Generiranje kritičnih prepisov s strojnim prevajanjem na ravni znakov

Katja Zupan, Tomaž Erjavec

Odsek za tehnologije znanja, Institut »Jožef Stefan« in
Mednarodna podiplomska šola Jožefa Stefana
Jamova cesta 39, 1000 Ljubljana
katja.zupan@ijs.si
tomaz.erjavec@ijs.si

Povzetek

Pomembnejši rokopisi so pogosto predstavljeni v dveh prepisih, diplomatičnem in kritičnem, kjer prvi sledi izvirniku, drugi pa ga interpretira, pri čemer tudi delno posodobi njegovo besedilo. Izdelava kritičnega prepisa je zamuden postopek, ki bi ga lahko olajšali z avtomatskimi metodami. V prispevku predstavimo orodje, ki temelji na statističnem strojnem prevajanju znakov in prevaja posamezne vrstice diplomatičnega prepisa v kritičnega. Metodo smo preizkusili na dveh pridigah (1825, 1829) A. M. Slomška in jo primerjali z več drugimi pristopi. Preizkusi pokažejo, da program, naučen na prvi pridigi in preizkušen na drugi, zmanjša delež razlik na ravni znakov za skoraj dve tretjini in deluje bolje kot preostale metode.

Generating Critical Transcriptions with Character-based SMT

Historic manuscripts are often represented in the form of two transcriptions, a diplomatic and a critical one. The former follows closely the original while the latter interprets it, also by partly modernising its text. Generating a critical transcription is a time-consuming process, which could be improved through automatic methods. The paper presents a tool that uses character-based statistical machine translation, translating each line of the diplomatic transcription into the critical one. The method has been tested on two sermons (1825, 1829) by A. M. Slomšek, and compared to several other approaches to text modernisation. The experiments show that training a program on the first sermon and testing it on the second one reduces the character error rate by nearly two thirds, performing better than other methods.

1 Uvod

Napredek jezikovno usmerjenih informacijskih tehnologij vpliva na razvoj številnih področij humanistike. Na področju besedilne kritike je digitalizacija besedil odprla nove možnosti zapisa, raziskovanja in prikaza in s tem spodbudila spremembo paradigme, kako besedilo predstaviti in posredovati uporabnikom.

Ponoven razcvet je tako doživela tudi metodologija znanstvenih oz. znanstvenokritičnih izdaj, kot se v literarnih vedah imenujejo tiste edicije – najpogosteje pomembnejših starejših rokopisov –, v katerih so izvirna besedila (oz. njihovi faksimili) pregledana, prepisana, rekonstruirana, komentirana in naposled objavljena po načelih tekstne kritike (Erjavec in Ogrin, 2004). Takšna besedila so običajno predstavljena v obliki dveh vrst prepisov, diplomatičnega in kritičnega.

Diplomatični prepis je v tiskani/digitalni obliki čim bolj natančen tipografski dvojnik rokopisa z vsemi avtorjevimi nedoslednostmi, napakami, avtorjevimi popravki, vrivki ipd. vred. Njegova funkcija je, da predstavlja izvirnik in tako olajša branje težje berljivih mest rokopisa. Kratice, nejasna in poškodovana mesta rokopisa pušča nerazrešena in nedopolnjena.

Kritični prepis je že tekstnokritična interpretacija besedila, ki izvirno besedilo redigira in nekoliko modificira na črkopisni in oblikoslovni ravni, vendar se pri tem ravna po ekspliciranih načelih, pridobljenih z indukcijo iz analize samega besedila. Vsebuje tudi ustrezen t. i. tekstnokritični aparat z opombami. (Dović, 2006, str. 210). Diplomatični prepis poskuša torej čim zvesteje slediti izvirniku, kritični pa z modificiranjem jezikovnega izraza skuša doseči ravnotežje med ohranjanjem avtentičnosti jezika in razumljivostjo sodobnemu bralcu oz. uporabniku. Modificirani jezikovni izraz torej ni nujno sodobni standardni jezik, temveč mu je samo približan. Stremi k ohranjanju arhaičnega vtisa diplomatičnega prepisa, na primer z (delnim) ohranjanjem starinskih besed, hkrati pa jih poda v sodobni pisni podobi, tj. v sodobnem črkopisu (npr. gajici).

