

Jezikoslovni viri starejšega slovenskega jezika IMP: digitalna knjižnica, korpus, slovar

Tomaž Erjavec

Odsek za tehnologije znanja
Institut »Jožef Stefan«

Pregled

1. ozadje
2. digitalna knjižnica IMP
3. korpus goo300k
4. slovar IMP
5. zaključki

Motivacija

Jezikovne tehnologije:

- izboljšanje avtomatskega prepoznavanja besedil (OCR)
- iskanje po polnem besedilu digitalnih knjižnic
- posodabljanje starejših besedil
- (postaranje sodobnih besedil)

Humanistika:

- digitalna knjižnica: o čem so včasih pisali
- korpus: kako so pisali in kako se je jezik spreminjal
- slovar: kaj pomenijo zastarele besede in kako so bile zapisane

Faksimile

Poflednjizh je Bog zhloveka stvaril. Stvar-
nik pogleda vse, kar je naredil, ino vse je bilo
prav prav dobro.

Ljubi otrozi! tukaj na nebu ino na semlji
lahko vidimo, kako je Bog mogozhen, dobro-
tljiv ino moder, ako ravno njega — nafhiga pre-
ljubiga Boga — she ne moremo viditi. Njegove
prelepe stvari velikokrat f' svetim veseljem pregle-
dujmo, ino veselimo se, de imamo tako dobrot-
ljiviga Boga.

OCR:

napake

starinske oblike

Poflednjizh je Bog zhloveka ItvariL Stvar-
nik pogleda vfe, kar je naredil ^ ino vfe je bilo
prav prav dobro.

Ljubi otrozi! tukaj na nebu ino na semlji
lahko vidimo, kako je Bog mogozhen, dobro-
ti jiv ino moder, ako ravno njega — nafhiga pre-
ljubiga Boga — ■ fhe ne moremo viditi. Njegove
prelepe ftvari velikokrat P fvetim vefeljem pregle-
dujmo, ino vefelimo fe, de imamo tako dobrot-
ljiviga Boga.

Problemi

- pomanjkanje ročno pregledanih (in stavljenih) transkripcij
- tokenizacija
 - *žnjo* → *z njo*
 - *po noči* → *ponoči*
- variabilnost:
 - pregibanje:
ljubezen ← *ljubezni, ljubeznijo*
 - starinske oblike:
ljubezen ← *lubesen, ljubesen, lubefn, ljubezin, ljubesin*
 - oboje:
ljubezen ← *ljubezni, ljubesni, lubesen, ljubesen, lubesni, lubefn, ljubeznijo, ljubezin, lubefne, lubefni, lubesne, ljubesnijo, ljubesin*
- zastarele besede:
 - *zajhen* / *cajhen* / *znamenje*

1. projekt: **AHLib** (2004–08)



- projekt „*Deutsch-slowenische/kroatische Übersetzung 1848–1918*“
- avstrijska akademija znanosti in Univerza Karl-Franzens v Gradcu
akad. prof. dr. Erich Prunč
- izdelava korpusa knjig, prevedenih iz nemščine
- 300+ faksimilov, 91 redigiranih transkripcij, 9 lematiziranih
- IJS: spletni vmesnik za redakcijo in lematizacijo
 - izdelana osnovna metodologija označevanja
- dostopnost:
 - UKF Gradec: konkordančnik [GAMS](#), digitalna knjižnica [TRADOK](#)
 - IJS: <http://nl.ijs.si/ahlib/>

ERJAVEC, Tomaž. An Architecture for Editing Complex Digital Documents. V: *Proceedings of INFuture'07 "Digital Information and Heritage"*. Zagreb, Faculty of Arts, Croatia, 2007, str. 105-114.

ERJAVEC, Tomaž. Slovenska prevodna književnost 1848-1918 : digitalna knjižnica in korpus AHLIB. V: *Meddisciplinarnost v slovenistiki*, (Obdobja, Simpozij, = Symposium, 30), 2011, str. 113-120

2. in 3.



EU IP IMPACT (razširjeni projekt 2010–2011)

- boljši OCR za starejša besedila + boljše iskanje po besedilih digitalnih knjižnic
- NUK (Alenka Kavčič Čolić, Ines Vodopivec): 5.000 strani besedil
- IJS: izdelava računalniškega leksikona

Nagrada Google (2011 + 2012)

- izdelava računalniških jezikoslovnih modelov za starejšo slovenščino
- ZRC SAZU (Matija Ogrin): transkripcije starih besedil
- IJS: referenčni korpus starejše slovenščine
- projekt Wikivir „Slovenska leposlovna klasika“ (Miran Hladnik) podpora za popravke transkripcij, izdelava konverzije Wiki v XML/TEI (Domen Kermc)

ERJAVEC, Tomaž, RINGLSTETTER, Christoph, ŽORGA DULMIN, Maja, GOTSCHAREK, Annette. A lexicon for processing archaic language : the case of XIXth century Slovene. V: *First International Workshop on Lexical Resources, An ESSLLI 2011 Workshop*, Ljubljana, Slovenia - August 1-5, 2011. *An ESSLLI 2011 Workshop*, str. 24-30.

