Slovenska akademska besedila: prototipni korpus in načrt analiz

Tomaž Erjavec,[*] Darja Fišer,[†*] Nikola Ljubešić,[*♣] Nataša Logar,[‡] Milan Ojsteršek[♠]

* Odsek za tehnologije znanja, Institut »Jožef Stefan

† Oddelek za prevajalstvo, Univerza v

♣ Filozofska fakulteta, Univerza v Zagrebu

‡ Fakulteta za družbene vede, Univerza v Ljubljani

♠ Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru

JTDH 2016
Ljubljana

Pregled predavanja

1. Uvod
2. Korpus KAS-proto
3. Načrti raziskav
4. Zaključki

Uvod

- Razvoj in uporaba slovenskega jezika v visokem šolstvu ter znanosti je zadnja leta eno osrednjih vprašanj slovenske jezikovne politike
- *Akcijski načrt za jezikovno izobraževanje 2015*: »razvijanje sporazumevalne zmožnosti v /slovenskem/ strokovnem jeziku« ter »izboljšanje položaja slovenščine kot jezika znanosti«
- Raziskovalni projekt ARRS »Slovenska znanstvena besedila: viri in opis« (2016-2018)
- Namen projekta je uresničiti del zgornjih izzivov, podatkovni temelj zanj pa predstavlja obsežen korpus pisnih besedil akademske slovenščine

Portal odprte znanosti

- Nacionalni portal odprte znanosti agregira vsebine iz repozitorijev slovenskih univerz, slovenskih raziskovalnih organizacij in drugih zbirk
- Skupni iskalnik, priporočilni sistem, detektor podobnih vsebin, izvoz metapodatkov, povezan s COBISS.SI
- Dostop do prek 124.000 slovenskih objav s širokega nabora strokovnih področij
- Ta dela so izjemno dragocen, a zaenkrat še pomanjkljivo izkoriščen vir podatkov o akademski slovenščini, kot tudi bogat vir terminologije

KAS-proto

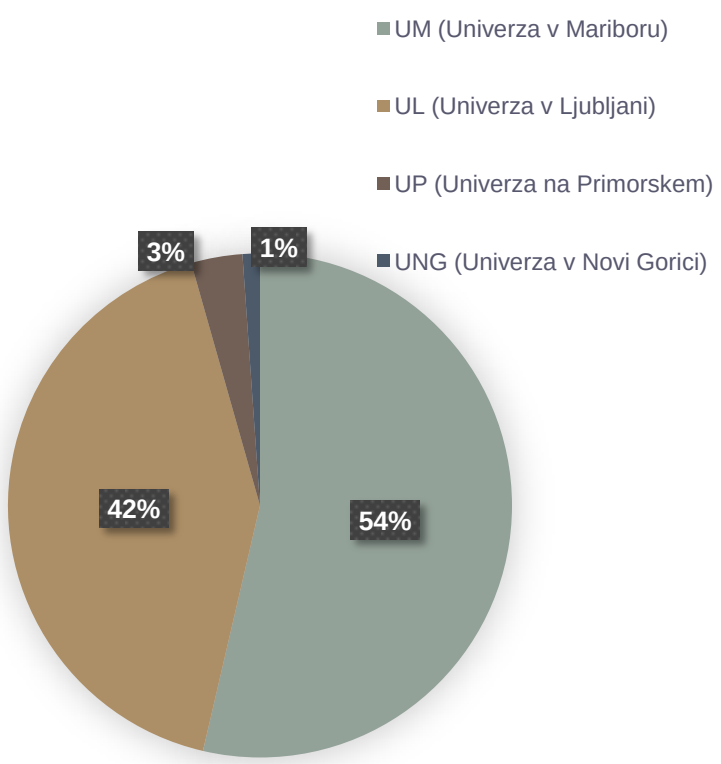
- Izvoz podatkovne baze Nacionalnega portala odprte znanosti v začetku leta 2016:
  PDF + besedilo + metapodatki
- Filtriranje: (razmeroma) dobro slovensko besedilo in metapodatki, od 2000 naprej
- Čiščenje znakov in strukturiranje besedil
- Jezikoslovno označevanje:
  tokenizacija, stavčna segmentacija, lematizacija in oblikoskladenjsko označevanje

