MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA

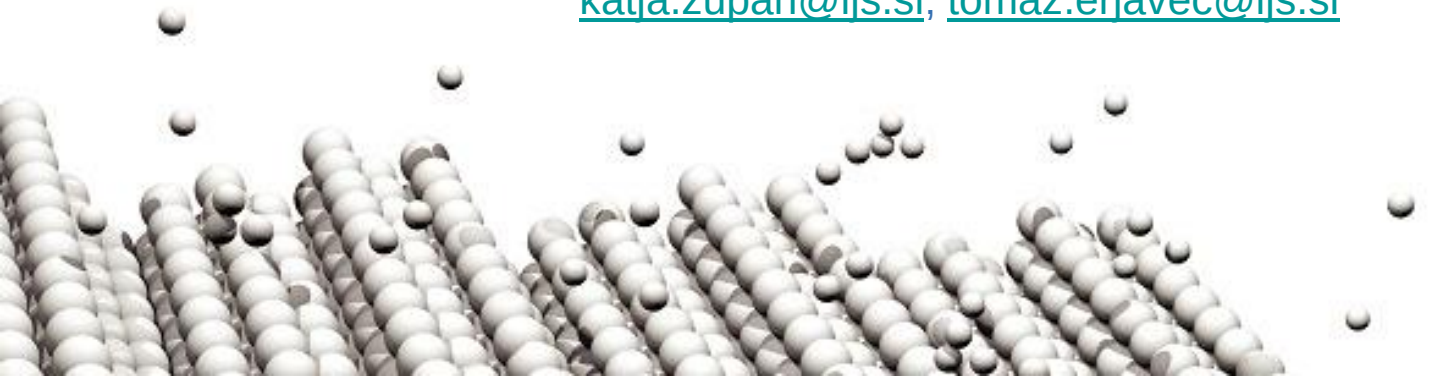
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Generiranje kritičnih prepisov s strojnim prevajanjem na ravni znakov

Katja Zupan in Tomaž Erjavec

Institut „Jožef Stefan“ in Mednarodna podiplomska šola Jožefa Stefana

katja.zupan@ijs.si; tomaz.erjavec@ijs.si



Struktura predstavitve

- Uvod in motivacija
- Diplomatični in kritični prepis
- (Pred)obdelava starejših besedil
- Strojno prevajanje na ravni znakov
- Primer uporabe: Slomškovi pridigi
- Sklepne ugotovitve in prihodnje delo

UVODNE MISLI

- Digitalizacija pisne kulturne dediščine
- Znanstvene oz. znanstvenokritične izdaje
- Rokopisi: pregled, prepis, rekonstrukcija, komentar
- Dve obliki prepisov

Prepisi

- DIPLOMATIČNI: natančen tipografski dvojnik rokopisa;

  avtorjeve nedoslednosti, napake, popravki, vrivki ohranjeni

- KRITIČNI: tekstnokritična interpretacija,

  besedilo približa sodobnemu bralcu s sistematičnimi spremembami na ortografski in oblikoslovno-leksikalni ravni

  >> zamudno, bi bilo mogoče delno avtomatizirati?