ERJAVEC, Tomaž, VODOPIVEC, Ines, KODRIČ, Maša. Izdelava korpusa starejših slovenskih besedil v okviru projekta IMPACT. V: *Meddisciplinarnost v slovenistiki*, (Obdobja, Simpozij, = Symposium, 30). 2011, str. 121-127

Vodila zbirke IMP

Razviti zbirko jezikovnih virov in orodij
za starejšo slovenščino,
ki:

- vsebuje čistopise in poravnane faksimile
- ima vsako besedo označeno z:
 1. sodobno ustreznico
 2. lemo (osnovno sodobno obliko)
 3. oblikoskladenjsko oznako
 4. za zastarele besede pa najbližje sodobne ustreznice
- je široko in prosto dostopna:
 - jezikoslovci: analitična orodja na spletu (slovar/konkordančnik/knjižnica)
 - jezikovni tehnologi: prenos (zapis po mednarodnih priporočilih: XML / TEI / MULTTEXT)

I. Digitalna knjižnica IMP

- osnova:
 - digitalni faksimili + ročno korigirane transkripcije
 - ni problemov z uporabo in nadaljnjim razširjanjem
- viri:
 - AHLib: slovenske knjige prevedene iz nemščine
 - NUK GTD: slovenske knjige, časopis KRN
 - ZRC SAZU: 3 vzorci (naj)starejših knjig
 - WIKIVIR: dokončana dela do 1918
- označevanje:
 - metapodatki (bibliografski, odgovornosti, dostopnost, oznake, ...)
 - prelom (poravnava faksimila z besedilom)
 - osnovna struktura (naslovi, odstavki)
 - ? kazala (razdelki, tabele, slike)
 - ? uredniški posegi (sic, tujejezično)

Zapis v XML / TEI

<body>

<pb n="[000]" facs="#FPG00396-000" xml:id="pb.001"/>

<p xml:id="p.1">»LJUDSKI ODER« II ZVEZEK</p>

<figure xml:id="figure.1">

<figDesc>Majhna bordura in dve črti </figDesc>

</figure>

<div type="level1" xml:id="div.1">

<head type="docTitle" xml:id="head.1">Krivoprisežnik. </head>

<p xml:id="p.2">Narodna igra s petjem v treh dejanjih (7 slikah). </p>

<p xml:id="p.3">Za slovenske ljudske odre priredil: Homunkulus. </p>

<p xml:id="p.4">Vse pravice pridržane. </p>

<figure xml:id="figure.2">

<figDesc>Majhna bordura </figDesc>

</figure>

<p xml:id="p.5">V Ljubljani 1911. </p>

<p xml:id="p.6">Založil »Ljudski oder«. </p>

<p xml:id="p.7">Natisnila Katoliška Tiskarna. </p>

<pb n="[1]" facs="#FPG00396-001" xml:id="pb.002"/>

<p xml:id="p.8">Važno! </p>

...

Na spletu (HTML)

AHLlib: Anzengruber, Ludwig. "Krivoprisežnik" (1911) - Mozilla Firefox

File Edit View History Bookmarks Tools Help

IMPACT GTD "Vezhna pratika od gos..." AHLlib: Anzengruber, Ludwig. "Krivop..."

nl.ijs.si/ahlib/dl/FPG_00396-1911.html#pb.001

Google Koledar zastarele besede SKE/Config/FullDoc... Up-conversion using X... Popword Editorial Manager® The Action ISO Standards Develop... IntelliText Corpus Que...

„LJUDSKI ODER“ II. ZVEZEK.

Krivoprisežnik.

Narodna igra s petjem v treh dejanjih (7 slikah).

Za slovenske ljudske odre priredil: Homunkulus.

Vse pravice pridržane.



V Ljubljani 1911.
Založil „Ljudski oder“.
Natisnila Katoliška Tiskarna.

[[0001]]

»LJUDSKI ODER« II ZVEZEK

Majhna bordura in dve črti

Krivoprisežnik.

Narodna igra s petjem v treh dejanjih (7 slikah).

Za slovenske ljudske odre priredil: Homunkulus.

Vse pravice pridržane.

Majhna bordura

V Ljubljani 1911.

Založil »Ljudski oder«.

Natisnila Katoliška Tiskarna.

[[1]]

Važno!

Prireditelje narodne igre Krivoprisežnik opozarjamo, da je vsaka uprizoritev navezana na tantijemo K 15 - Ta znesek naj se pošlje na Katoliško Bukvarno v Ljubljani, ki v tem oziru posreduje in ima knjižico tudi v prodaji.

Označeni znesek se mora vplačati vsaj tri dni pred uprizoritvijo, ker bi sicer prireditelj poleg tantijeme zapadel tudi visoki konvencionalni globi.

Važno!