Uredniško delo pri ustvarjanju kritičnega prepisa tako zajema tudi razmeroma preproste in sistematične posege, kot je posodabljanje črkopisa, oz. sistematične posege na oblikoslovno-leksikalni ravni. V pričujočem prispevku raziskujemo, kako bi bilo mogoče tovrstne posege izvajati strojno, s pomočjo prilagojene računalniške metode strojnega prevajanja znakov, ki bi na nezahteven način samodejno predlagala ustrezno (delno) posodabljanje besed ter s tem poenostavila in pospešila uredniško delo.

2 Strojna normalizacija besed

Za posodabljanje (ali, širše, normalizacijo) starejših besedil je bilo razvitih že več računalniških metod (Piotrowski, 2012), od uporabe ročno napisanih pravil za pretvorbo starejših oblik v sodobni oz. normalizirani zapis (Erjavec, 2015), prek avtomatske izpeljave tovrstnih pravil (Bollmann, 2012) do uporabe statističnega strojnega prevajanja na ravni znakov, ki »prevaja« arhaične zapise besed v sodobne oz. normalizirane (Tiedemann, 2009; Scherrer in Erjavec, 2015).

Vsem tem metodam je skupno, da je posodabljanje izvedeno na ravni posamezne pojavnice (besedne oblike) in da se zanaša na obstoj velikega leksikona sodobnega jezika, da izloči hipoteze, ki jih sistem generira, ker so sicer sistemsko možne, a niso del besedja določenega jezika.

Obe domnevi sta problematični, če ju uporabimo pri izdelavi kritičnega prepisa starejših besedil. Pogosta razlika med arhaičnim in posodobljenim zapisom je namreč zapis skupaj oz. narazen, tj. kot dve besedi na ortografski ravni. Vsak sistem, ki besedilo najprej razdeli na pojavnice in nato

izvede njihovo pretvorbo, bo v teh primerih zatajil. Druga težava, ki še posebej velja za kritične izdaje, pa je neuporabnost leksikona sodobnega jezika, saj v njem ne bo arhaično obarvanih besed, ki jih običajno najdemo v kritičnih prepisih.

3 Metodologija

3.1 Strojno prevajanje na ravni znakov

Statistično strojno prevajanje (SSP oz. ang. Statistical Machine Translation, SMT) (Brown et al. 1993, Koehn, 2010) je sklop metod, ki temelji na učenju (kombinacije) modela prevajanja in modela ciljnega jezika. Prvega gradimo na vzporednih korpusih, ki so tokenizirani in večinoma poravnani po povedih, drugega pa na enojezičnem korpusu jezika, v katerega prevajamo. Naučeni sistem je nato sposoben prevajati povedi izvornega jezika v ciljnega. Na tem principu temelji Googlov prevajalnik, v okviru več (tudi zelo velikih) projektov pa je bil izdelan odprtokodni sistem Moses, ki ga je razmeroma preprosto namestiti na računalnik in uporabiti tako za učenje kot tudi izvajanje prevajanja. Moses je kompleksno orodje, ki omogoča prilagojene nastavitve veliko parametrov in izbiro različnih metod pri prevajanju, kjer je od lastnosti obeh jezikov in razmerja med njima odvisno, katere dajo najboljše rezultate. Tudi za slovenščino so bile metode SSP že uporabljene (Vičič, 2002; Sepesy Maučec et al., 2006; Brest, 2009; Verdonik, 2013; Sepesy Maučec et al., 2013; Dugonik, 2013).

Predlagamo preprosto metodo, ki temelji na SSP, a s spremenjenim načinom prevajanja odpravlja zgoraj opisane pomanjkljivosti. Pri eksperimentih smo uporabili Moses (Koehn et al., 2007), najpogosteje uporabljani odprtokodni sistem za strojno prevajanje.