(Pred)obdelava starejših besedil

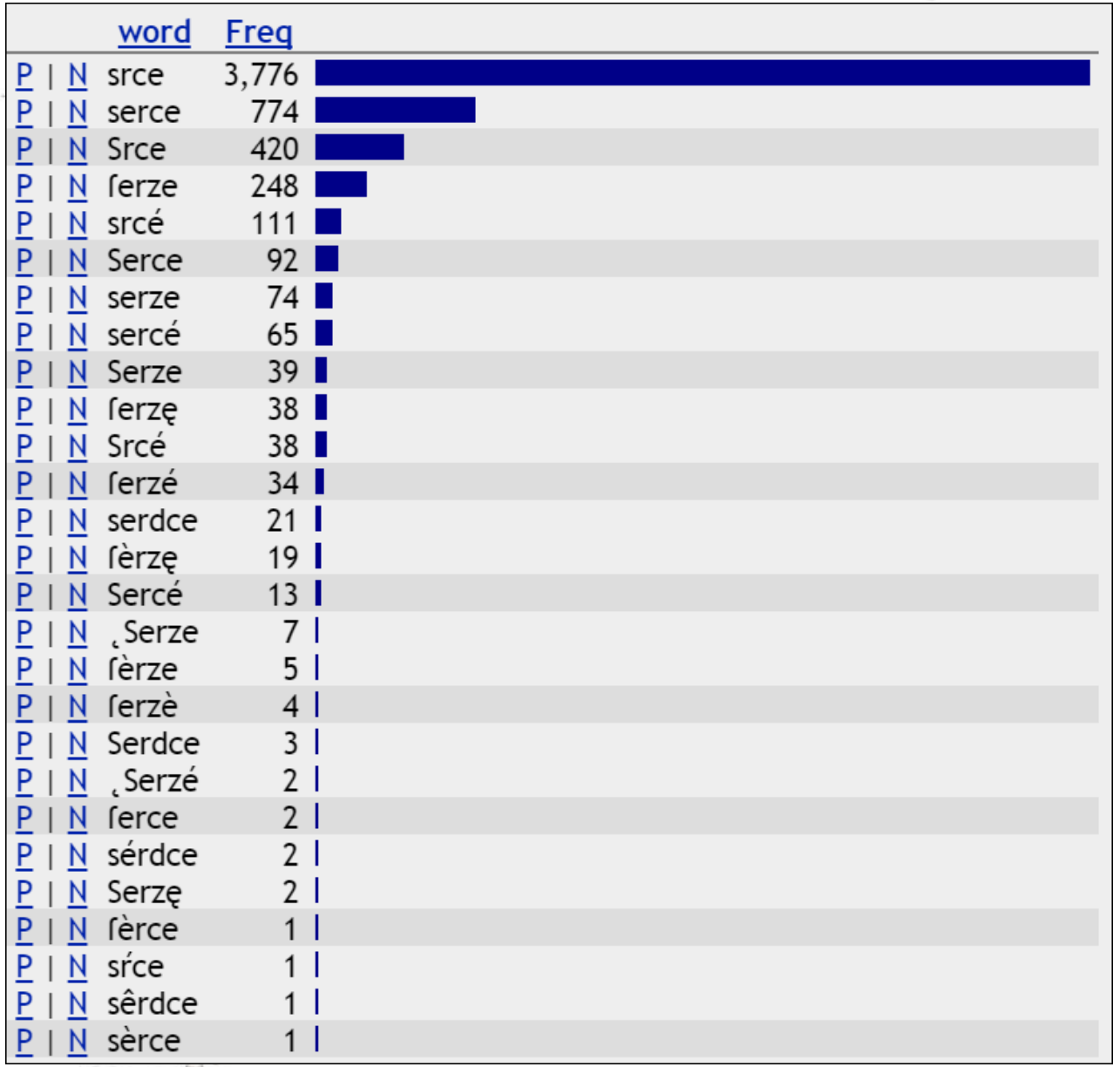
- Odsotnost pisne norme: različen zapis besed
- Za koga je to težava?

❑za sodobne bralce: težje razumevanje in iskanje po besedilu

❑za orodja za obdelavo naravnega jezika: slabša lematizacija in oblikoskladenjsko označevanje

Pomemben korak v (pred)obdelavi: normalizacija/modernizacija/kanonikalizacija

Samostalnik ‚srce‘ v imenovalniku ednine:



	word	Freq
P | N	srce	3,776
P | N	serce	774
P | N	Srce	420
P | N	ſerze	248
P | N	srcé	111
P | N	Serce	92
P | N	serze	74
P | N	sercé	65
P | N	Serze	39
P | N	ſerzę	38
P | N	Srcé	38
P | N	ſerzé	34
P | N	serdce	21
P | N	ſèrzę	19
P | N	Sercé	13
P | N	˛Serze	7
P | N	ſèrze	5
P | N	ſerzè	4
P | N	Serdce	3
P | N	˛Serzé	2
P | N	ſerce	2
P | N	sérdce	2
P | N	Serzę	2
P | N	ſèrce	1
P | N	sŕce	1
P | N	sêrdce	1
P | N	sèrce	1

Kako normalizirati?

