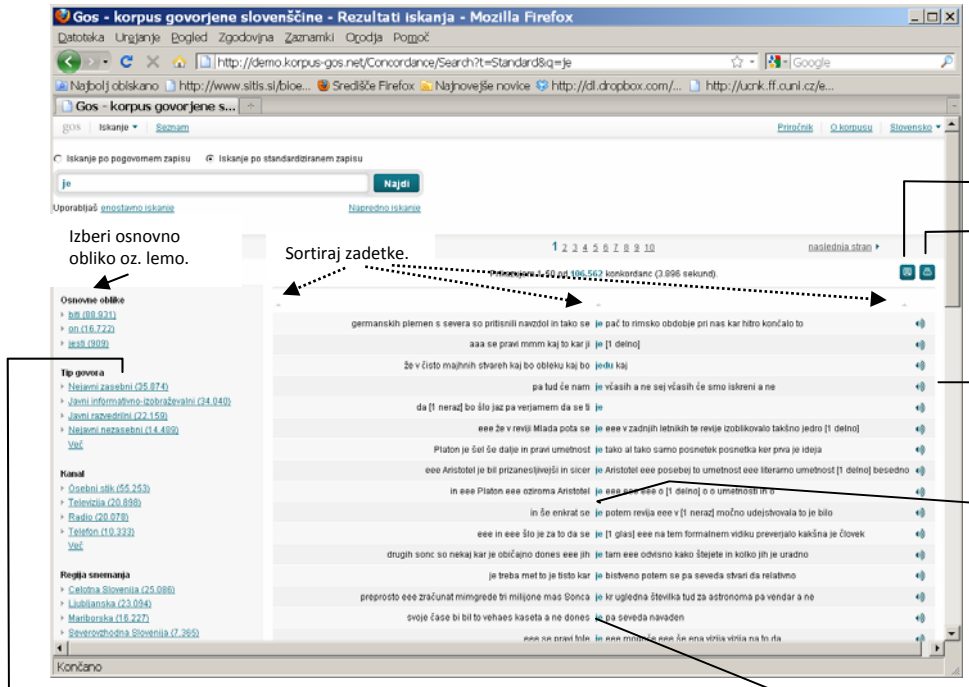
Za dodatne informacije o iskanju po seznamu glej zavihek Priročnik na spletni strani www.korpus-gos.net.

Zadetki pri enostavnem in naprednem iskanju

Zadetki se prikažejo VEDNO V POGOVORNEM ZAPISU.

Ko se prikažejo zadetki, je na voljo več možnosti:

- zadetke lahko s pomočjo izbir levo od zadetkov filtriraš glede na lastnosti pogovora in govorcev
- posamezno izjavo, v kateri je zadetek, lahko s klikom na zvočnik desno od zadetka poslušaš
- zadetke lahko sortiraš s puščicami nad njimi
- zadetke lahko natisneš ali izvoziš s pomočjo ikon desno nad zadetki
- o posameznem zadetku izveš dodatne informacije, če klikneš osrednjo besedo v zadetku



Zadetke lahko filtriraš, tako da klikneš eno od navedenih izbir in s tem izbereš samo tisto ali pa klikneš **več** in odkljukaš več izbir hkrati.

V zgornjem delu okna so izbire, ki se nanašajo na lastnosti pogovora, spodaj (uporabi drsnik okna) sledijo izbire, ki se nanašajo na lastnosti govorca.

Ko se postaviš z miško na zadetek, se pokaže podatek o tipu govora:
JI – javni informativni
JR – javni razvedrilni
NN – nejavni nezasebni
NZ – zasebni

