Sveučilište u Zagrebu

Filozofski fakultet

Ivan Dunđer

SUSTAV ZA STATISTIČKO STROJNO PREVOĐENJE I RAČUNALNA ADAPTACIJA DOMENE

DOKTORSKI RAD

Zagreb, 2015.

Sveučilište u Zagrebu

Filozofski fakultet

Ivan Dunđer

SUSTAV ZA STATISTIČKO STROJNO PREVOĐENJE I RAČUNALNA ADAPTACIJA DOMENE

DOKTORSKI RAD

Mentor:
prof. dr. sc. Sanja Seljan

Zagreb, 2015.

University of Zagreb
Faculty of Humanities and Social Sciences

Ivan Dunđer

STATISTICAL MACHINE TRANSLATION SYSTEM AND COMPUTATIONAL DOMAIN ADAPTATION

DOCTORAL THESIS

Supervisor:
Full Professor Sanja Seljan, PhD

Zagreb, 2015

…supruzi Ivani,
mami Jeleni, tati Josipu, sestri Nikolini

Sažetak

Statističko strojno prevođenje temeljeno na frazama jedan je od mogućih pristupa automatskom strojnom prevođenju. U radu su predložene metode za poboljšanje kvalitete strojnog prijevoda prilagodbom određenih parametara u modelu sustava za statističko strojno prevođenje. Ideja rada bila jest izgraditi sustave za statističko strojno prevođenje temeljeno na frazama za hrvatski i engleski jezik. Sustavi su trenirani za dva jezična smjera, na dvije domene, na paralelnim korpusima različitih veličina i obilježja za hrvatsko-engleski i englesko-hrvatski jezični par, nakon čega proveden postupak ugađanja sustava. Istraženi su hibridni sustavi koji objedinjuju značajke obiju domena. Time je ispitan izravan utjecaj adaptacije domene na kvalitetu automatskog strojnog prijevoda hrvatskog jezika, a nova saznanja mogu koristiti pri izgradnji novih sustava. Provedena je automatska i ljudska evaluacija (vrednovanje) strojnih prijevoda, a dobiveni rezultati uspoređeni su s rezultatima strojnih prijevoda dobivenih primjenom postojećih web servisa za statističko strojno prevođenje.

Abstract

Phrase-based statistical machine translation is one of possible automatic machine translation approaches. This work proposes methods for increasing the quality of machine translation by adapting certain parameters in the statistical machine translation model. The idea was to build phrase-based statistical machine translation systems for Croatian and English language. The systems were be trained for two directions, on two domains, on parallel corpora of different sizes and characteristics for Croatian-English and English-Croatian language pair, after which the tuning procedure was conducted. Afterwards, hybrid systems which combine features of both domains were investigated. Thereby the direct impact of domain adaptation on the quality of automatic machine translation of Croatian language was explored, whereas new findings can be utilised for building new systems. Automatic and human evaluation of machine translations were carried out, while obtained results were compared with results obtained from applying existing statistical machine translation web services.

Prošireni sažetak

Višejezična komunikacija jedan je od najvažnijih prioriteta u današnjem globaliziranom svijetu. Budući da je ljudsko prevođenje vremenski intenzivan, skup i neefikasan način osiguravanja dostupnosti informacija na stranom jeziku, jedno o mogućih rješenja tog problema jest primjena paradigmi automatskog strojnog prevođenja koje se uvelike oslanjaju na saznanja koja proizlaze iz informacijskih znanosti. Sustavi za automatsko strojno prevođenje danas imaju sve raširenu primjenu u svakodnevnoj komunikaciji, poslovnom i znanstveno-istraživačkom svijetu te postoje za šire govorene jezike, dok se za manje govorene jezike takvi sustavi rjeđe razvijaju. Statističko strojno prevođenje temeljeno na frazama jedan je od mogućih pristupa automatskom strojnom prevođenju. Takav sustav segmentira izvorne rečenice u fraze, prevodi svaku frazu, premješta ih ukoliko je to potrebno te od tih prijevoda fraza sastavlja rečenice u ciljnom jeziku. Treba pritom naglasiti kako se ne radi nužno o lingvističkim frazama, već o skupu proizvoljnih riječi koje s određenom vjerojatnošću čine odgovarajuće nizove riječi. Ovaj pristup strojnom prevođenju uvelike se oslanja na količinu i kvalitetu paralelnih korpusa, pri čemu heterogenost podatkovnih skupova za treniranje i ugađanje ima znatan utjecaj na kvalitetu strojnih prijevoda, pogotovo kada se radi o morfološki bogatim jezicima kao što je hrvatski.