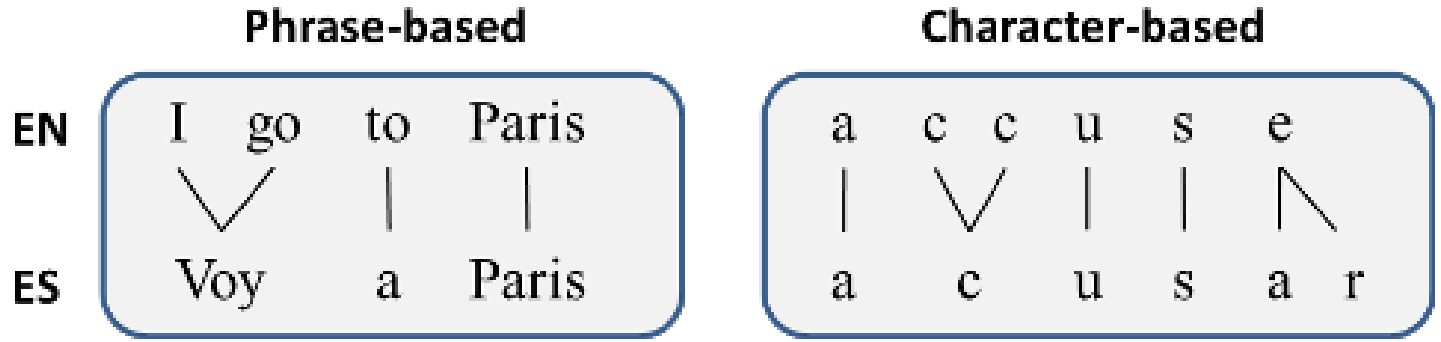
Več pristopov:

- izdelava leksikona parov besedne oblike;
- izdelava pravil: ročno ali avtomatsko;
- fonetično ujemanje;
- računanje razdalje med oblikama;
- statistično strojno prevajanje.

Statistično strojno prevajanje na ravni znakov

prilagojena metoda statističnega strojnega prevajanja na ravni besednih zvez

>> znaki kot besede, vmesni presledki



Prednosti, slabosti in omejitve

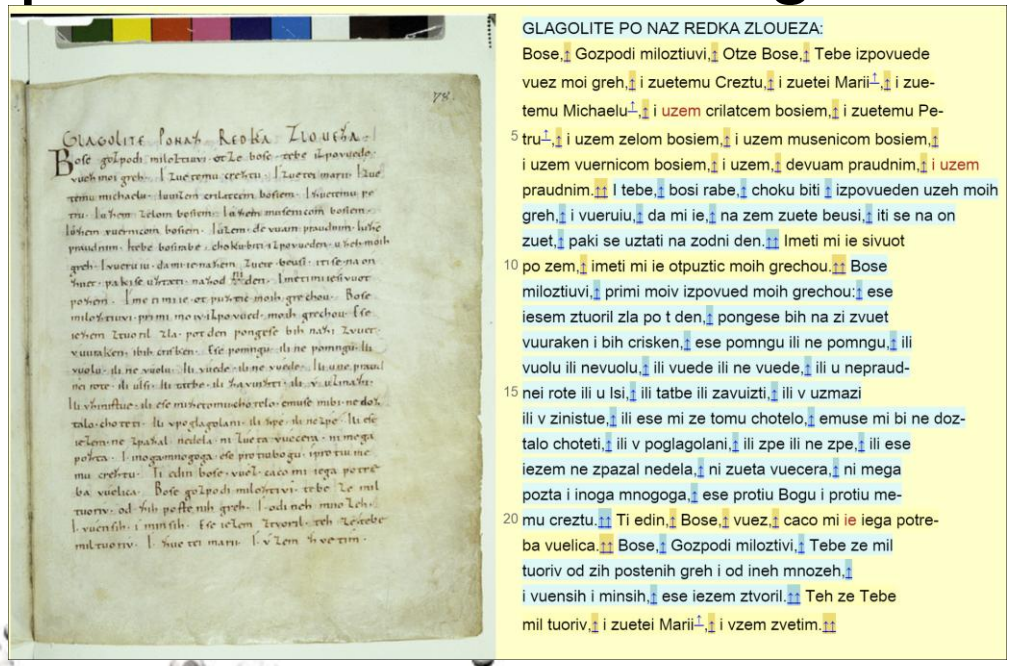
+ zadostuje manjša učna množica kot pri klasičnem strojnem prevajanju, zna prevajati besede, ki jih ni v učni množici