Medtem ko je v klasičnem pristopu k strojnemu prevajanju osnovna enota beseda (pojavnica) oz. besedna zveza (SSP-B), pa statistično prevajanje na ravni znakov (SSP-Z) deluje tako, da znake obravnava, kot bi bili besede, potrebne prevoda. Na praktični ravni to pomeni, da besedilo prilagodimo za obdelavo v Mosesu enostavno tako, da med posamezne črke vstavimo presledke, nekdanje presledke (meje med besedami) pa označimo s posebnim znakom, npr. podčrtajem. Pomanjkljivost SSP-B je tudi ta, da zna prevajati samo besede oz. besedne zveze, ki so bile vključene v učno množico. SSP-Z sicer prevaja na enak način, a ker je nabor znakov veliko manjša končna množica kot nabor besed posameznega jezika, bodo v veliki večini že zastopani v učnem modelu in jih bo tako sistem znal prevesti. Kot rečeno, je tovrstni znakovni pristop že uveljavljen način za prevajanje zelo sorodnih jezikov oz. jezikovnih različic, kjer so razlike večinoma omejene na ortografsko raven (Vilar, 2007; Nakov in Tiedemann, 2012; Scherrer in Erjavec, 2013; Pettersson et al., 2013; Sánchez-Martínez et al., 2013).

Kar loči našo metodo od obstoječih pristopov, je način prevajanja znakov. Običajni način prevaja posamezne črke (znake), iz katerih je sestavljena beseda (pojavnica), ne upošteva pa ko(n)teksta. Beseda je namreč zanj zaključena enota, kot bi bil v klasičnem prevajanju stavek, zato ne sega prek besedne meje. Naša metoda poskuša preseči to omejitev, zato kot »stavek« ne vzame besede, temveč širšo besedilno enoto. V našem primeru smo se odločili, da bodo to posamezne vrstice, saj so te praviloma označene (in s tem poravnane) tako v diplomatičnem kot v kritičnem prepisu.

Metoda se opira na predpostavko, da ima del diplomatičnega prepisa že izdelan svoj kritični prepis oz. »prevod«, ta vzporedni korpus pa je nato uporabljen kot učna množica, na kateri se bo učil prevajalski model. Ciljni jezikovni model je naučen na že izdelanem delu kritičnega prepisa, s predpostavko, da bo deloval bolje brez opiranja na zunanje vire, kot sta korpus in/ali leksikon sodobnega jezika.

3.2 Slomškovi pridigi kot učna in testna množica

Da bi preizkusili delovanje metode, smo izvedli niz eksperimentov na podatkovni množici, vzeti iz digitalne znanstvenokritične izdaje »Treh pridig o jeziku« (Faganel et al., 2004), delo Antona Martina Slomška (1800–1862), znanega slovenskega škofa, pedagoga in pisatelja, pa tudi pomembnega reformatorja slovenskega kulturnega, narodnostnega in verskega življenja – zlasti v vzhodni Sloveniji in na Koroškem –, ki je opozarjal na vpliv agresivne germanizacije in skušal s svojo dejavnostjo zmanjšati ta vpliv (Erjavec in Ogrin, 2004).

Digitalna izdaja je del širše zbirke, imenovane *Elektronske znanstvenokritične izdaje slovenskega slovstva* (eZISS, http://nl.ijs.si/e-zrc/), ki se je kot skupni projekt začela graditi leta 2001 pod vodstvom Matije Ogrina z Inštituta za slovensko literaturo in literarne vede ZRC SAZU ter Tomaža Erjavca z Odseka za tehnologije znanja na IJS, avtor obeh vrst prepisov Slomškovih pridig pa je Jože Faganel.

Vse digitalne izdaje eZISS so zapisane v skladu s smernicami za kodiranje besedil TEI (Konzorcij TEI), torej v zapisu XML, usklajenem s shemo, ki je parametrizacija smernic TEI za namene projekta eZISS.