Find: Match case

NEW ERM-DESIGN.COM

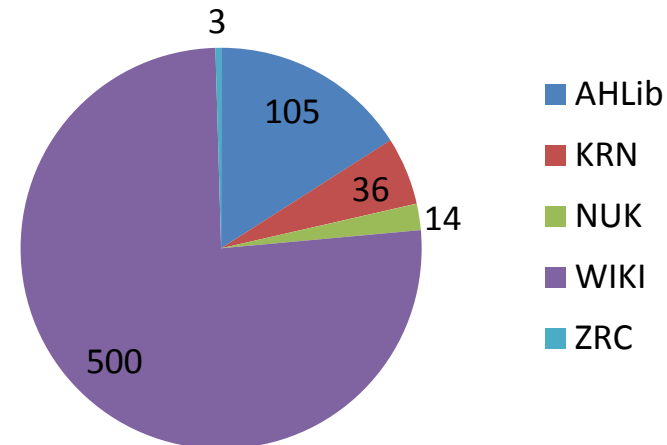
Najstarejša dela

1. Dalmatin, Jurij. *Biblia*. (1584) 8 str. [ZRC_00001]
2. Janez Svetokriški. *Sacrum promptuarium*. (1695) 27 str. [ZRC_00002]
3. Canisius, Petrus; Parhamer, Ignaz; Pohlin, Marko. *Ta male katechismus*. (1768) 130 str. [NUK_13105]
4. Sailer, Sebastian; Pohlin, Marko. *Marianske Kempensar, ali Dvoje bukuvze*. (1769) 188 str. [NUK_13130]
5. Hevenesi, Gabriel; Rupnik, Gašpar. *Ta Kristusovimu terpleinu posvezheni post*. (1773) 177 str. [NUK_10199]
6. Pohlin, Marko. *Branja, inu evangeliumi*. (1777) 374 str. [NUK_10187]
7. Japelj, Jurij. *Svetu pismu noviga testamenta*. (1784) 40 str. [ZRC_00003]
8. Rupnik, Gašpar. *Peisme od kershanskiga vuka po versti tega katechisma*. (1784) 57 str. [NUK_10095]
9. Pohlin, Marko. *Kratkozhasne uganke*. (1788) 46 str. [NUK_13057]
10. Breznik, Anton. *Vezhma pratika od gospodarstva*. (1789) 93 str. [NUK_13067]
11. Pohlin, Marko. *Abecedika ali Plateltof*. (1789) 173 str. [NUK_10299]
12. Linhart, Anton Tomaž. *Shupanova Mizka in Ta vesseli dan, ali: Matizhek se sheni*. (1790) 187 str. [NUKR10214]
13. Anton Janša. *Popolnoma podvuzhenje sa vsse zhebellarje*. (1792) 199 str. [WIKI00523]
14. Japelj, Jurij. *Pridige sa vse nedele skusi lejt*. (1794) 397 str. [NUK_10224]
15. Vodnik, Valentin. *Kuharske Bukve*. (1799) 207 str. [NUKR10221]
16. *To je tapravi inu tazieli Colemone-Shegen*. (1800) 273 str. [NUK_10029]
17. več avtorjev. *Osem, inu šestdeset sveteh pesm*. (1800) 247 str. [WIKI00004]
18. Alvian, Fecit Franciscus. *Praemium sa male otrozhizhke*. (1810) 10 str. [NUK_07541]
19. Lhomond, Charles Francois; Vodnik, Valentin. *Pozhetki gramatike*. (1811) 116 str. [NUK_10220]
20. Schmid, Christoph von. *Sgodbe svetiga pisma za mlade ljudi*. (1830) 196 str. [FPGN04488]
21. Janez Cigler. *Štiri poslednje reči*. (1831) 209 str. [WIKI00014]
22. Peter Dajnko. *Čelarstvo*. (1831) 245 str. [WIKI00299]

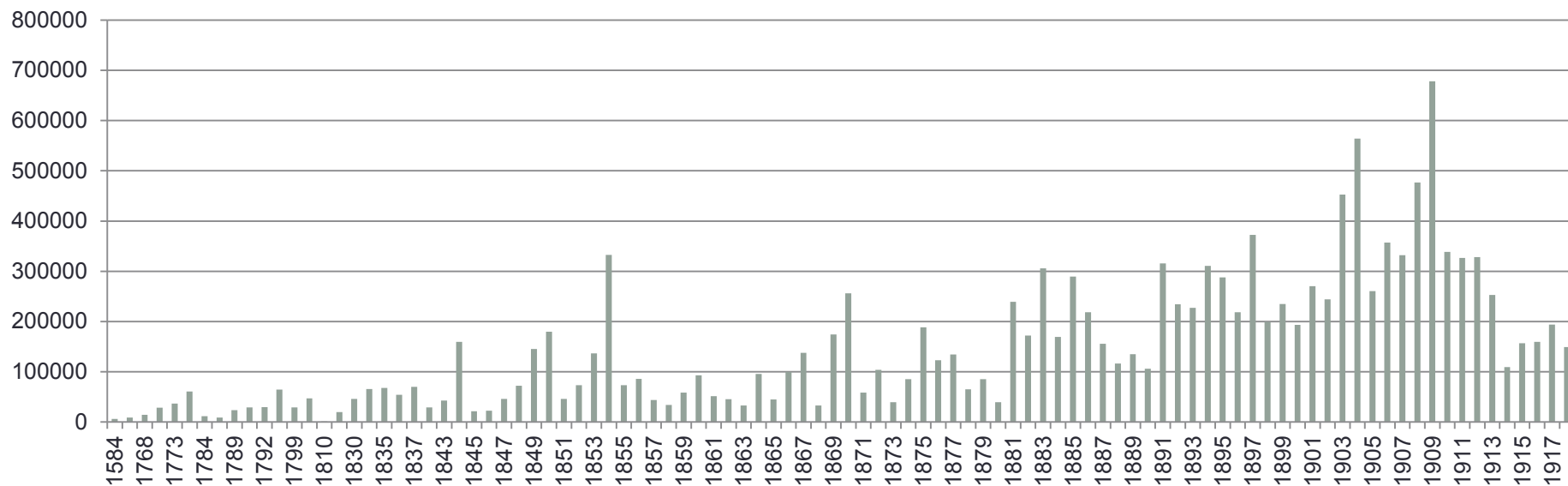
Obseg DK IMP

Vir	enot	strani	besed
AHLib	105	12.854	2.783.569
KRN	36	555	551.599
NUK	14	2.428	338.874
WIKI	500	30.296	10.761.913
ZRC	3	75	20.631
Σ	658	46.208	14.456.586