U ovom istraživanju odgovoreno je na pitanje koliko su dobri vlastiti hrvatsko-engleski sustavi za statističko strojno prevođenje razvijeni u sklopu ovog doktorskog rada za područje općenite domene i područje vezano uz računalni softver. Provedena je evaluacija kvalitete strojnih prijevoda pomoću automatskih metrika i ljudske prosudbe. Analizirano je kakvi su novoizgrađeni sustavi u usporedbi s postojećim online servisima za strojno prevođenje. Provedeno je ispitivanje utjecaja relativno malenih podatkovnih skupova korištenih u ovom istraživanju te njihova uloga u izgradnji dobrih sustava za strojno prevođenje za hrvatski i engleski jezik. Istraženo je da li tehnike adaptacije domene mogu poboljšati performanse sustava za hrvatsko-engleski jezični par. U ovoj doktorskoj disertaciji predložene su metode

za povećanje kvalitete automatskog strojnog prijevoda pomoću prilagodbe određenih parametara u modelu sustava za statističko strojno prevođenje.

Izgrađeno je ukupno osam sustava za statističko strojno prevođenje temeljeno na frazama: četiri za hrvatsko-engleski smjer te četiri za englesko-hrvatski smjer. Sustavi su trenirani za oba smjera, na dvije domene, s paralelnim korpusima različitih veličina i karakteristika, nakon čega je izvršen postupak ugađanja modela. Zatim su istraženi hibridni sustavi, koji kombiniraju značajke iz obje domene te time modificiraju logiku modela statističkog strojnog prevođenja.

Utjecaj adaptacije domene na kvalitetu automatskih strojnih prijevoda za hrvatsko-engleski jezični par time je istražen, a nova saznanja su iskorištena pri izgradnji novih sustava. Evaluacija strojnog prijevoda izvršena je i za općenitu domenu i za domenu računalnog softvera. Provedena je automatska i ljudska evaluacija strojnih prijevoda, a generirani strojni prijevodi uspoređeni su s rezultatima prikupljenim za vrijeme primjene postojećih online servisa za strojno prevođenje. Statistička značajnost evaluacijskih rezultata također je analizirana.

Ključne riječi: statističko strojno prevođenje, adaptacija domene, automatska evaluacija kvalitete strojnog prijevoda, ljudska evaluacija, rangiranje sustava za strojno prevođenje, računalna obrada prirodnog jezika, jezične tehnologije, informacijske znanosti

Extended abstract

Multilingual communication has become a top priority in today's globalised world. As human translation is a time-consuming, expensive and non-efficient way of satisfying the availability of information in a foreign language, one of the possible solutions to this problem is to apply the paradigms of automatic machine translation which heavily rely on the findings that stem from information sciences. Automatic machine translation systems today are widely used in everyday communication, in the world of business, science and research, and exist for widely spoken languages, while for less spoken languages such systems are less developed. Phrase-based statistical machine translation is one of the possible automatic machine translation approaches. Such a system segments source sentences into phrases, translates each phrase, reorders phrases if needed and composes sentences from these phrase translations in the target language. It should be emphasised that it is not necessarily about linguistic phrases, but a set of arbitrary words that with a certain probability constitute corresponding sequences of words. This machine translation approach relies heavily on the amount and quality of parallel corpora, whereas heterogeneity of training and tuning datasets has a major impact on the quality of machine translations, especially when it comes to morphologically rich languages, such as Croatian.

In this research the question on how good the own Croatian-English statistical machine translation systems developed in this doctoral research for the general domain and the field of computer software are, was answered. Evaluation of the quality of machine translations in terms of automatic metrics and human judgment was carried out. The performance of newly built systems when compared to existing online machine translation services was analysed. An examination on the impact of relatively small datasets used in this research and their role in building well-performing statistical machine translation systems for the Croatian and English language was conducted. If domain adaptation techniques can improve system performance for the Croatian-English language pair, was also investigated. This doctoral

thesis proposes methods for increasing quality of automatic machine translation by adapting certain parameters in the statistical machine translation system model.

Eight phrase-based statistical machine translation systems were built in total, four for the Croatian-English direction and four for the English-Croatian direction. The systems were trained for both directions, on two domains, with parallel corpora of different sizes and characteristics, after which the model tuning procedure was conducted. Afterwards, hybrid systems, which combine features of both domains and therefore modify the statistical machine translation logic, were investigated.

The impact of domain adaptation on the quality of automatic machine translation of the Croatian-English language pair was thereby explored, whereas new findings were utilised for building new systems. Machine translation evaluation trials were conducted on both general and computer software domain. Automatic and human evaluation of machine translations were carried out, while generated machine translations were compared with results obtained from applying existing online machine translation services. Statistical significance of the evaluation results was also analysed.