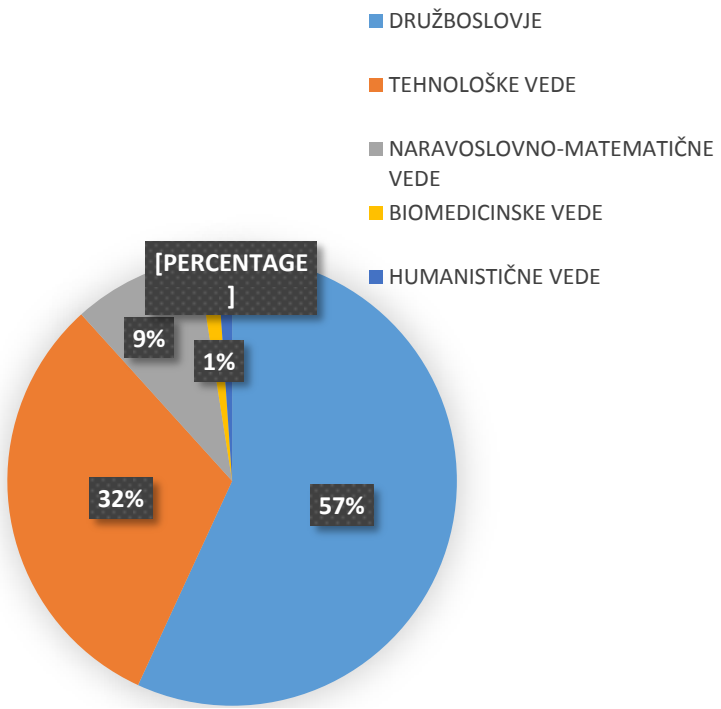
Zapis korpusa

<document xml:id="kas-10000" doc_id="10000" text_id="16514" cobiss_id="7078419"
title="Uravnoteženi sistem kazalnikov v poslovni banki X"
author="Aver, Goran" supervisor="Bernik, Mojca" year="2012"
publisher_abbr="UM FOV" publisher="Fakulteta za organizacijske vede" place="Kranj"
url="http://dkum.uni-mb.si/Dokument.php?id=28143"
type="Diplomsko delo" udc="005" udc_desc="Menedžment"
pdf_url="http://nl.ijs.si/project/kas/pdf/000/kas-10000.pdf">
<page xml:id="pb1" n="1"
pdf_url="http://nl.ijs.si/project/kas/pdf/000/kas-10000.pdf#page=1"
facs_url="http://nl.ijs.si/kas/facs/000/kas-10000/p0001-Pr4U.png">
<p xml:id="pb1.p1" xml:lang="sl">
<s>
<w lemma="diplomski" ana="jos:Agpnsn">Diplomsko</w>
<c> </c>
<w lemma="delo" ana="jos:Ncnsn">delo</w>
…