Število del / vir



Besed / leto



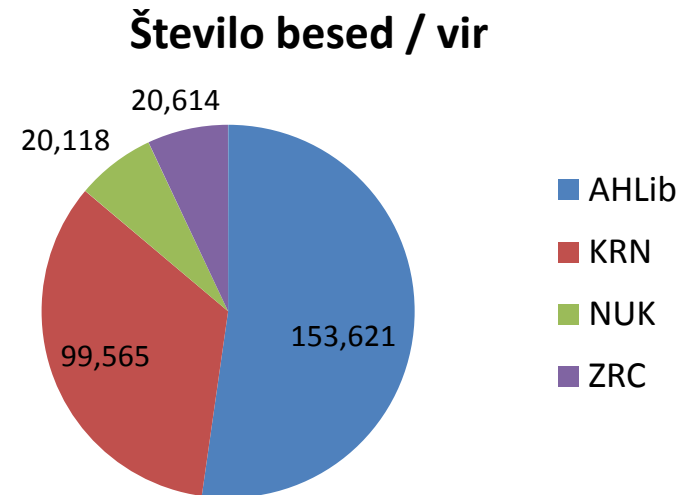
II. Referenčni korpus goo300k

- ročno označeni referenčni korpus starejše slovenščine
- vzorčen iz digitalne knjižnice IMP (iz 2010)
- dodatno popravljena transkripcija
- vsaka beseda v korpusu jezikoslovno označena
- posamezne strani povezane s faksimili in digitalno knjižnico

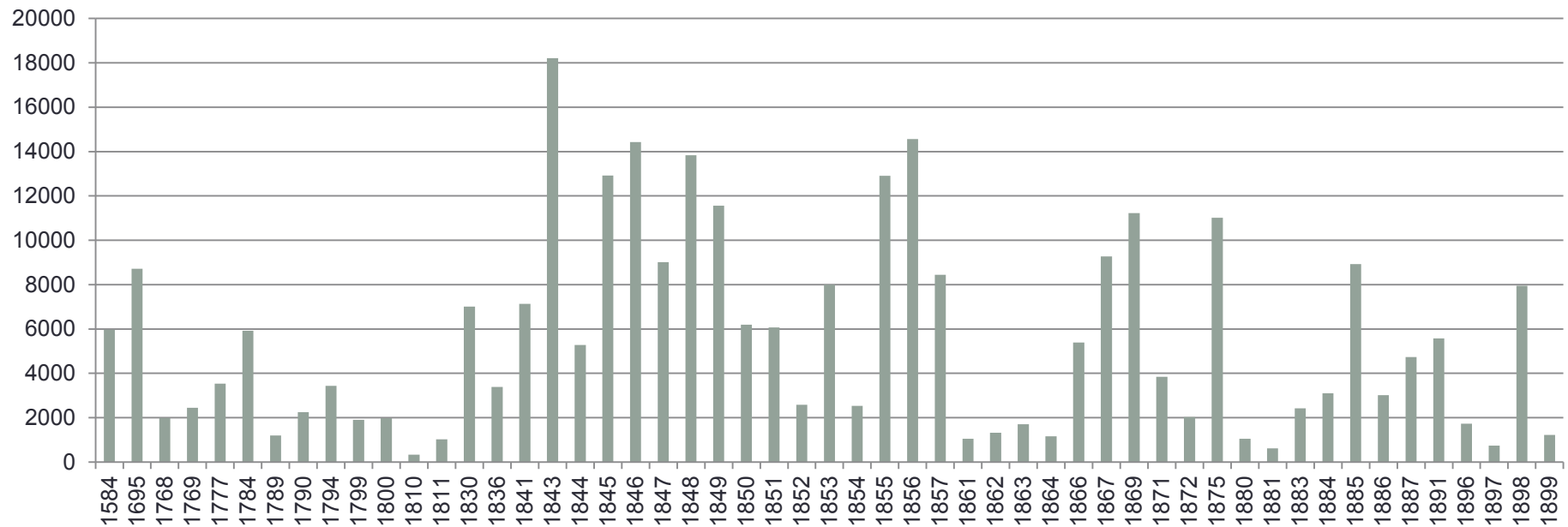
ERJAVEC, Tomaž. The goo300k corpus of historical Slovene. V: *8th International Conference on Language Resources and Evaluation*, 21-27 May 2012, Istanbul, Turkey. *LREC 2012: proceedings*. Istanbul: ELRA, 2012, str. 2257-2260.

Obseg goo300k

Vir	enot	strani	besed
AHLib	65	770	153.621
KRN	11	106	99.565
NUK	10	150	20.118
ZRC	3	74	20.614
Σ	89	1.100	293.918



Besed / leto



Izdelava korpusa

- INL CoBaLT: Corpus Based Lexicon Tool
- skupina anotatorjev
- priročnik, „kuharica“, FAQ, dopisni seznam, sestanki...
~ 2.000-4.000 ur

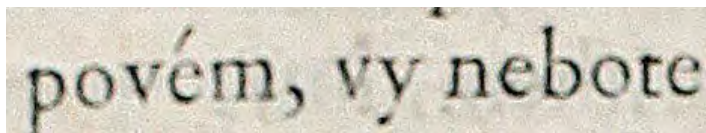
The screenshot displays the INL CoBaLT interface. At the top, there is a search bar with the query 'gospod%' and a 'Filter lemmata...' checkbox. Below the search bar, a table lists the word forms and their lemmas. The word 'gospodje' is highlighted in yellow.