Keywords: statistical machine translation, domain adaptation, automatic evaluation of machine translation quality, human evaluation, machine translation system ranking, computational natural language processing, language technologies, information sciences

Sadržaj

Popis slika

Popis tablica

1. UVOD

Informacija predstavlja najvrjedniji resurs u informacijskom dobu. No, informacija mora biti na raspolaganju pravovremeno, stoga i ne čudi što su dostupnost informacije, mogućnosti pretraživanja i indeksiranja informacija pa čak i na stranom jeziku neki od imperativa suvremenog poslovanja. U današnjem globaliziranom svijetu povećava se potreba za višejezičnom komunikacijom. Naime, bez komunikacije, bila ona interna ili ona s klijentima i partnerima, poslovanje se ne može odvijati (Dillinger, 2010), a s obzirom da se danas u svijetu govori više od 7000 jezika (Stüker i Waibel, 2008), ne iznenađuje potreba za raznim jezičnim tehnologijama. Takve tehnologije omogućuju učinkovitije upravljanje resursima, kvalitetniju razmjenu znanja te „recikliranje" već prevedenih dokumenata. Pored toga, one nastoje povećati konzistentnost prijevoda i efikasnost rada te time smanjuju troškove prevođenja. Iako uvođenje jezičnih tehnologija zahtijeva određene financijske izdatke, njihovom se kvalitetnom implementacijom uložena sredstva relativno brzo mogu povratiti (eng. *return-on-investment, ROI*) (Dillinger i Marciano, 2012; Dillinger, 2010).

Jedan od mogućih pristupa osiguravanja informacija na stranom jeziku, posebno važan za manje govorene jezike, jest primjena sustava za automatsko strojno prevođenje. Pod pojmom strojno prevođenje (eng. *machine translation*) podrazumijeva se primjena računala za automatiziranje (dijela) procesa prevođenja s jednog jezika na drugi (Folajimi i Omonayin, 2012). Cilj sustava za strojno prevođenje je što brže računalno generirati velik broj prijevoda prihvatljive kvalitete uz minimalan trošak.

Razvoj takvih sustava od izuzetne je važnosti i za uključivanje Republike Hrvatske u međunarodne znanstveno-istraživačke projekte, za akademsku suradnju, za svakodnevno poslovanje te napredak u gospodarstvu i industriji (Seljan i Pavuna, 2006). Međutim, izgradnja i evaluacija, tj. vrednovanje takvih sustava vrlo su zahtjevni zadatci (Unnikrishnan et al., 2010), što će biti detaljno opisano u narednim poglavljima.

S obzirom na stupnjeve automatizacije i uključenost čovjeka u proces prevođenja, prevođenje se može svrstati u jednu od tri kategorije (Calude, 2004):

- ljudsko prevođenje (eng. *human translation, HT*)
- ljudsko prevođenje potpomognuto strojem, tj. računalom (eng. *machine-assisted human translation, MAHT*)
- strojno prevođenje potpomognuto čovjekom (eng. *human-assisted machine translation, HAMT*)
- automatsko strojno prevođenje (eng. *fully automatic machine translation, FAMT*)

Ljudsko prevođenje potpomognuto strojem (računalom) i strojno prevođenje potpomognuto čovjekom nazivaju se i računalno-potpomognuto prevođenje (eng. *computer-assisted translation, CAT*). Automatsko strojno prevođenje može se zasnivati na jezičnim formalizmima, tj. gramatičkim pravilima (eng. *rule-based machine translation*, *RBMT*) ili na empirijskim opažanjima (España-Bonet i Gonzàlez, 2014). Pod sintagmom „empirijska opažanja" podrazumijeva se ideja prema kojoj računalo može naučiti pravila strojnog prevođenja na temelju velike količine dvojezičnih tekstova (eng. *bitexts*) (Koehn, 2010). To su redovito reprezentativni podatkovni skupovi koji se sastoje od ljudski prevedenih tekstova visoke kvalitete i pripadaju određenoj domeni, kao što su npr. politika, sport, meteorologija, specifikacija bijele tehnike ili računalnih komponenti itd.

Pristup „empirijsko opažanje" naziva se i pristup upravljan podatcima (eng. *data-driven approach*), s obzirom da se svi procesi temelje na velikoj količini podataka. Strojno prevođenje koje se temelji na empirijskim opažanjima može se nadalje podijeliti u dvije kategorije (España-Bonet i Gonzàlez, 2014): statističko strojno prevođenje (eng. *statistical machine translation*, *SMT*) i strojno prevođenje temeljeno na primjerima (eng. *example-based machine translation*, *EBMT*).

Strojno prevođenje koje se zasniva na empirijskim opažanjima nastalo je kao posljedica nezadovoljstva strojnim prevođenjem koje se temelji na jezičnim formalizmima. Naime, pristup zasnovan na jezičnim formalizmima nije dao željene rezultate (Way i Hassan, 2009), a pojava velike količine dostupnih i računalno čitljivih tekstualnih korpusa ubrzala je razvoj novih paradigmi u strojnom prevođenju.