Izdaja Slomškovih del je v zbirki zastopana s tremi pridigami kot primeri retorske proze v prvi polovici 19. stoletja, ki še posebno simbolizirajo njegovo delo in prizadevanja za širšo rabo slovenskega jezika. Gre za naslednje pridige:

1) »Za krščansko govorjenje« (1825);
2) »Jezik je vir dobrega in zla« (1829);
3) »Svoj jezik je treba spoštovati« (1838).

Prvi dve pridigi sta ohranjeni v avtorjevem rokopisu, rokopis tretje, ki je najbolj znana, pa je izgubljen, a je bilo besedilo pridige kmalu po nastanku dvakrat natisnjeno. Elektronska izdaja vsebuje predgovor, faksimile, diplomatični prepis (razen za 3. pridigo), 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.

Ker za tretjo pridigo ni na voljo diplomatičnega prepisa, smo jo izločili iz eksperimenta. Prvi dve pridigi sta nastali v približno istem časovnem obdobju, ko se v slovenskem jeziku še ni oblikoval standardni zapis besed, niti se še ni začel ključni proces poenotenja zapisa, saj se je to zgodilo šele sredi stoletja s t. i. novimi oblikami.

Za ponazoritev jezika, ki ga je uporabljal Slomšek, in njegovega kritičnega prepisa poda slika 1 tri vrstice iz prve pridige.

Diplomatični prepis	Kritični prepis
3.She enkrat bel ſe' veſeli duſha kerſhanſka! Kader ſvete	3.Še enkrat belj se vesêli duša krščanska! Kader svete
godove ſvetnikov obhajaſh, s'akaj ravno oni ſo tebi taji‸	godove svetnikov obhajaš, zakaj ravno oni so tebi taji-
ſte ſvetle s'ves'de, ki ſe is'ſ. nebeſ na tebe os'irajo, tebi	ste svetle zvezde, ki se iz sv. Nebes na tebe ozirajo, tebi

Slika 1: Primer diplomatičnega in kritičnega prepisa pridige A. M. Slomška v izdaji eZISS.

Primer pokaže, da razlike med prepisoma po eni strani niso omejene samo na prečrkovanje bohoričice v gajico, pač pa zajemajo tudi druge premene v besedah (npr. *kerſhanſka (→ keršanska) → krščanska*) kot tudi spremembe ločil (npr. ‸ → -), po drugi strani pa ponazarja, da kritični prepis ne uporablja sodobnega standarda, pač pa v mnogo primerih še vedno ohrani starinsko obarvani zapis besed (*Kader, tajiste*). Dodatno je pomembno, da kritični prepis ohrani deljene besede na koncu vrstic, kar je tudi problematično za klasične metode, ki se opirajo na posodabljanje posameznih besed.

Za preizkus predlagane metode smo izdelali učni in testni korpus. Da bi bila evalvacija bolj realistična, nismo premešali vrstic obeh pridig, pač pa smo kot učni korpus vzeli prvo pridigo, kot testnega pa drugo.

Iz izvornega zapisa TEI smo odstranili vse oznake XML (npr. poudarjeno besedilo) ter upoštevali avtorjeve popravke besedila. S tem smo dobili dva vzporedna korpusa, ki sta poravnana po vrsticah in vsebujeta samo golo besedilo, točno tako, kot je predstavljeno v sliki 1.

Velikost učnega in testnega korpusa je podana v tabeli 1, ki pokaže, da je testna pridiga več kot za tretjino večja od učne, pri čemer prva vsebuje približno 2.000 besed, druga pa 3.300.

	Učna pridiga (1825)		*Testna pridiga (1829)*	
Prepis:	*Dipl.*	*Krit.*	*Dipl.*	*Krit.*
Vrstic:	252	252	362	362
Pojavnic:	2.036	2.081	3.393	3.346
Znakov:	13.958	12.809	23.353	20.732
DR-Z	20,68		22,26	

Tabela 1: Velikost uporabljene podatkovne množice in delež razlike na ravni znakov med dipl. in krit. prepisom.