Word form	Frequency	Lemma
gospod	5	gospod, <gospod>, Ncm
gospoda	2	gospod, <gospoda>, Ncm gospoda, <gospoda>, Ncf
gospodarji	1	gospodar, <gospodarji>, Ncm
gospodarstvo	1	gospodarstvo, <gospodarstvu>, Ncn
gospode	1	gospoda, <gospode>, Ncf
gospodje	2	gospod, <gospodje>, Ncm
gospodo	1	gospoda, <gospodo>, Ncf

Below the table, there is a section titled 'This word was viewed last by you at Wed 01 Feb 2012, 15:45:16'. It shows two example sentences from a corpus:

je pravil — Jak . Takí fo le ti **Gospodje** ; fladki , koker mēd , pak »
 . Naj bo tedej jih . Vifsoku vuzhēni **Gospodje** ! oní bodo lohka saftopili , de »

KENTNER, Tom, ERJAVEC, Tomaž, ŽORGA DULMIN, Maja, FIŠER, Darja. Lexicon construction and corpus annotation of historical language with CoBaLT editor. V: *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2012)*. Avignon: ACL, 2012.



Jezikoslovne oznake

povem	,	vy	nebote
povem		vi	ne+boste
povedati		ti	ne+biti
Vme		P	Q+Va

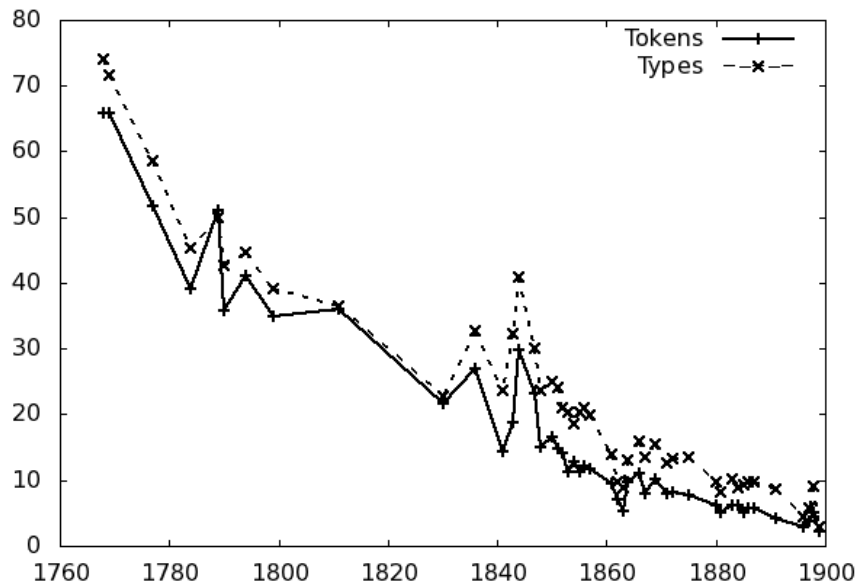
zahtevno:

mejni primeri, dvoumnosti, izjeme, ...

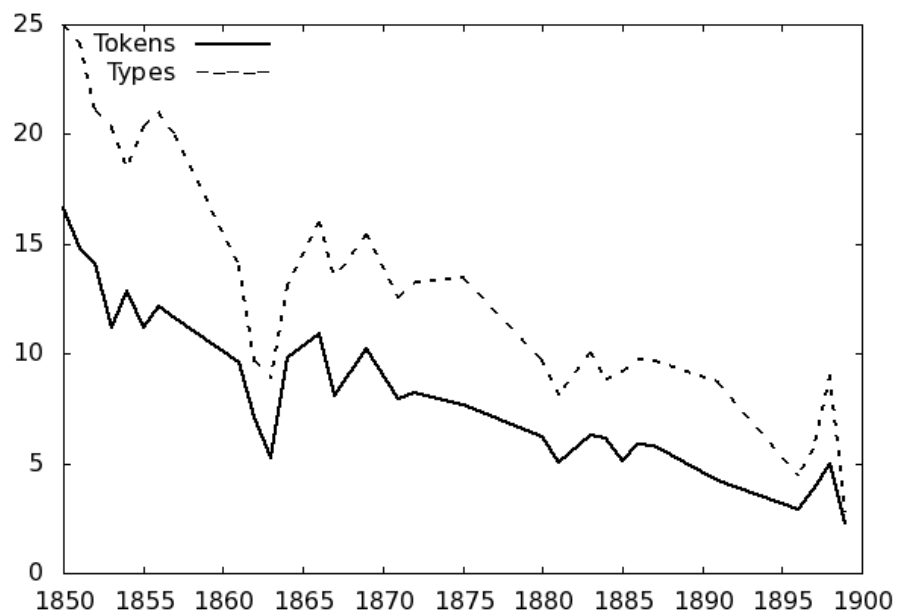
```

<w lemma="povedati" ana="#Vme">povem</w>
<pc>,</pc>
<c> </c>
<choice>
  <orig>
    <w>vy</w>
  </orig>
  <reg>
    <w lemma="ti" ana="#P">vi</w>
  </reg>
</choice>
<c> </c>
<choice>
  <orig>
    <w>nebote</w>
  </orig>
  <reg>
    <w lemma="ne" ana="#Q">ne</w>
    <w lemma="biti" ana="#Va">boste</w>
  </reg>
</choice>