Statističko strojno prevođenje temelji se upravo na velikoj količini sravnjenih paralelnih korpusa (Koehn, 2010), zbog čega se svrstava u kategoriju korpusno-temeljenog strojnog prevođenja (eng. *corpus-based machine translation*) (Tripathi i Sarkhel, 2010). U statističkom strojnom prevođenju prijevodi se generiraju pomoću statističkih modela čiji parametri izravno proizlaze iz karakteristika paralelnih korpusa. Drugim riječima, budući da kvaliteta paralelnih korpusa izravno

utječe na kvalitetu automatskog strojnog prijevoda, tj. na performanse sustava za strojno prevođenje, očita je potreba za izgradnjom kvalitetnih i opsežnih paralelnih korpusa, što međutim, predstavlja dugotrajan i zahtjevan proces (Callison-Burch i Osborne, 2003). Načelno vrijedi, što su podatkovni skupovi koji se koriste pri izgradnji sustava za strojno prevođenje kvalitetniji i reprezentativniji, to će i strojni prijevodi u konačnici biti točniji i precizniji (Koehn i Haddow, 2012). Primjer kvalitetnog dvojezičnog, ljudski prevedenog, teksta u domeni politike dan je u nastavku (isječak iz SETimes korpusa):

The event was highly relevant for several reasons. The timing agreed on will be respected. EU funds should be included in the multiannual budgeting of finance ministries. Other Balkan countries are also banking on increased financial assistance from the EU for their development efforts. Economic stagnation in the EU fuels protectionist tendencies, impeding a successful outcome to the Doha trade round. To be clear, the word „corrupt" doesn't exist for this parliament.	Ovaj je događaj iznimno znakovit iz nekoliko razloga. Dogovoreni će rokovi biti ispoštovani. Sredstva EU morala bi biti uključena u višegodišnje nacrte budžeta ministarstava financija. Ostale balkanske zemlje također računaju na povećanu financijsku potporu EU njihovom razvitku. Gospodarska stagnacija u EU potpiruje protekcionističke sklonosti, sprječavajući uspješan ishod trgovačkih pregovora u Dohi. Ukratko, riječ „korupcija" nije primjenjiva u odnosu na ovaj parlament.
Boycott of parliament	Bojkot parlamenta
It is the law on using the state and the national symbols.	Riječ je o zakonu o korištenju državnih i nacionalnih obilježja.

Na temelju dvojezičnog teksta (kao u gornjem primjeru) moguće je rečenično sravniti jedan paralelni korpus, pri čemu se jednom segmentu na izvornom jeziku, najčešće rečenici, pridružuje segment ili rečenica na ciljnome jeziku (Yu et al., 2012).

Rečenično sravnjivanje korpusa može se izvršiti ručno ili automatski pomoću raznih metoda, kao npr. Gale-Church metoda (Manning i Schütze, 1999). Radi se o empirijskom opažanju da se semantički ekvivalentne rečenice, tj. izvorna rečenica i njen prijevod, podudaraju u duljini, tj. broju riječi (Gale i Church, 1991). Odnosno, dulje rečenice u izvornom jeziku odgovaraju duljim

rečenicama u ciljnom jeziku. Jednako tako, kraće rečenice u jednom jeziku inkliniraju kraćim prijevodima u drugom jeziku.

Primjer rečenično sravnjenog paralelnog korpusa prikazan je u nastavku:

The event was highly relevant for several reasons.	Ovaj je događaj iznimno znakovit iz nekoliko razloga.
The timing agreed on will be respected.	Dogovoreni će rokovi biti ispoštovani.
EU funds should be included in the multiannual budgeting of finance ministries.	Sredstva EU morala bi biti uključena u višegodišnje nacrte budžeta ministarstava financija.
Other Balkan countries are also banking on increased financial assistance from the EU for their development efforts.	Ostale balkanske zemlje također računaju na povećanu financijsku potporu EU njihovom razvitku.
Economic stagnation in the EU fuels protectionist tendencies, impeding a successful outcome to the Doha trade round.	Gospodarska stagnacija u EU potpiruje protekcionističke sklonosti, sprječavajući uspješan ishod trgovačkih pregovora u Dohi.
To be clear, the word „corrupt" doesn't exist for this parliament.	Ukratko, riječ „korupcija" nije primjenjiva u odnosu na ovaj parlament.
Boycott of parliament	Bojkot parlamenta
It is the law on using the state and the national symbols.	Riječ je o zakonu o korištenju državnih i nacionalnih obilježja.

Ipak, ideja koja je primijenjena u paralelnim korpusima, a to je zapisivanje jednog teksta u više različitih jezika i/ili pisama nije nova. Primjerice, kamen iz Rosette vrlo je dobar primjer paralelnog korpusa. Otkriven je 1799. i omogućio je dešifriranje drevnih egipatskih pisama, s obzirom da je na njemu ispisano na desetke redaka kodiranih pomoću tri različita pisma: egipatskih hijeroglifa, demotskog i starogrčkog pisma (Giménez, 2008; Tufiş i Barbu, 2001).

S obzirom na različite pristupe strojnom prevođenju sa specifičnim prednostima i nedostatcima, postoji realna potreba za prosudbom, tj. evaluacijom kvalitete ishoda procesa strojnog prevođenja. Posebno poglavlje u ovom radu bavit će se pitanjem, mogu li računala razlikovati dobre prijevode od loših? Nadalje, može li strojni prijevod zadovoljiti ljudskog

evaluatora i njegove potrebe? Treba naglasiti da strojni prijevod ni ne mora uvijek biti savršen. Prema namjeni, ali istovremeno i prema brzini i kvaliteti prevođenja, strojno prevođenje se razvrstava u jednu od tri kategorije (Koehn, 2010):

- asimilacija (eng. *assimilation*) – strojno prevođenje za osnovno razumijevanje sadržaja na stranom jeziku,
- komunikacija (eng. *communication*) – strojno prevođenje poslovne korespondencije, e-pošte itd.,
- diseminacija (eng. *dissemination*) – strojno prevođenje visoke kvalitete s namjenom publiciranja.