Za evalvacijo bomo uporabili delež razlik na ravni znakov (DR-Z), ki meri povprečno Levenshteinovo razdaljo (razliko) med avtomatsko generiranim kritičnim prepisom in ročno izdelanim kritičnim prepisom pridige, uporabljene kot testne množice. Razdalja se meri v številu t. i. posegov – kamor spadajo vstavljanje, brisanje ali zamenjava znaka (Levenshtein, 1966) – ki je potrebno, da izhodiščno besedilo »popravimo« v ciljnega. Razmerje med številom posegov in številom znakov v referenčnem (ciljnem) besedilu izrazimo v odstotnih deležih, ki bodo naša evalvacijska mera učinkovitosti delovanja posamezne metode.

Kot izhodiščno mero evalvacije, ki jo bomo skušali izboljšati, smo vzeli razdaljo med diplomatičnim prepisom, če v njega ne posegamo z nobeno izmed metod (in ga tako obravnavamo kar kot neke vrste strojno generirani kritični prepis), ter med ročno izdelanim kritičnim prepisom. Kot pokaže tabela 1, delež razlik v znakih med njima v testnem korpusu znaša 22,62 odstotka, kar je nekaj več kot v učni množici, kjer je 20,68 odstotka.

3.3 Opis eksperimenta

Preizkusili smo tri metode:

- *statistično strojno prevajanje na ravni znakov (SSP-Z)*, pri čemer je prevodna enota ena vrstica;
- *statistično strojno prevajanje na ravni besed* (SSP-B) in
- *kombinacijo več normalizacijskih metod v Normi.*

Normo (Bollmann et al., 2012) smo se odločili uporabiti za primerjavo s SSP, ker gre za eno bolj znanih odprtokodnih orodij za normalizacijo. Orodje združuje več metod normalizacije, vendar pa je sestavljeno iz različnih zunanjih modulov, imenovanih »normalizatorji«, ki predstavljajo različne metode normalizacije (mdr. avtomatsko naučena pravila, Levenshtein, historični leksikon) in jih je mogoče uporabiti ali izključiti iz procesne verige. Privzete nastavitve za svoje delovanje potrebujejo dvojezični historični leksikon (diplomatična in njena kritičnoprepisna besedna oblika) ter enojezični leksikon sodobnega jezika.

Pri obeh metodah strojnega prevajanja smo preizkusili tri različice modelov ciljnega jezika, in sicer smo kot učne modele vključili:

A. zgolj besedilo pridige, uporabljene za učno množico;
B. poleg A tudi kritične prepise avtorjevih drugih del (tj. zbrana dela A. M. Slomška);
C. poleg A tudi jos100k (Erjavec et al., 2010), referenčni korpus za jezikoslovno označevanje slovenskega jezika.

Poleg različic jezikovnega modela smo pri strojnem prevajanju preizkusili tudi različne stopnje jezikovnega modela, t. i. n-grame. Ti zaznamujemo dolžino enot, ki jih sistem obravnava kot potencialne besedne zveze, kar v našem primeru pomeni dolžino niza zaporednih znakov (črk, cifer, ločil ipd.). Dolžino smo omejili v razponu od dveh do petih zaporednih znakov (tj. 2-, 3-, 4- in 5-grami), saj so preizkusi modelov z enim znakom in več kot petimi pokazali, da ti delujejo slabše kot modeli v izbranem razponu.

4 Rezultati

Rezultati, podani v tabeli 2, pokažejo, da nam izhodiščni delež razlik uspe zmanjšati z vsemi tremi metodami, tako z obema vrstama strojnega prevajanja kot z Normo. Najboljši rezultat (DR-Z = 7,59 %) je dosegla predlagana metoda, tj. SSP-Z z osnovnimi nastavitvami sistema Moses, najpreprostejšim modelom ciljnega jezika (zgolj besedilo učne pridige) in najnižjo stopnjo modela (2-grami). V obravnavanem primeru to pomeni, da če bi si prepisovalec pomagal s SSP-Z, bi si prihranil dve tretjini dela, kar zadeva popravke na ravni posameznih znakov, v primerjavi s tem, da bi za izhodišče pri izdelavi kritičnega prepisa druge pridige vzel kar diplomatični prepis in ga nato popravljal povsem ročno.