– sistemske spremembe na ravni znakov

- običajno: posamezna pojavnica (ni ko(n)teksta, zapis skupaj/narazen)
- normalizacija = modernizacija
- Naš poskus: vrstica kot osnovna enota, normalizacija v posodobljeni, a ne sodobni jezik

Primer uporabe: Slomškovi pridigi

- eZISS, elektronske znanstvenokritične izdaje slovenskega slovstva: http://nl.ijs.si/e-zrc/
- izbrana slovenska besedila v integraciji faksimilov, prepisov in znanstvenega komentarja
- večinoma rokopisi



GLAGOLITE PO NAZ REDKA ZLOUEZA:

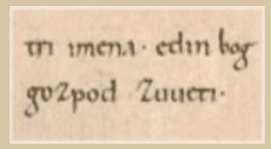
Bose,‡ Gozpodi miloztiuvi,‡ Otze Bose,‡ Tebe izpovuede
vuez moi greh,‡ i zuetemu Creztu,‡ i zuetei Marii‡–,‡ i zue-
temu Michaelu‡–,‡ i uzem crilatcem bosiem,‡ i zuetemu Pe-
tru‡–,‡ i uzem zelom bosiem,‡ i uzem musenicom bosiem,‡
i uzem vuernicom bosiem,‡ i uzem,‡ devuam praudnim,‡ i uzem
praudnim.‡‡ I tebe,‡ bosi rabe,‡ choku biti ‡ izpovueden uzeh moih
greh,‡ i vueruiu,‡ da mi ie,‡ na zem zuete beusi,‡ iti se na on
zuet,‡ paki se uztati na zodni den.‡‡ Imeti mi ie sivuot
po zem,‡ imeti mi ie otpuztic moih grechou.‡‡ Bose
miloztiuvi,‡ primi moiv izpovued moih grechou:‡ ese
iesem ztuoril zla po t den,‡ pongese bih na zi zvuet
vuuraken i bih crisken,‡ ese pomngu ili ne pomngu,‡ ili
vuolu ili nevuolu,‡ ili vuede ili ne vuede,‡ ili u nepraud-
nei rote ili u lsi,‡ ili tatbe ili zavuizti,‡ ili v uzmazi
ili v zinistue,‡ ili ese mi ze tomu chotelo,‡ emuse mi bi ne doz-
talo choteti,‡ ili v poglagolani,‡ ili zpe ili ne zpe,‡ ili ese
iezem ne zpazal nedela,‡ ni zueta vuecera,‡ ni mega
pozta i inoga mnogoga,‡ ese protiu Bogu i protiu me-
mu creztu.‡‡ Ti edin,‡ Bose,‡ vuez,‡ caco mi ie iega potre-
ba vuelica.‡‡ Bose,‡ Gozpodi miloztivi,‡ Tebe ze mil
tuoriv od zih postenih greh i od ineh mnozeh,‡
i vuensih i minsih,‡ ese iezem ztvoril.‡‡ Teh ze Tebe
mil tuoriv,‡ i zuetei Marii‡–,‡ i vzem zvetim.‡‡

eZISS

ELEKTRONSKE ZNANSTVENOKRITIČNE IZDAJE SLOVENSKEGA SLOVSTVA

IZDAJE

BRIŽINSKI SPOMENIKI



Najstarejša slovenska besedila, prvi dokument slovenske kulture.