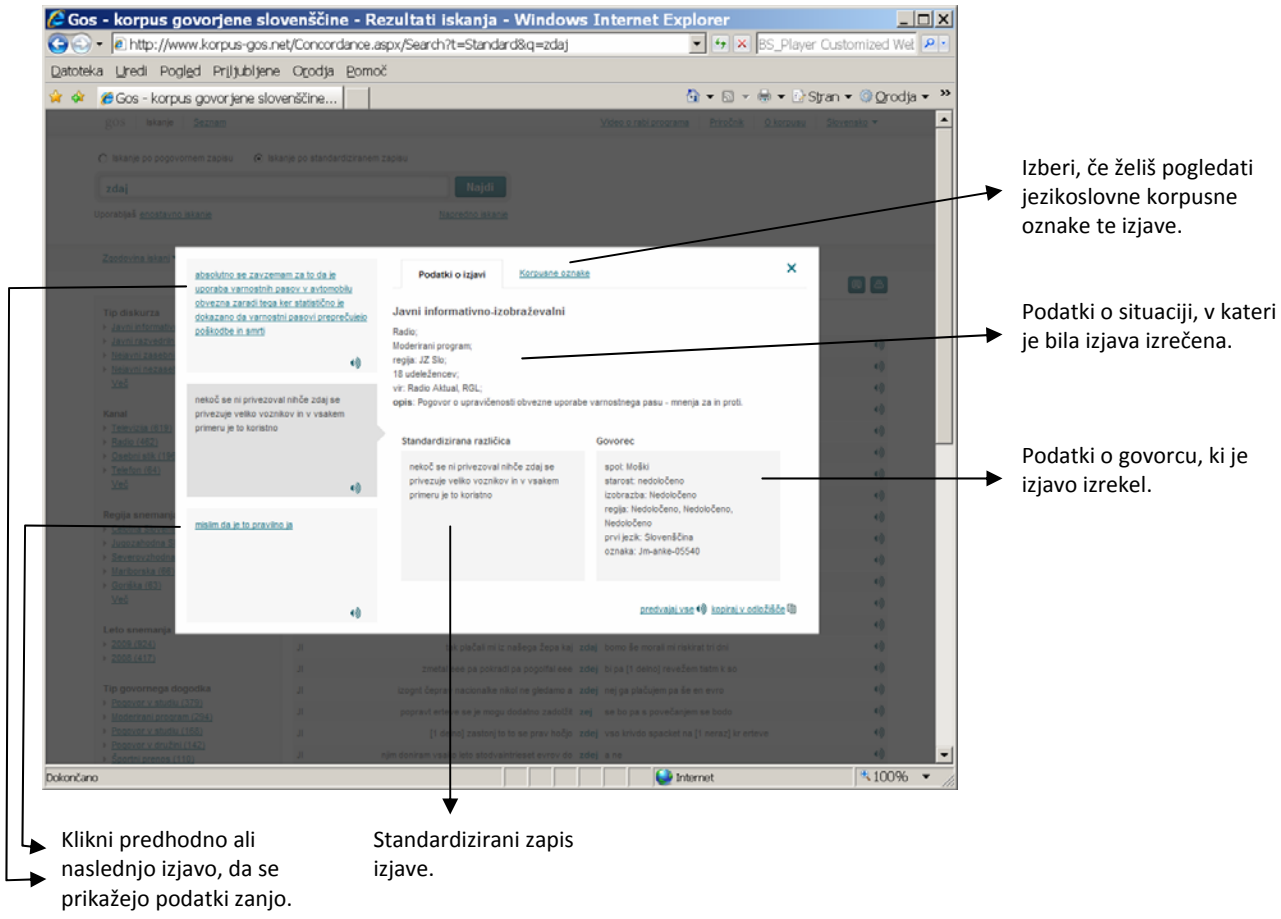
Za dodatne informacije o zadetkih glej zavihek Priročnik na spletni strani www.korpus-gos.net.

Več informacij o zadetku pri enostavnem in naprednem iskanju

Če klikneš osrednjo (označeno) besedo v zadetku, se prikaže okno z več informacijami o tem zadetku:

- predhodna izjava in naslednja izjava
- možnost poslušanja vsake posamezne izjave ali vseh izjav skupaj
- podatki o situaciji, v kateri je bila posamezna izjava izrečena
- podatki o govorcu, ki je izrekel posamezno izjavo
- standardizirani zapis posamezne izjave
- pod **korpusne oznake** si ogledaš dodane jezikoslovne podatke o izjavi

Prikazane podatke lahko kopiraš v odložišče.