Medtem ko z najosnovnejšim in najkompleksnejšim jezikovnim modelom (A in C) najbolje delujejo 2-grami kot osnovna enota prevajanja, pa se pri učnem modelu s širšim naborom Slomškovih del kot najučinkovitejši kažejo nizi treh zaporednih znakov, daljši nizi od treh znakov pa v vseh primerih dosegajo slabše rezultate, kar nakazuje na to, da je večina sprememb med diplomatičnim in kritičnim prepisom omejena na razpon treh znakov. Če upoštevamo, da je najbolj tipični uredniški poseg posodobitev črkopisa, je rezultat mogoče pojasniti s tem, da je tipična sprememba bohoričice v gajico omejena na zamenjavo največ dveh grafemov z enim (npr. *zhaft* > *čast*). Če primerjamo posamezne jezikovne modele, ugotovimo, da je pri vseh tipih n-gramov najslabši model C, kar kaže na to, da sodobni jezik kot model ciljnega jezika ni primeren za kritični prepis, saj ta ne stremi k izrazni podobi sodobnega jezika.

Da bi določili, ali je za generiranje strojnega prepisa zares bolj smiselno uporabiti prilagojeno metodo strojnega prevajanja, tj. prevajanje znakov, smo preizkusili tudi, kakšne rezultate daje strojno prevajanje na ravni besed (SSP-B). Generirani prepis smo zaradi primerljivosti evalvirali na enak način, tj. glede na delež razlik v znakih, in ne kot odstotek napačno prevedenih besed, kar je običajna metoda evalvacije klasičnih strojnoprevajalskih sistemov. Eksperiment je pokazal, da klasična metoda za obravnavani primer ni ustrezna, saj deluje za polovico slabše kot SSP-Z: medtem ko prevajanje znakov prihrani do dve tretjini dela, ga prevajanje besed samo tretjino. Spreminjanje stopnje jezikovnega modela je imelo pri SSP-B zanemarljiv vpliv na rezultat, saj je ta pri vseh n-gramih praktično enak; izjema je nekoliko boljši rezultat 2-gramov, kar potrjujejo že izsledki pri SSP-Z, le da gre tu za niz dveh besed in ne znakov. Poleg tega osnovna enota prevajanja ni bila več celotna vrstica, temveč smo besednim oblikam iz diplomatičnega prepisa samodejno določili njihove kritičnoprepisne ustreznice s pomočjo Gize++, orodja za poravnavo vzporednih prevodov.

S pomočjo poravnavnih tabel Gize (t. i. datoteka e2f) smo na ta način tudi strojno izdelali dvojezični leksikon in ga skupaj z leksikonom sodobnih besednih oblik Sloleks (Dobrovoljc et al., 2015) dodali v Normo, sicer pa uporabili privzete nastavitve. Orodje deluje za skoraj tri odstotne točke bolje od SSP-B, a še vedno skoraj šest odstotnih točk slabše kot SSP-Z.

	2-grami	*3-grami*	*4-grami*	*5-grami*
SSP-Z, model A	**7,59**	8,58	8,71	9,18
SSP-Z, model B	8,05	**8,14**	8,21	8,80
SSP-Z, model C	**8,26**	8,45	8,48	9,50
SSP-B	**16,84**	16,87	16,86	16,87
Norma:	14,19			
Izhodišče:	22,26			

Tabela 2: Odstotni delež razlik v znakih glede na različico in stopnjo modela ciljnega jezika na ravni znakov oz. besed ter prim. z Normo in izhodiščem.

5 Sklepne misli in prihodnje delo

Preizkus strojnega generiranja kritičnega prepisa je pokazal, da je mogoče že s privzetimi nastavitvami odprtokodnega sistema za strojno prevajanje Moses ter preprostim prevajalskim in jezikovnim modelom, delujočim na ravni znakovnih nizov, uredniku izdaje prihraniti do dve tretjini dela, če ga ovrednotimo kot število potrebnih posegov v besedilo.