ŠKOFJELOŠKI PASIJON

Najzgodnejše ohranjeno dramsko delo v slovenščini.

MESTNA PRISEŽNA BESEDILA

eZISS

Elektronske znanstvenokritične izdaje slovenskega slovstva ponujajo izbrana slovenska besedila v integraciji faksimilov, prepisov in znanstvenega komentarja, mestoma tudi avdiovizualnih posnetkov.

Projekt eZISS poskuša ustvariti sintezo treh prvin. Prva je tradicija slovenskega slovstva v razponu od srednjeveškega rokopisa in ljudske pesmi do estetske literarne umetnine. Druga prvina je ekdotika – bogata tradicija filološkega proučevanja besedil in njihovih predstavitev v znanstvenokritičnih izdajah. Tretja prvina so sodobne informacijske tehnologije. Spoj dveh tradicij in sodobne tehnologije, ki je na področju obdelave besedil že presegla raven eksperimentov, oblikuje jasne standarde in se uveljavlja kot pomemben posrednik literarnega izročila. Elektronske znanstvene izdaje ne odpravljajo klasičnih tiskanih izdaj, ponujajo le nekatera pomembna dopolnila in pot k bolj raznoliki bralski recepciji.

Kompleksna predstavitev besedil s faksimili, prepisi, kritičnim aparatom in avdiovizualnimi posnetki je mogoča z dosledno rabo sodobnih standardov odprtokodnega označevanja besedil: Unikod, XML, smernice konzorcija TEI. Le ta podlaga ustreza zahtevi, naj bodo naše izdaje odporne na tehnološke spremembe, neodvisne od programske opreme in združljive z drugimi standardiziranimi elektronskimi viri. Za branje teh izdaj ne potrebujemo drugega kakor navaden brskalnik.

Izdaje so namenjene trajni javni uporabi, zato so vsakomur dostopne na svetovnem spletu. Omogočen je tudi prenos vsake izdaje na vaš računalnik, kjer jo lahko uporabljate brez spletne povezave. Ob upoštevanju licenčnih določil lahko izdaje posredujete tudi drugim na digitalnih nosilcih.

O projektu

V projektu eZISS "Elektronske znanstvenokritične izdaje slovenskega slovstva" sodelujeta Inštitut za slovensko literaturo in literarne vede ZRC SAZU (vodja projekta Matija Ogrin) in Odsek za tehnologije znanja IJS (vodja projekta na IJS Tomaž Erjavec).

Priprava elektronskih izdaj eZISS je v slovenskem jeziku dokumentirana v naslednjih objavah (angleške so dostopne tule):

A. M. SLOMŠEK: TRI PRIDIGE O JEZIKU

- Izdaja v HTML
  - I. Za krščansko govorjenje
  - II. Jezik je vir dobrega in zla
  - III. Svoj jezik je treba spoštovati
- Prenos na vaš računalnik:
  - celotna izdaja (9MB .zip → 11MB)
  - brez faksimilov (225kB .zip → 3MB)

Anton Martin Slomšek
TRI PRIDIGE O JEZIKU

- Uredništvo, diplomatični in kritični prepis: Jože Faganel
- Uredništvo, opombe, redakcija zapisa TEI: Matija Ogrin
- Računalniški zapis in redakcija TEI: Tomaž Erjavec
- Izdajatelj: Inštitut za slovensko literaturo in literarne vede ZRC SAZU
- Izdaja 1.6, 2007-04-06
- Mesto objave: http://nl.ijs.si/e-zrc/slomsek/
- Kolofon TEI

Anton Martin Slomšek (1800–1862), slovenski škof, pisatelj, kulturni delavec, zaslužen še zlasti za preporod slovenske narodne zavesti v vzhodni Sloveniji in na Koroškem. Papež Janez Pavel II. ga je leta 1999 razglasil za blaženega.