```



Starejše / sodobne oblike



Konkordančnik noSketchEngine

Konkordance - Mozilla Firefox

File Edit View History Bookmarks Tools Help

AHLib digital library TOMAZ ERJAVEC [05023] Konkordance

nl.ijs.si/noske/si-ref.cgi/view?q=aword%2C[word%3D("%3F")ljubezen"]&lemma%3D("%3F")ljubezen";q=f;corpname=goo300k&refs=%3Dtext.display&iquery=ljubezen

Google Koledar MySQL database on Li... Git - Recording Chang... PP Funding > Latest Info ... IMPACT@JSI SKE SKE/Config/FullDoc... Editorial Manager@ The Action ISO Standards Develop... IntelliText Corpus Que...

NoSketchEngine

Išči v Pomoč

User: defaults Korpus: goo300k (starejša besedila, ročno označena) Išči ljubezen v goo300k (starejša besedila, roč

Konkordance
Seznam

Stran 1 od 7 Pojdi Naslednja | Zадnja

Oče naš (1885)
Kmetijske in rokodelske n... (1843)
Marianske Kempensar, ali ... (1769)
Blagomir puščavnik (1853)
Občno vzgojeslovje (1887)
Sgodbe svetiga pisma za m... (1830)
Branja, inu evangeliumi (1777)
Oče naš (1885)
Pridige sa vse nedele sku... (1794)
Deborah (1883)
Občno vzgojeslovje (1887)
Genovefa (1841)
Oče naš (1885)
Ta male katechismus (1768)
Roza Jelodvorska (1855)
Divica Orleanska (1848)
Divica Orleanska (1848)
Branja, inu evangeliumi (1777)
Sgodbe svetiga pisma za m... (1830)

zastran tega, kar se je zgodilo, in prejeto **ljubezen** v hvaležnim spominu ohraniti. „Barba,“ poviklfovati, bom perviga pévza med nami, pévza **ljubésni** naprofil, de , kar ni meni mogozhe, tvojo se vonder poftavi te „XXV. POSTAVA. „Leftna **Lubefn** je fa zhloveka narnavarnehshe perlifuvanje unkraj groba terpela, naj se naj_na otroka v **ljubezni** zedinita in vzameta, da bota tako dva imenitna vesti. „Z vestnostjo v najožji vzajemnosti je **ljubezen** do resnice (resnicoljubje). Bistvo resničnosti ino ajdje fo fe ji perdrushevali. Vera ino **ljubesen** fte te duhe s' Jesufam ino med feboj fklépale dolshnufti nyma odvsame, de fe sakonske **lubesni** po fvojih shelah nafmeta dapolniti, dokler glasom rekel: „Pokazali smo svojo vero, svojo **ljubezen** ; ljube hčere, ljubi sinovi, odprite upanju fvojim fovrashnikam eno pravo saftavo fvoje **lubēsni** inu fprave, jim odpufy vfe „govory kir živel in trpel s človeštvom ter ga učil **ljubezni** . „Ana. Da bi mogla tudi jaz trpeti za človeštvo obedve se vstrajno držite svojega predmeta, — **ljubezen** zato, da ga more uživati, — sovraštvo zato molila: „O moj bôshji odrefhenik, ki fi is **ljubésni** do mene na krishu umerl, to Tvoje snamnje Ménart na_to, pogledovaje poln očetovske **ljubézni** in skrbí otroka, ki je tiho in prijazno reshalliti. „Koku sturemo mi pak fadosti **lubefni** nashega blifhnega? „Kader njemu ufte dobrou sovraštvo izvira iz pekla; prijaznost in **ljubezen** pa iz nebes! „Strašimir je danes velike serca v persih ni. „Montgomeri. „Po zakonu **ljubezni** svetodelnimu, Katerimu je vsak podložen meni sveti niso, ne častitlivi. Nič od vezi **ljubezni** klicane ne vem, In nikdar znati nečem njenih vudeh fally toku , de fe namoreta sakonske **lubesne** dolhnufti nekol prov dapolniti. „15. Nezh

Find: Match case

NEW EUP-DESIGN.COM

Konkordančnik CUWI

Iskalnik CUWI in goo300k for [word="ljubezen" | lemma="ljubezen" %c] - Mozilla Firefox

File Edit View History Bookmarks Tools Help

AHLib digital library TOMAŽ ERJAVEC [05023] Konkordance Iskalnik CUWI in goo300k for [word=... x +

nl.ijs.si/cuwi/goo300k/simple?simple_query=ljubezen&peer=goo300k&rnd=3453

Google Koledar MySQL database on Li... Git - Recording Chang... PP Funding > Latest Info ... IMPACT@JSI SkE SkE/Config/FullDoc... Editorial Manager® The Action ISO Standards Develop... IntelliText Corpus Que...

CUWI Izvoz zadetkov Zahtevno iskanje jeziki Pomoč

Iskalnik CUWI • goo300k (starejša besedila, ročno označena)

Preprosto iskanje: Korpus:

Zadetki 1 do 26 od skupno 128 zadetkov za poizvedovanje [word="ljubezen" | lemma="ljubezen" %c] v 0,045 s.