Strojni prijevod za asimilacijske potrebe najmanje je kvalitete, budući da treba prenijeti samo srž, tj. osnovno značenje poruke (eng. *gisting*) (Koehn, 2010), bez obzira na jezične konvencije i ustaljena gramatička pravila standardnoga jezika. Takav oblik strojnog prevođenja danas je implementiran u brojne web servise, kao što su Bing Translator[1], Google Translate[2] (Koletnik Korošec, 2011), SYSTRANet[3], Yandex Translate[4] i drugi (Hampshire i Porta Salvia, 2010; Gaspari i Hutchins, 2007).

Brojni su pristupi i mogućnosti poboljšanja kvalitete strojnog prijevoda. Jedan pristup nalaže uporabu kontroliranog jezika (eng. *controlled language*), tj. jezika s određenim pravilima konstrukcije rečenica, što proces strojnog prevođenja znatno olakšava (Hutchins, 2003). Nadalje, kvaliteta strojnog prijevoda u pravilu je uvijek visoka ukoliko se prevodi tekst koji je vrlo sličan onome kojim se sustav za strojno prevođenje trenirao (Haddow i Koehn, 2012; Koehn, 2010). Najčešće su takvi tekstovi iz vrlo specifične domene (npr. pravne, tehničke itd.), koja primjerice uključuje zakonodavne dokumente ili razne oblike uputa za korištenje uređaja (eng. *user guides, user manuals*).

Cilj strojnog prevođenja je izgraditi sustave koji potpuno automatizirano proizvode strojne prijevode najviše kvalitete (eng. *fully-automatic high quality machine translation, FAHQMT*) (Koehn, 2010; Hutchins, 2001). Pored toga, istražuju se mogućnosti integracije različitih tehnologija sa strojnim prevođenjem, poput govornih tehnologija, kao što su računalna sinteza govora (eng.

[1] https://www.bing.com/translator/
[2] https://translate.google.com/
[3] http://www.systranet.com/translate
[4] https://translate.yandex.com/

speech synthesis) i prepoznavanje govora (eng. *speech recognition*). Na taj način računalno prepoznati govor koji je na izvornom jeziku može poslužiti kao ulazna varijabla u sustav za strojno prevođenje, a zatim se na temelju generiranog strojnog prijevoda može računalno proizvesti govor na ciljnome jeziku.

Optičko prepoznavanje znakova (eng. *optical character recognition*) također ima primjenu u strojnom prevođenju. Ono se može iskoristiti u procesu digitalizacije tiskane dokumentacije radi izrade paralelnog korpusa (Phillips et al., 2010), ali isto tako u procesu strojnog prevođenja, što je posebno pogodno za mobilne uređaje. Tekst na stranom jeziku može se pomoću kamere preuzeti na uređaj, nakon čega se strojni prijevod vrši lokalno ili se proces prevođenja delegira drugim računalima ili poslužiteljima.

U ovom doktorskom radu, u fokusu istraživanja jesu statističko strojno prevođenje, s posebnim naglaskom na strojno prevođenje temeljeno na frazama te računalna adaptacija domene kojom se nastoji poboljšati kvaliteta strojnog prijevoda u specifičnoj domeni i za određeni jezični par. Doktorski rad podijeljen je u dvije osnovne cjeline: teorijski i praktični dio. U teorijskom dijelu naveden je znanstveni doseg te je detaljno pojašnjena teorijska podloga koja omogućuje razvoj sustava za statističko strojno prevođenje, dok su u praktičnom dijelu disertacije opisani postupci izgradnje sustava za automatsko strojno prevođenje. Nadalje, praktični dio obuhvaća opis provedenih istraživanja, eksperimentalne rezultate s pripadajućom diskusijom i analizom ishoda eksperimenata.

Doktorska disertacija sastoji se od šest tematskih poglavlja koja pokrivaju uvodna razmatranja problematike, povijesni pregled razvoja strojnog prevođenja, opis modela statističkog strojnog prevođenja i metoda evaluacije, opis doktorskog istraživanja s pripadajućim eksperimentalnim rezultatima te zaključno poglavlje. Nakon toga slijede popis korištenih bibliografskih jedinica (sedmo poglavlje), detaljan prikaz rezultata metode ponovnog uzorkovanja (Dodatak A), rezultati ljudske evaluacije kvalitete statističkih strojnih prijevoda (Dodatak B), deskriptivna statistika rezultata ljudske evaluacije (Dodatak C) te životopis autora disertacije s popisom radova.