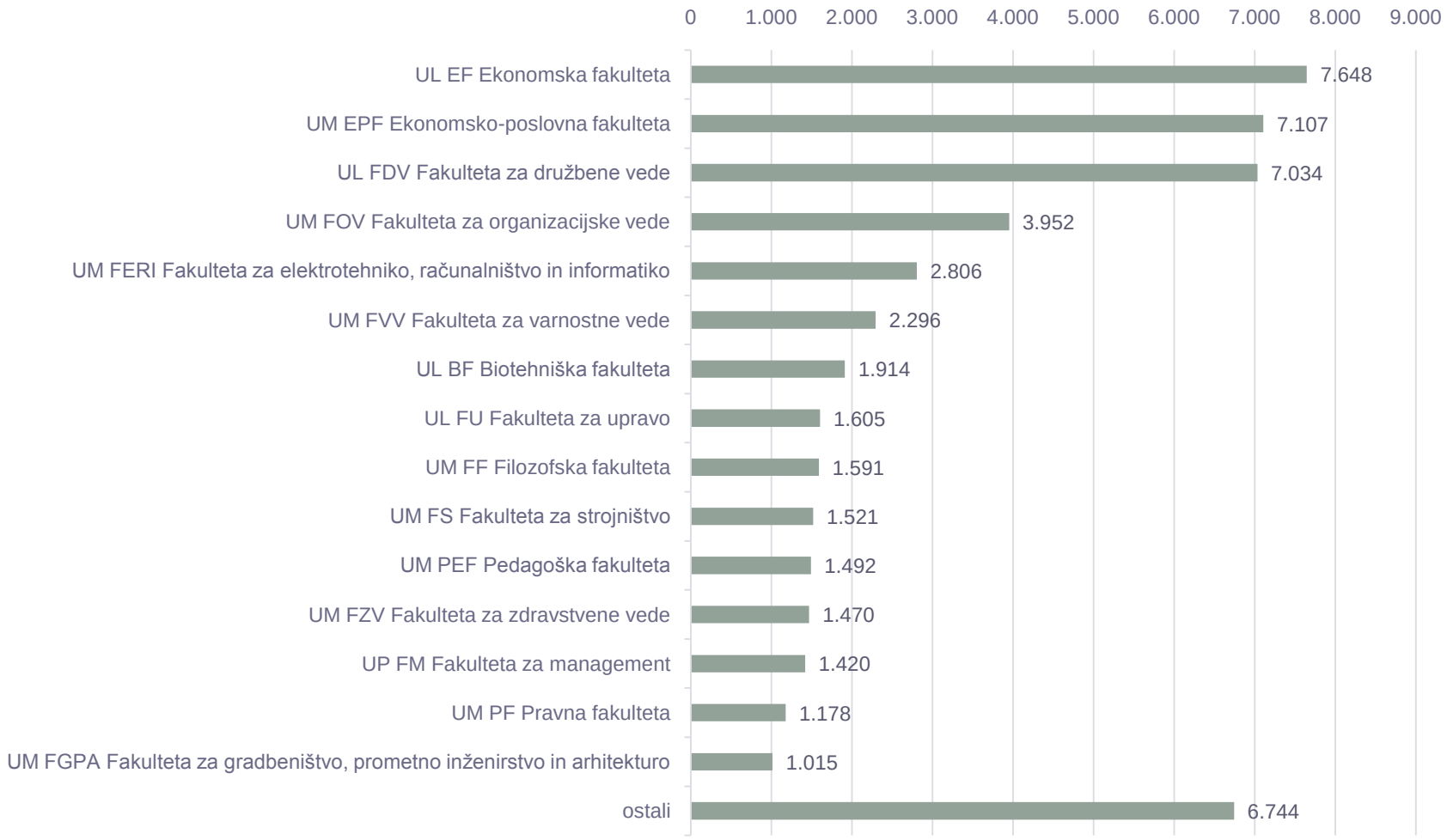
Sestava korpusa

- 50.000 besedil, 4 milijone strani, milijarda besed
- 72 % dipl. 20 % mag., 4 % dr., 4 % ostalo