[Nazaj](#) [1](#) [26](#) [51](#) [76](#) [101](#) [126](#) [Naprej](#)

[1] Sacrum prom...	, ter premislite pamet ,	lubesan	, inu poterpeshlivost , katero
[2] Ta male kat...	prov ostudni , inu ozhitnu	lubefni	, inu pravizi tega blišnjega
[3] Ta male kat...	Koku sturemo mi pak šadosti	lubefni	nashega blišnjega ? Kader njemu
[4] Ta male kat...	vire , upanja , inu	lubefne	tudi popolnema grevengo zhes uffe
[5] Ta male kat...	greshil , al ne is	lubefne	bofhje tolkajn , koker is
[6] Ta male kat...	, de Mohor je od	lubefne	pruti Xtusu unèt , inu
[7] Marianske K...	, koker da uralhenja te	lubesne	se beffediti . 2. En
[8] Marianske K...	, temuzh v' eni sveti	lubefni	. Varvej se shepetavzov ,
[9] Marianske K...	inu vilhej tvoj jesek :	lubefn	, inu bofhje strah tvoje
[10] Marianske K...	mano . Tukaj bode tvoja	lubefn	na dan pershla , aku
[11] Marianske K...	pershla , aku bosh to	lubefn	pruti tebi v' temu bofhjemu
[12] Marianske K...	XXV . POSTAVA . Leftna	Lubefn	je ša zhloveka narnavarnehshe perlifuvanje

Find: Match case New Era-Design.com

Več kot le konkordance

The screenshot shows the NoSketch Engine web interface in a Mozilla Firefox browser. The browser's address bar shows the URL `nl.ijs.si/noske/sl-ref.cgi/struct_wordlist`. The page title is "Frequency list - Mozilla Firefox". The browser's tab bar shows several tabs, including "Slovar starejšega slovenskega knjižne...", "Frequency list", and "Digitalna knjižnica IMP".

The NoSketch Engine interface has a blue header with the logo and a search bar. Below the header, the user is logged in as "defaults" and the corpus is set to "IMP (starejša besedila)". The subcorpus is set to "Older". The search bar is set to "in IMP (starejša besedila)".

The main content area is titled "Frequency list". It shows a frequency limit of 1 and a "Set limit" button. Below this, there is a table of word frequencies with horizontal bars representing the frequency of each word.

	<u>nform</u>	<u>Freq</u>
<u>p/n</u> nevarnoft		10
<u>p/n</u> namarnoft		7
<u>p/n</u> sanikernoft		5
<u>p/n</u> navarnoft		5
<u>p/n</u> fupernoft		4
<u>p/n</u> supernoft		3
<u>p/n</u> fanikernoft		2
<u>p/n</u> sanizhemernoft		2
<u>p/n</u> myrnoft		2

At the bottom right of the page, there is a footer with the text "Lexical Computing Ltd. LEX.COM" and "Sketch Engine (ver:open-2.59.1-open-2.91.13)". Below this, there is a line of text: "Interface language: English | česky | 简体中文 | 繁體中文 | Gaeilge | Slovenski".

III. Slovar IMP

- Vsebuje samo atestirane in ročno pregledane primere iz korpusa:
 1. celotni korpus goo300k (1.100 strani, 300.000 besed)
 2. dodatni korpus (4.700 strani, 3.000.000 besed, < 1900),
 - po straneh vzorčen iz digitalne knjižnice IMP
 - tu ročno pregledane in v slovar zajete samo besedne oblike, ki:
 - jih ni v goo300k
 - se pojavijo vsaj 2x
- Namen slovarja:
 - za pregledovanje (pretvorba v HTML, postavitve na splet)
 - kot vir podatkov za jezikovnotehnološka orodja
- Struktura:
 - zaglavje (lema, oznaka, razlaga?)
 - paradigma sodobnih besednih oblik
 - atestirane variante s primeri

Na spletu

Slovar starejšega slovenskega knjižnega jezika IMP: starinske besedne oblike, Stran 53 - Mozilla Firefox

File Edit View History Bookmarks Tools Help

AHLlib digital library TOMAŽ ERJAVEC [05023] Slovar starejšega slovenskega knjižnega jezika

nljjs.si/imp/imp25k/html-m/imp25k-m-0052.html#lex.d3a7c38ae578224967c80c386ff96472

Google Koledar MySQL database on Li... Git - Recording Chang... PP Funding > Latest Info ... IMPACT@JSI SKE SKE/Config/FullDoc... Editorial Manager® The Action ISO Standards Develop... IntelliText Corpus Que...

▶ **zasajenje** samostalnik, občno_ime, srednji (zastarela beseda) ↘ noSkE ↗ SSKJ, Pleteršnik

sadika; kontekst

- zasajenje
 - safajenje

Svetu pismu noviga t...(1784): flifhali? || 13. On pak je odgovóril, inu je djal: Vľaku *safajenje*, katěru ny mój nebeľhki Ozha safadil, bó s' kórenam isruvanu
 - safajenje

Biblia(1584): beľsedo flifhali? On je pak odgovoril, inu je djal: Vľaku *safajenje*, kateru nej moj Nebelki Ozha safadil, bo vunkaj iftrěblenu: Puftite

▶ **zase** zaimek ↘ noSkE ↗ SSKJ, Pleteršnik

- zase
 - sase

To je tapravi inu ta...(1800): sanjo stori karje tabei le unagozh sakai tiľta dusha sama *sase* namora nezheser dovi ispusati al ti kna bosh sanjo perbuegi