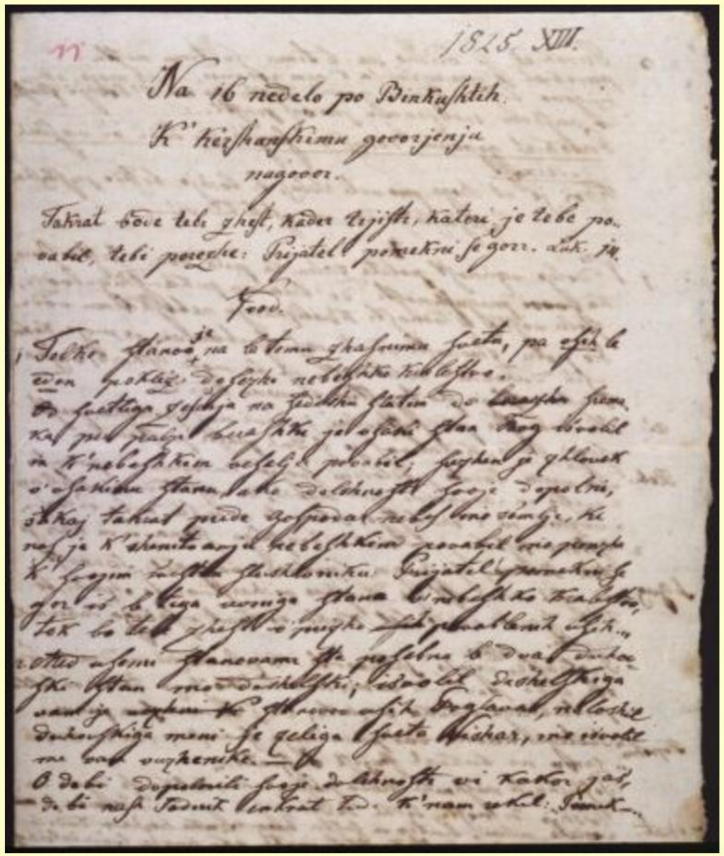
Tri pridige o jeziku A. M. Slomška so nastale v času od leta 1825 do 1841. Dve od izbranih treh besedil sta ohranjeni v avtorjevem rokopisu. Rokopis tretjega besedila, ki je najbolj znano, je izgubljen, a kmalu po nastanku dvakrat natisnjen, nato je bila znamenita pridiga večkrat delno objavljena in citirana. Vsa tri besedila povezuje v celoto skupna tematika. Besedila so govornikove predloge za pridiganje, formulirane v celoti kot dokončno besedilo, torej literarnozvrstno verski govor, točneje pridiga oziroma t.i. retorska proza.

Elektronska izdaja vsebuje predgovor, faksimile, diplomatični prepis, kritični prepis in urednikove opombe. Vsi prepisi so povezani s faksimili, medsebojno pa so povezani po vrsticah. To omogoča vzporedni prikaz faksimilov s prepisi, kakor tudi vzporedni prikaz faksimila z obema prepisoma. Slomškova robna sklicevanja na svetopisemske vire so povezana s spletnim portalom Biblija.net, kjer je njegovo slovenjenje bibličnih odlomkov moč primerjati z raznimi prevodi Svetega pisma v slovenščino.

Izdajo lahko uporabljate preko spleta, lahko pa si jo prenesete tudi na svoj računalnik. Preneseno datoteko shranite in dekompresirate (unzip), nato v dobljeni mapi kliknete na index.html. Tako lahko uporabljate izdajo brez spletne povezave. Skupaj z datotekami HTML in (če izberete prenos celotne izdaje) s faksimili boste dobili tudi izvorne datoteke XML/TEI, primerne za računalniške obdelave besedila; najdete jih v mapi tei .

http://nl.ijs.si/e-zrc/slomsek/index-sl.html

„Za krščansko govorjenje“ (1825)