Metodo je mogoče uporabiti tudi v primerih, ko ima sorazmerno majhen del diplomatičnega prepisa že izdelan svoj kritični prepis, saj metoda SSP-Z za svoje delovanje ne potrebuje velike učne množice, kar dokazuje tudi obravnavani primer relativno kratkega pridižnega besedila, kjer SSP-B deluje za polovico slabše, le nekoliko boljša je kombinacija »neprevajalskih« metod v Normi.

V širšem smislu bi izhodno besedilo metode SSP-Z lahko uporabili tudi za preučevanje sistemskih in nesistemskih posegov v diplomatični prepis pri »posodabljanju« za kritični prepis, torej postopkov na ortografski ravni, uporabljenih za približanje razumevanja besedila sodobnemu bralcu. Vpogled v sistemskost bi bil uporaben tudi za izboljšanje kakovosti procesiranja starejših besedil z avtomatskimi orodji za jeziko(slov)no označevanje. Pri nadaljnjem razvoju metode nameravamo preizkusiti metodo na novih primerih uporabe in dodatnih možnostih združevanja učnih modelov v okviru orodja Moses, nato pa uporabiti še nekoliko zahtevnejše nastavitve parametrov, ki jih omogoča Moses, kot je npr. uglaševanje parametrov namesto privzetih nastavitev.

6 Zahvala

Avtorja se zahvaljujeta anonimnim recenzentom za koristne pripombe. Raziskava, opisana v prispevku, je bila opravljena v okviru raziskovalnega programa Tehnologije znanja P2-0103 in programa Mladi raziskovalci (št. 37487), ki ju financira ARRS.

7 Literatura

Marcel Bollmann. 2012. (Semi-)Automatic Normalization of Historical Texts using Distance Measures and the Norma tool. V: *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*, str. 3–14, Lizbona, Portugalska. https://www.linguistics.ruhr-uni-bochum.de/comphist/pub/acrh12.pdf.

Janez Brest. 2009. Statistično strojno prevajanje iz slovenščine v angleščino. V: *Zbornik povzetkov delavnic "Algoritmi po vzorih iz narave" v študijskem letu 2008/2009*, str. 15, Ljubljana, Slovenija.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, Robert L. Mercer. (1993) The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics,* 19(2): 263–311. MIT Press.

Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec in Miro Romih. 2015. *Morfološki leksikon Sloleks 1.2*. Slovenski repozitorij jezikovnih virov CLARIN.SI. http://hdl.handle.net/11356/1039.

Marijan Dović. 2006. Tekstualna tradicija in elektronski medij: od digitalne slikovne reprodukcije do znanstvenokritične izdaje. V: *Sborník prací Filozofické fakulty brněnské univerzity, Studia minora Facultatis philosophicae Universitatis Brunensis*. 9: 208–215.

https://digilib.phil.muni.cz/bitstream/handle/11222.digilib/102980/X_SlavicaLitteraria_09-2006-1_25.pdf?sequence=1.

Jani Dugonik. 2013. *Uglaševanje parametrov pri statističnem strojnem prevajanju*. Magistrsko delo, Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru. http://dkum.uni-mb.si/IzpisGradiva.php?id=40979.

Tomaž Erjavec in Matija Ogrin. 2004. E-Slomšek: elektronska znanstvenokritična izdaja retorske proze 19. stoletja po standardu XML TEI. V: *Jezikovne tehnologije: zbornik B*, str. 87–93. http://nl.ijs.si/e-zrc/bib/sdjt04-16erjavec.pdf.

Tomaž Erjavec. 2015. The IMP historical Slovene language resources. *Language resources and evaluation*, 49(3): 753–775. doi: 10.1007/s10579-015-9294-7.