To je tapravi inu ta...(1800): te bukve kaker je mena poviedano daso oni ko sami *sase* imeli shribane inu te bukve soble niem ukradane inu tu
 - safe

Kmetijske in rokodel...(1843): perdelk s pridam v dnar ľpravit, kako in kje snajo *safe* potrebne rezhi, to je orodje, shivesh, obleko in t. d.
 - sa ľe (združena oblika)

Kmetijske in rokodel...(1843): ľépe travnike preobernili, de ľo bili s potrebnim shivesham prevideni *sa ľe*, in sa vľe fvoje vojľhne trume, ki ľo ľe

Find: Match case

NEW ERM-DESIGN.COM

Zastarele besede

- razlaga / in vir
 - *kuha / jed, hrana (sskj)*
 - *četveriti / voditi dvovprežno konjsko vprego (Pleteršnik)*
 - *čeznemoč / močnejši kot oni (kontekst)*
 - *cartan / dragocen (splet)*
 - *cedel / list (nemško-slovenski slovar)*
- kratka razlaga, po možnosti sinonimi
- zastarelih lem: 4.177 (SSKJ 1.600, Pleteršnik 770)
- zastarele besede označene že v korpusu
- razrešene dvoumnosti → več gesel z isto lemo
 - npr. *cveten*: „fin, gladek“ (Pleteršnik)
 - *Pifheta s' zvetnim ohravtam*
 - *v' en pifker deni dve shlize zvętne make*

Obseg slovarja

Velikost: kaj štejemo	lem	sodobnih oblik	zgodovinskih oblik
XL: vse	28.121	58.870	81.537
L: besede	21.707	51.154	73.093
M: starinske oblike	12.244	24.776	35.756
S: zastarele besede	4.177	6.175	6.900
XS: stičnost	892	2.348	5.189

IV. ToTrTaLe

Orodje za procesiranje starejših besedil:

1. Tokenizacija: razdelitev besedila na besede
 2. Transkripcija: posodabljanje besednih oblik
 3. Tagiranje: oblikoskladenjsko označevanje
 4. Lematizacija: pripis osnovne oblike
- uporablja slovar IMP, pravila za transkripcijo (*eh#:ih#*) in modela za označevanje in lematizacijo sodobne slovenščine
 - DK IMP → korpus IMP (tudi v konkordančnikih)

ERJAVEC, Tomaž. Automatic linguistic annotation of historical language : ToTrTaLe and XIX century Slovene. V: *5th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, June 19-24, 2011, Portland, Oregon, USA. *LaTeCH 2011*. Portland: Association for Computational Linguistics, 2011, str. 33-38

Dva pristopa

k jezikoslovno-infrastrukturnim projektom

- „klasičen“ pristop: daljše državno financiranje, načrtovanje, narejen & zaključen, dostopen za branje: SSKJ, TURS
- IMP: mednarodno in postopno financiranje, ciklično-kaotično-celostni pristop, več napak, a večji obseg in večja dostopnost ... množicanje?

Zaključki

Predstavljeni trije obsežni, medsebojno povezani in označeni viri starejše slovenščine:

- za jezikoslovce (digitalna knjižnica, konkordančnik, pregledovalnik slovarja)
 - se uporabljajo na FF, mogoče tudi na kakšni gimnaziji in OŠ
- za razvoj jezikovnih tehnologij (licenca CC-BY)
 - prvo (IMPACT) verzijo slovarja uporabljajo na dLib.si
- javno dostopno:
 - DL: 2013-03-24
 - goo300k, imp: 2012-06-15 (!)
 - slovar: 2012-12-26

ERJAVEC, Tomaž. Jezikoslovni viri starejše slovenščine. V: VODOPIVEC, Ines (ur.). *Ljubljana v BiTiH - BiTi v Ljubljani : prispevki iz prvega ljubljanskega kongresa digitalizacije kulturne dediščine: tematska številka*, (Knjižnica, 56, 3). Ljubljana: Zveza bibliotekarskih društev Slovenije: NUK, 2012, str. 205-221.

Nadaljnje delo

- številke v predavanju iz trenutnega stanja
- naslednjo izdajo (v1.0) načrtujemo 2013-06
 - čistimo razširjeni korpus in slovar
 - dopolnjujemo metapodatke in uvajamo taksonomijo
 - pišemo uvod v slovar in korpus
 - ko bosta goo300k in slovar končana, bomo ponovno naučili ToTrTaLe:
 - na novo označiti korpus IMP → konkordančnik
- raziskovalno delo na čim boljših avtomatskih metodah za normalizacijo in oblikoskladenjsko označevanje starejše slovenščine (+ tvitov)
- izziv: statistično strojno prevajanje med starejšo in sodobno slovenščino (besedotvorje, skladnja)
- in seveda - jezikoslovne študije!

Pri izdelavi jezikovnih virov IMP so sodelovali:

Kozma Ahačič, Tina Benčina, Katja Cingerle, Metod Čepar, Darja Fišer, Miran Hladnik, Alenka Jelovšek, Urška Kamenšek, Alenka Kavčič Čolić, Domen Kermc, Maša Kodrič, Simon Krek, Nina Mikulin, Matija Ogrin, Daša Pokorn, Zala Šmid, Ines Vodopivec in Maja Žorga Dulmin.

Hvala Vam na posornosti!