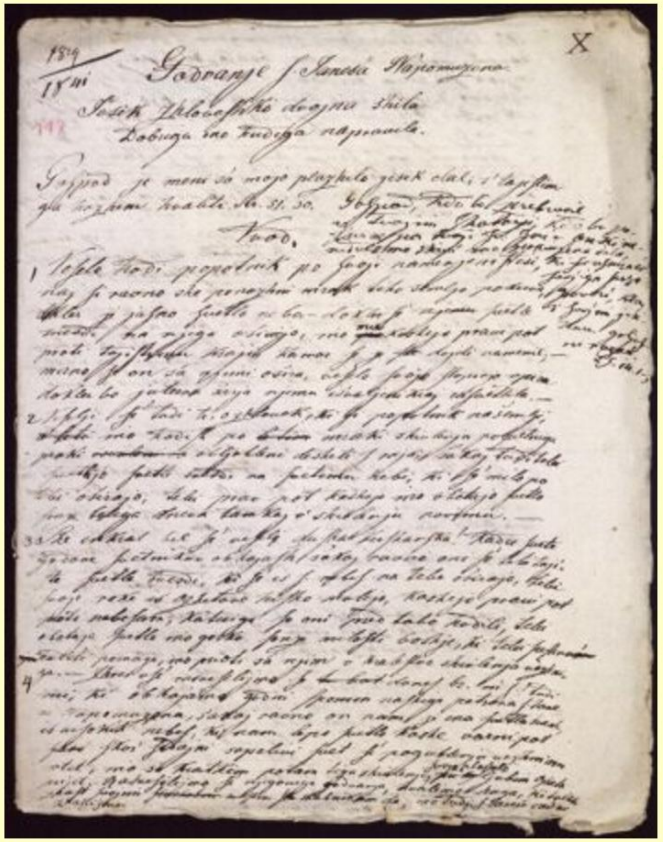
I.

Za krščansko govorjenje

Vzporedni prikaz

1825. XIII	
1825. XIII	1.1 op. [1]
Na 16 nedelo po Binkuſhtih.	
Na 16. nedeljo po Binkuštih.	1.2 op. [2-
K'kerſhanſkimu govorjenju	
K krščanskemu govorjenju	
nagovor.	
nagovor	1.4 op. -2]
Takrat bode tebi zheſt, kader tijiſti, kateri je tebe po=	1.5
Takrat bode tebi čest, kader tijisti, kateri je tebe po-	1.5
vabil, tebi porezhe: Prijatel, pomekni ſe gori. *Luk. 14.*	
vabil, tebi poreče: Prijatelj, pomekni se gori. Luk. 14. [10]	
Vvod.	
Uvod	
1. Tolko ſtánov je na le temu zhaſnimu ſvetu, pa vſih le	
1. Tolko stánov je na le-tem časnem svetu, pa vseh le	
eden pokliz: doſezhi nebeshko kraleſtvo.	
eden poklic: doseči nebeško kral[j]estvo.	
Od ſvetliga zeſarja na ſedeshu ſlatim do ~~berazha~~ ſroma=	1.10
Od svetlega cesarja na sedežu zlatem do sroma-	1.10
ka per palzi beraſhki je vſaki ſtan Bog is'volil	
ka pri palci beraški je vsaki stan Bog izvolil	

„Jezik je vir dobrega in zla“ (1829)



II.