U prvom poglavlju dana je motivacija za provođenjem doktorskog istraživanja te je naglašena važnost automatskog strojnog prevođenja u današnjem globaliziranom svijetu, posebno u kontekstu Republike Hrvatske. Navedene su osnovne podjele područja strojnog prevođenja: prema načinu definiranja „pravila" strojnog prevođenja, prema namjeni te time i brzini i kvaliteti te prema stupnjevima automatizacije procesa prevođenja. Istaknuta je uloga tekstualnih podatkovnih skupova u obliku rečenično-sravnjenih paralelnih korpusa u izgradnji sustava za statističko strojno prevođenje koje se temelje na „empirijskim opažanjima". Navedene su i

prednosti te cilj i namjena sustava za automatsko strojno prevođenje. Određeni trendovi u području istraživanja također su spomenuti.

Drugo poglavlje daje povijesni pregled razvoja područja strojnog prevođenja. Istaknuto je kako je povijest strojnog prevođenja obilježena brojnim usponima i padovima. Ovo poglavlje navodi događaje koji su omogućili napredak i primjenu raznih pristupa strojnom prevođenju. Ustanovljeno je da je razvoj strojnog prevođenja 1940-ih i 1950-ih godina obilježen radom kibernetičara te uvjetovan definiranjem kriptografskih i kriptoanalitičkih metoda te Shannonove informacijsko-komunikacijske teorije. Utvrđeno je kako je zatim uslijedilo razdoblje koje u središte istraživanja postavlja formalne gramatike, tj. jezične formalizme. Pored toga, spomenuto je i ALPAC izvješće iz 1966. koje se negativno odrazilo na daljnji razvoj strojnog prevođenja. Potom su navedeni primjeri sustava za strojno prevođenje koji su, unatoč „lošoj" istraživačkoj klimi, razvijeni u određenim zemljama. Kasnih 1980-ih godina razvoj računalne tehnologije, sve veća dostupnost osobnih računala te pojava interneta ubrzavaju napredak automatskog strojnog prevođenja koje se temelji na statističkim paradigmama. Istaknuto je kako je 21. stoljeće razdoblje velikog optimizma u području strojnog prevođenja, s velikim brojem istraživača i investitora. Poglavlje također navodi aktualne istraživačke teme u područjima povezanima sa strojnim prevođenjem te ukazuje na određene nedoumice i sumnje u kvalitetu strojnog prijevoda.

Treće poglavlje prikazuje model statističkog strojnog prevođenja, detaljno opisuje elemente modela sustava za strojno prevođenje te pojašnjava ulogu sastavnih dijelova, tj. značajki u modelu sustava za statističko strojno prevođenje. Definirana je terminologija, a pomoću brojnih primjera opisana je logika modela sustava za statističko strojno prevođenje. Posebno su analizirani jezični i prijevodni model te dekoder, s obzirom da se radi o temeljnim komponentama modela sustava za statističko strojno prevođenje. Važnost statističkih paradigmi i temeljnih razvojnih procesa posebno je naglašena te su analizirane karakteristike i posebnosti jezičnog modela, problemi u razvoju jezičnih modela te određene zakonitosti u tekstualnim korpusima koje ograničavaju mogućnosti n-gramskih jezičnih modela. Potom su istražene metode treniranja jezičnog i prijevodnog modela, metode izglađivanja n-gramskog jezičnog modela, metode ugađanja modela sustava za statističko strojno prevođenje, postupci sravnjivanja i premještanja riječi te heuristike ekstrakcije, uparivanja fraznih parova i pohranjivanja ekstrahiranih fraza. Problematika IBM-ovih modela pomno je analizirana, a procesi i algoritmi dekodiranja strojnog prijevoda detaljno su opisani. Uloge sastavnih dijelova modela sustava za statističko strojno prevođenje u automatskom prevođenju morfološki bogatih jezika poput hrvatskoga posebno su istaknute. Ovo poglavlje obuhvaća i pregled ranije provedenih istraživanja i mogućnosti proširenja standardnog modela statističkog strojnog prevođenja. Radi se pritom o log-linearnom obliku modela sustava za

statističko strojno prevođenje koji sastavne dijelove modela opisuje logaritamskim funkcijama značajki modela. Takav modificirani model dozvoljava pridruživanje težina značajkama modela sustava koje se zatim mogu ugađati metodom diskriminativnog učenja. Uloge preostalih značajki u modificiranom modelu statističkog strojnog prevođenja također su pojašnjene, a problemi u pronalaženju, tj. generiranju točnih strojnih prijevoda naglašeni. Nadalje, mogućnosti implementacije jezičnoga znanja u model sustava za statističko strojno prevođenje također su analizirane. Posebna važnost u trećem poglavlju dana je raznim metodama adaptacije domene u modelu sustava za statističko strojno prevođenje, poput ugađanja domenski specifičnim podatkovnim skupom, kombinacije domenskih i izvandomenskih značajki, metode alternativnog puta dekodiranja i primjene *back-off* modela. Analizirani su i problemi koji proizlaze iz primjene određenih tehnika adaptacije domene.