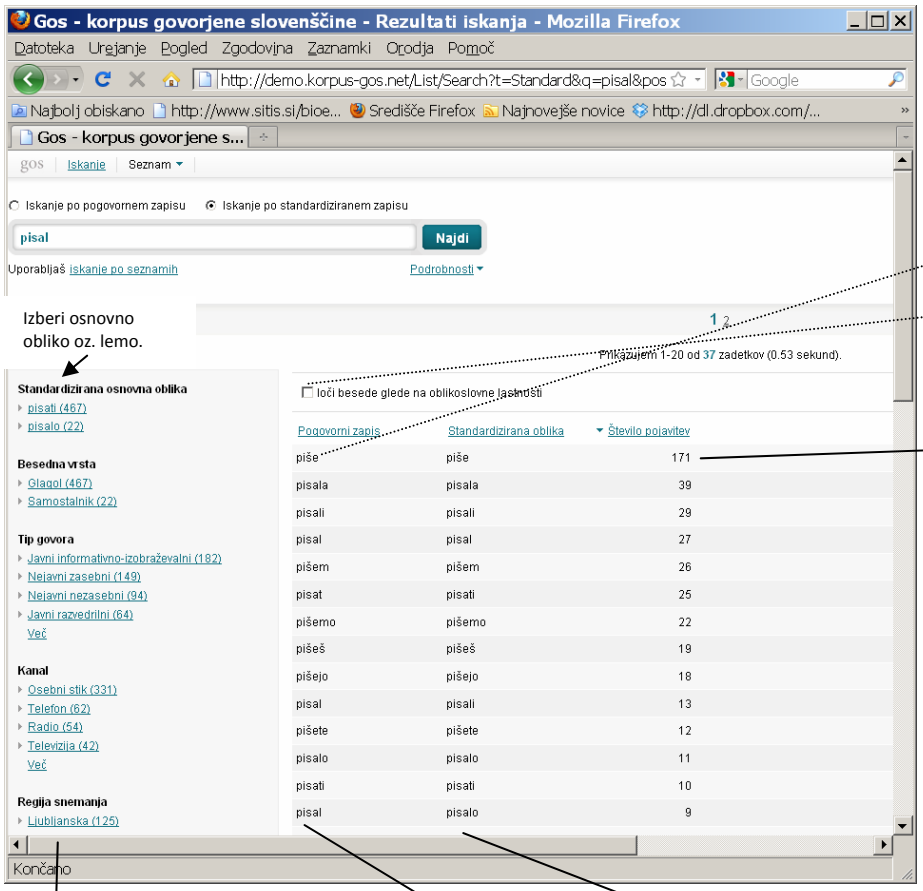
Zadetki pri iskanju po seznamu

Pri iskanju po seznamu se zadetki prikažejo v obliki seznama besed v več stolpcih, ki vsebujejo:

- pogovorni zapis
- standardizirani zapis (stolpec standardizirana oblika)
- število zadetkov za vsako posamezno kategorijo zadetkov

Zadetke lahko:

- filtriraš
- pogledaš vse pojavitve izbranega zadetka v kontekstu (prehod na prikaz rezultatov v enaki obliki kot pri enostavnem in naprednem iskanju)
- pogledaš oblikoslovne lastnosti izbranega zadetka
- shraniš
- natisneš



Izberi zadetek v stolpcu **pogovorni zapis** in odkljukaj ta kvadratek, da se prikažejo oblikoslovne lastnosti besede.

Klikni podatek o številu pojavitev, da se prikažejo zadetki te vrstice v kontekstu, tako kot pri enostavnem in naprednem iskanju.

Zadetke lahko filtriraš, tako da klikneš eno od navedenih izbir in s tem izbereš samo tisto ali pa klikneš **več** in odkljukaš več izbir hkrati.

V zgornjem delu okna so izbire, ki se nanašajo na lastnosti pogovora, spodaj (uporabi drsnik okna) sledijo izbire, ki se nanašajo na lastnosti govorca.

Podatek, kako je beseda izgovorjena v govoru.

Podatek, kako je beseda zapisana v standardiziranem zapisu.