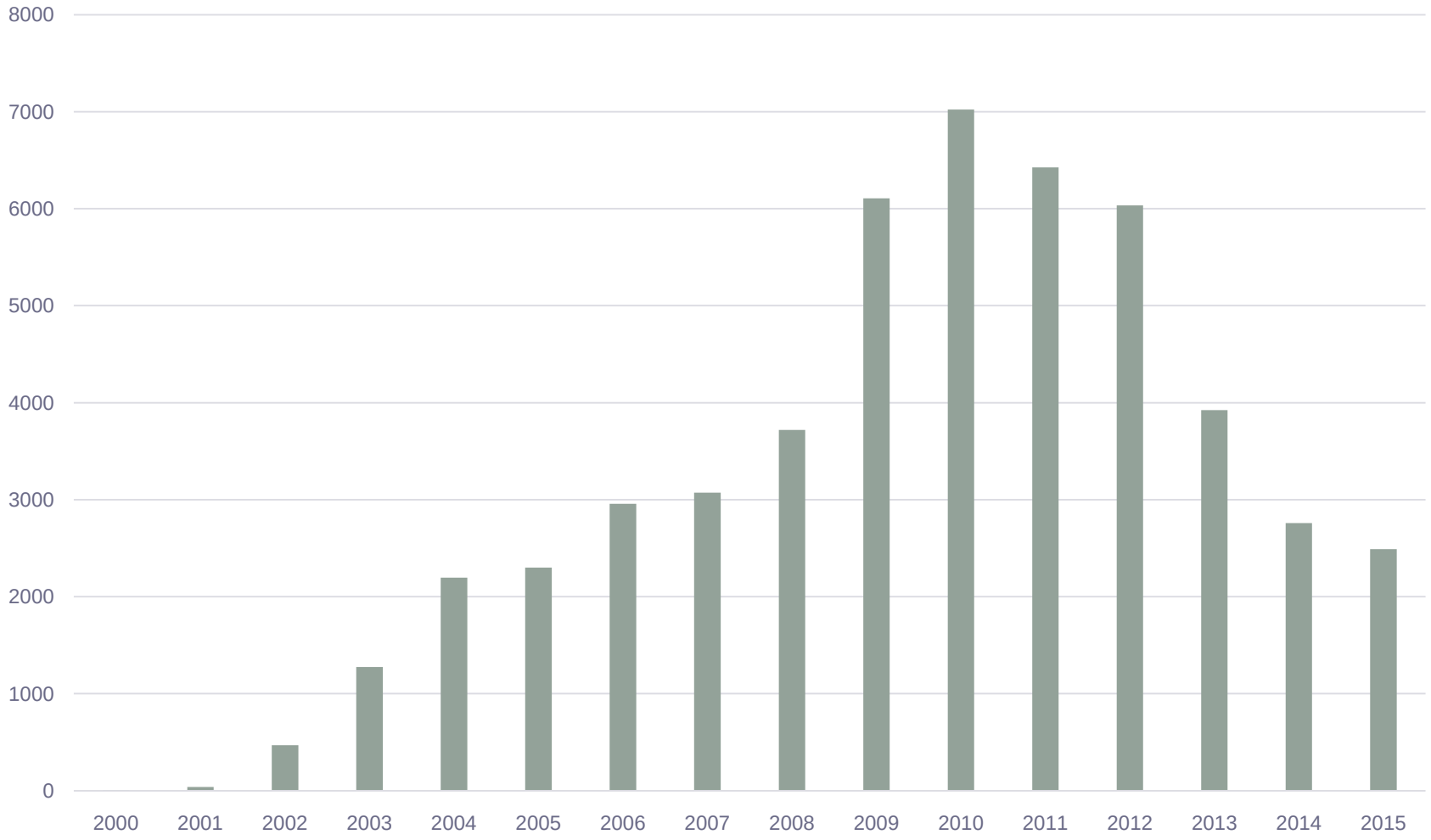




Razporeditev po fakultetah



Razporeditev po letih



Raziskave: klasifikacija besedil (UM)

Motivacija: omogočiti boljše iskanje po vsebinah portala odprte znanosti in nadgradnjo vmesnika za knjižničarje, ki v univerzitetne repozitorije vnašajo nova besedila

- Razvoj metod za klasifikacijo besedil in za luščenje ključnih besednih zvez
- Kot glavno taksonomijo za klasifikacijo bomo uporabili UDK, saj je z njo že opremljena večina dokumentov v repozitorijih
- Učna množica: besedila v korpusu, ki že vsebujejo oznako UDK

Raziskave: orodja za delo s terminologijo (FF)

Motivacija: korpus lahko služi kot bogat vir terminologije, ki pa jo je najprej potrebno identificirati

- Razvili bomo avtomatske metode za luščenje terminologije:
  1. enotskost: zaznavanje jezikovnih prvin, ki sestavljajo večbesedno enoto
  2. terminološkost: razvrščanje po verjetnosti, da so izluščeni termini z določenega področja
  3. variantnost: združevanje pomensko in konceptualno povezanih terminov
  4. prevajanje: identifikacija prevodnih ustreznic terminov v drugem jeziku
- Izvedli bomo analizo uporabe terminov po strokovnih področjih in časovnih obdobjih
- Izluščene term. kandidate bomo objavili v spletnem slovarskem urejevalniku, ki bo slovenskim znanstvenim in strokovnim skupnostim omogočal upravljanje s terminologijo lastnih področij

Raziskave: opis slovenskega akademskega jezika (FDV)

Podatki iz korpusa nam bodo omogočili pripravo opisa sodobnega slovenskega jezika, kakršen se rabi v akademskem okolju.

Opis bo nastal s sintezo rezultatov treh vrst analiz:

1. leksikalne analize (npr. frekvenčni profil)
2. besediloslovne in slovnične analize (Logar in dr. 2016)
3. stilne analize

Zaključki

- Predstavili smo korpus KAS-proto, njegovo izdelavo in kvantitativni vpogled v njegovo sestavo ter pregled načrtovanih raziskav
- Javno dostopna različica korpusa KAS-proto:
  - brez <front> in <back>
  - povezava na posamezno stran, ne pa na PDF
  - KAS, KAS-dipl, KAS-mag, KAS-dr
  - dostopen na noSketchEngine @ nl.ijs.si: http://nl.ijs.si/noske/
  - tudi http://www.clarin.si/kontext/