Četvrto poglavlje opisuje evaluaciju kvalitete strojnog prijevoda. Budući da evaluacija predstavlja važnu i neophodnu fazu u razvoju sustava za strojno prevođenje analizirane su, kako ljudske metode evaluacije, tako i automatske metode koje vrlo bitnu ulogu imaju i pri ugađanju značajki modela sustava. Pojašnjeni su kriteriji i skale ljudske prosudbe kvalitete strojnih prijevoda, a naglašeni su i problem interpretacije rezultata evaluacije te važnost reprezentativnosti i konzistentnosti evaluacijskih rezultata. Karakteristike primijenjenih automatskih evaluacijskih metrika su detaljno objašnjene, a pored toga istaknuti su i problemi automatske evaluacije te statističke značajnosti rezultata evaluacije.

Peto poglavlje odnosi se na eksperimentalno istraživanje i opis eksperimentalnog okruženja. Prvo je dan znanstveni doprinos doktorskog istraživanja, a zatim su opisani ciljevi istraživanja te definirane tri ključne hipoteze. Dio petog poglavlja posvećen je primijenjenoj metodologiji te tijeku istraživanja. Opisane su metode pretprocesiranja i normalizacije ulaznih podatkovnih skupova, faze razvoja sustava te karakteristike izgrađenih sustava za strojno prevođenje. Zatim su pojašnjene posebnosti hibridnih sustava koji kombiniraju domenske i izvandomenske značajke modela sustava. Primjenom hibridnih sustava, naime, nastojale su se poboljšati performanse sustava za automatsko strojno prevođenje, tj. kvaliteta statističkih strojnih prijevoda. Navedeni su upotrijebljeni alati i resursi te razvijena programska rješenja, a zatim su detaljno kvantificirani i analizirani korišteni podatkovni skupovi za treniranje, ugađanje i testiranje sustava za statističko strojno prevođenje. Utjecaj karakteristika podatkovnih skupova na logiku modela sustava također je naglašen. Sam postupak izgradnje sustava za statističko strojno prevođenje temeljeno na frazama posebno je detaljno opisan uz izračun perpleksnosti i udjela nepoznatih riječi. Testiranje performansi svih osam izgrađenih sustava, odnosno četiri za hrvatsko-engleski i četiri za englesko-hrvatski smjer, vršeno je pomoću osam različitih automatskih metrika i pomoću istog

podatkovnog skupa za testiranje. Pored ljudske evaluacije koja je provedena s obzirom na kriterije adekvatnosti i fluentnosti, analizirani su i tipovi pogrešaka u strojnim prijevodima. Konzistentnost među evaluatorima istražena je Cronbach alphom, a korelacija ljudske evaluacije i automatskih metrika analizirana je pomoću Pearsonovog koeficijenta korelacije. Performanse izgrađenih sustava za strojno prevođenje zatim su komparirane u odnosu na performanse online prevodilačkih servisa, također testiranih istim podatkovnim skupom za testiranje. Pritom su hipoteze ovog doktorskog istraživanja ili potvrđene ili opovrgnute. Statistička značajnost istraživačkih rezultata također je ispitana, a sustavi su odgovarajuće rangirani s obzirom na rezultate evaluacije.

U šestom poglavlju dan je zaključak provedenog doktorskog istraživanja s jasnim i konciznim pregledom primijenjenih metoda, alata i resursa, opisom faza razvoja sustava te provedenih eksperimenata. Potom su prodiskutirani najvažniji istraživački rezultati sa zadanom razinom statističke značajnosti, a ključni pronalasci su istaknuti. Rezultati ispitivanja istraživačkih hipoteza posebno su pojašnjeni, a uspješni pristupi povećanju kvalitete automatskog strojnog prijevoda u općenitoj domeni i domeni računalnog softvera za hrvatsko-engleski i englesko-hrvatski jezični par posebno su naglašeni.

2. POVIJESNI PREGLED RAZVOJA STROJNOG PREVOĐENJA

Povijest strojnog prevođenja puna je uspona i padova (Koehn, 2010). 1930-ih godina Georges Artsrouni prijavljuje prve patente za „strojeve koji prevode“ (eng. *translation machines*) (Hutchins, 2001b). Radilo se o automatskom dvojezičnom rječniku zapisanom na bušenim trakama. Ruski istraživač Petr Smirnov-Troyanskij iste godine, neovisno o istraživanju Georgesa Artsrounija, prijavljuje opsežniji patent za također dvojezični rječnik koji se, međutim, oslanja i na metodu prepoznavanja gramatičkih uloga u raznim jezicima (Hutchins, 2004).

Za vrijeme Drugog svjetskog rata, matematičar Alan Turing radio je na kriptoanalizi Enigme, pritom koristeći elektromehaničke strojeve koji su se upotrebljavali za dekodiranje znakova. Prvo programibilno računalo ENIAC (eng. *Electronic Numerical Integrator And Computer*) predstavljeno je javnosti 1946. te ubrzo nakon toga započinju istraživanja na području strojnog prevođenja.