Jezik je vir dobrega in zla

Vzporedni prikaz

1829		*X*
1829		*X* 2.1 *op. [19-*
1841	Godvanje ſ. Janes'a Napomuzena.	
1841	Godvanje sv. Janeza Napomucena.	2.2 *op. -19]*
	Jes’ik zhloveſhki dvojna s’hila	
	Jezik človeški dvojna žila	
	Dobriga ino hudiga napravila.	
	dobrega ino hudega napravila.	
	Goſpod je meni s’a mojo plazhilo jesik dal; s’ tajiſtim	2.5
	Gospod je meni za moje plačilo jezik dal; s tajistim	2.5
	ga hozhem hvaliti.*Sir. 51,30*	
	ga hočem hvaliti. *Sir. 51, [22 (30)]*	2.6 *op. [20]*
	Goſpod, kdo bo prebival	
	Gospod, kdo bo prebival	2.7 *op. [21-*
	v' tvojim ſhotorji, kdo bo po-	
	v tvojem šotorji, kdo bo po-	
	zhival na tvoji ſ. Gori? On ki ne-	
	čival na tvoji sv. gori? On ki ne-	
	nedolshno shivi ino pravizhno dela,	2.10
	dolžno živi ino pravično dela,	2.10
	ki ſ reſnizo is’	
	ki resnico iz	

Metodologija

1. pridiga kot učna, 2. kot testna množica

Primerjava 3 metod:

- Norma (leksikon, mere razdalje)
- SSP: klasično
- SSP-Z: znaki

Orodje: Moses s privzetimi nastavitvami (KenLM) in Giza++ za poravnavo

Rezultati

Izhodišče:

delež razlik v znakih (povprečna Levenshteinova razdalja) = **22,26 %**

DRZ (%)	2-grami	3-grami	4-grami	5-grami
SSPZ-JM1	**7,59**	8,58	8,71	9,18
SSPZ-JM2	8,05	**8,14**	8,21	8,80
SSPZ-JM3	**8,26**	8,45	8,48	9,50
SSP	**16,84**	16,87	16,86	16,87
Norma	**14,19**			

Jezikovni model 1 (JM1): kritični prepis 1. pridige

JM2: JM1 + Slomškova zbrana dela

JM3: JM1 + jos 100k

Primer dobrega in slabega prevoda

Izvirnik	Ročni prepis	Strojni prevod
1.Veſelo hodi popotnik po ſvoji namenjeni ſtes'i	1. Veselo hodi popotnik po svoji namenjeni stezi	1.Veselo hodi popotnik po svoji namenjeni stezi (1 razlika)
proti tajiſstimu kraju, kamor ſe je ſte doiti namenil;-	proti tajistemu kraju, kamor se je doiti namenil;-	proti tajistimu kraju, kamor se je ste dojiti namenil; (6 razlik)

Sklepne ugotovitve

- Najbolje deluje predlagana metoda SSP-Z, prihrani dve tretjini dela (22,26 > 7,59%).
- Majhen in preprost jezikovni model, privzete nastavitve zadoščajo.
- Prevajanje vrstic namesto pojavnic.
- Metoda je uporabna, potreben pa vsaj del kritičnega prepisa oz. podoben prepis.

Prihodnje delo

- Kvalitativna analiza
- Poskusi z uglaševanjem uteži

LM order	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10	AVG
untuned	6,833	6,833	6,833	6,833	6,833	6,833	6,833	6,833	6,833	6,833	6,833
2-grams	6,254	5,834	6,321	6,424	5,920	6,104	6,021	6,125	6,017	6,273	6,129
3-grams	6,038	5,924	5,864	5,967	5,823	6,037	5,878	5,864	5,956	5,924	5,928
4-grams	5,892	5,966	5,838	5,986	5,949	5,793	5,854	5,899	6,038	5,936	5,915
5-grams	5,870	6,162	5,889	6,110	5,815	5,794	5,750	6,032	5,814	5,908	5,914
6-grams	6,161	6,014	5,846	6,008	5,954	6,019	5,958	6,158	5,942	6,103	6,016
7-grams	5,951	6,037	6,170	5,948	5,974	6,108	6,080	5,953	5,996	5,977	6,019
8-grams	5,953	5,817	5,834	6,170	5,993	5,833	5,909	5,962	6,094	5,865	5,943
9-grams	5,834	5,931	5,820	5,887	5,879	5,813	5,877	5,840	5,998	5,809	5,869
10-grams	5,732	5,994	5,988	5,959	6,011	5,751	6,016	5,972	5,878	6,114	5,942

Jas tabei razham ***hualo*** noi sakvalo inu te sahvalem …

Vpraſhanja?