Tomaž Erjavec, Darja Fišer, Simon Krek in Nina Ledinek, 2010. The JOS Linguistically Tagged Corpus of Slovene. V: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Malta. http://www.lrec-conf.org/proceedings/lrec2010/pdf/139_Paper.pdf.

Jože Faganel, Matija Ogrin in Tomaž Erjavec. 2004. Anton Martin Slomšek: Tri pridige o jeziku: elektronska znanstvenokritična izdaja. Inštitut za slovensko literaturo in literarne vede, ZRC SAZU, Ljubljana. http://nl.ijs.si/e-zrc/slomsek/.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, Praga, Češka.

Philipp Koehn. 2010. Statistical Machine Translation. Cambridge University Press.

Konzorcij TEI (ur.). *Guidelines for Electronic Text Encoding and Interchange*. http://www.tei-c.org/P5/.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8): 707–710.

Preslav Nakov in Jörg Tiedemann. 2012. Combining word-level and character-level models for machine translation between closely-related languages. V: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, str. 301–305, Stroudsburg, Pennsylvania, ZDA. Association for Computational Linguistics. http://anthology.aclweb.org/P/P12/P12-2.pdf#page=329.

Eva Pettersson, Beáta Megyesi in Jörg Tiedemann. 2013. An SMT approach to automatic annotation of historical text. V: *Proceedings of the Workshop on Computational Historical Linguistics at NODALIDA 2013, NEALT Proceedings Series*, 18: 54–69. http://www.ep.liu.se/ecp/087/005/ecp1387005.pdf.

Michael Piotrowski. 2012. *Natural language processing for historical texts* (Synthesis Lectures on Human Language Technologies, ur. Graeme Hirst, vol. 17). Morgan & Claypool Publishers.

Felipe Sánchez-Martínez, Isabel Martínez-Sempere, Xavier Ivars-Ribes in Rafael C. Carrasco. 2013. An open diachronic corpus of historical Spanish: annotation criteria and automatic modernisation of spelling. ArXiv:1306.3692. http://www.dlsi.ua.es/~fsanchez/pub/pdf/sanchez-martinez13a.pdf.

Yves Scherrer in Tomaž Erjavec. 2013. Modernizing historical Slovene words with character-based SMT. V: *BSNLP 2013-4th Biennial Workshop on Balto-Slavic Natural Language Processing*, Sofija, Bolgarija. https://halshs.archives-ouvertes.fr/hal-00838575/document.

Yves Scherrer in Tomaž Erjavec. 2015. Modernising historical Slovene words. *Natural Language Engineering*, doi: 10.1017/S1351324915000236.

Mirjam Sepesy Maučec, Janez Brest in Zdravko Kačič. 2006. Statistical alignment models in machine translation from Slovenian to English. *Elektrotehniški vestnik*, 73(5): 273–78. http://www.dlib.si/details/URN:NBN:SI:DOC-GDCH7YIE.

Mirjam Sepesy Maučec, Gregor Donaj in Zdravko Kačič. 2013. Improving statistical machine translation with additional language models. V: *Human language technologies as a challenge for computer science and linguistics: proceedings / 6th Language & Technology Conference*, Poznanj, Poljska.

Jorg Tiedemann. 2009. Character-based PSMT for closely related languages. V: *Proceedings of the 13th Annual Conference of the European Association for Machine Translation (EAMT)*, str. 12–19, Barcelona, Španija. http://stp.lingfil.uu.se/~joerg/published/eamt09_related.pdf.

Darinka Verdonik. 2013. Strojno prevajanje z Mosesom. *Življenje in tehnika*, 64(7/8): 48–64.

Jernej Vičič in Tomaž Erjavec. 2002. Statistično strojno prevajanje na osnovi vzporednih korpusov. V: *Zbornik enajste mednarodne Elektrotehniške in računalniške konference ERK 2002*, str. 217–220, Portorož, Slovenija.

David Vilar, Jan-T. Peter in Hermann Ney. 2007. Can we translate letters? V: *Proceedings of the Second Workshop on Statistical Machine Translation*, str. 33–39. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1626360.