Među pionirima svakako treba spomenuti Warrena Weavera koji je predložio da se prevođenju pristupa računalno. 1947. u korespondenciji s kibernetičarom Norbertom Wienerom predložio je upotrebu digitalnog računala za prevođenje dokumenata između dva prirodna (ljudska) jezika (Hutchins, 2001b). Na nagovor svojih kolega, Weaver počinje intenzivno istraživati područje strojnog prevođenja, a dvije godine kasnije objavljuje svoje rezultate istraživanja u memorandumu „Translation“ (Chéragui, 2012). U memorandumu se iznose ciljevi strojnog prevođenja te potencijali digitalnih računala u prevođenju prirodnih jezika. Weaver razmatra razne pristupe strojnom prevođenju, pa tako primjerice predlaže da se kriptografske i kriptoanalitičke metode primijene u procesu automatskog prevođenja, što uvelike doprinosi stvaranju pozitivne i financijski stabilne istraživačke klime (Koehn, 2010):

Also knowing nothing official about, but having guessed and inferred considerable about, powerful new mechanized methods in cryptography – methods which I believe succeed even when one does not know what language has been coded – one naturally wonders if the problem of translation could conceivably be treated as a problem in

cryptography. When I look at an article in Russian, I say "This is really written in English, but it has been coded in some strange symbols. I will now proceed to ***decode****."* (Hutchins, 1997)

Godine 1948. Claude Shannon objavljuje rad „A Mathematical Theory of Communication", što se smatra početkom razvoja teorije informacije (Shannon, 1948). U središtu navedene teorije jest problematika prijenosa informacije kroz kanal sa šumom (Specia, 2010). Naime, za pouzdanu komunikaciju kanalom sa šumom, komunikacija se treba odvijati u skladu s tzv. kapacitetom kanala. Nadalje, prema Shannonu, informacija se može kvantificirati u bitove (eng. *bits*) potrebne za opisivanje ishoda neizvjesnog događaja. Potreban broj bitova opisan je svojom entropijom (Manning i Schütze, 1999). Stoga se kao ključna mjera informacije i danas upotrebljava entropija koja mjeri količinu neizvjesnosti (Koehn, 2010) i koja se definira kao prosječan broj bitova potrebnih za pohranjivanje ili komunikaciju jednog znaka u poruci.

Treba naglasiti da su neki od ranih principa u strojnom prevođenju uspostavljenih još 1940-ih godina i danas vrlo aktualni (Koehn, 2010). Pa se tako primjerice još uvijek govori o dekodiranju izvornog jezika u ciljni jezik primjenom kanala sa šumom kao metodom modeliranja složenih sustava.

Yehoshua Bar-Hillel, znanstvenik s MIT-a (eng. *Massachusetts Institute of Technology*), prvi je istraživač koji je u punom radnom vremenu istraživao strojno prevođenje potpomognuto rječnicima. Godine 1952. organizirao je i prvu međunarodnu konferenciju o strojnom prevođenju. Kasnijih je godina izrazio sumnju u budućnost i mogućnosti strojnog prevođenja (Hutchins, 2001c). 1954. održan je Georgetown eksperiment (eng. *Georgetown-IBM experiment*) u uredima korporacije IBM (eng. *International Business Machines Corporation*) u New Yorku (Hutchins, 2004b). Radilo se zapravo o prvom javnom predstavljaju sustava za strojno prevođenje koji je prevodio s ruskog jezika na engleski. Međutim, radilo se o vrlo jednostavnom sustavu s ograničenim vokabularom od 250 riječi, 6 gramatičkih pravila i 49 odabranih rečenica iz područja kemije. Sam eksperiment je bio vrlo dobro medijski popraćen, što je privuklo dodatne investitore.

1950-ih i 1960-ih godina Noam Chomsky istražuje mogućnosti modeliranja znanja i jezika primjenom formalnih gramatika. Općenito, navedeno razdoblje obilježili su sustavi za strojno prevođenje temeljeni na dvojezičnim rječnicima i pravilima za održavanje ispravnog poretka riječi u rečenici. S obzirom na metode i pristupe prevođenju, tadašnji sustavi za strojno prevođenje rabili su (Stein, 2013; Chéragui, 2012, Koehn, 2010):

- direktnu metodu (eng. *direct method*) – koja pomoću jednostavnih pravila uparuje riječi,
- metodu transfera (eng. *transfer method*) – sofisticiranija metoda koja koristi morfološku i sintaktičku analizu,
- metodu međujezika (eng. *interlingua*) – koja koristi apstraktnu reprezentaciju značenja.

Godine 1964. vlada Sjedinjenih Američkih Država (eng. *Federal Government of the United States*) naredila je ispitivanje stanja i brzinu napretka područja strojnog prevođenja, s obzirom na velike financijske izdatke. Odbor ALPAC (eng. *Automatic Language Processing Advisory Committee*) tada sastavljen od sedmero znanstvenika objavio je 1966. izvješće o stanju u području strojnog prevođenja (Hutchins, 1996; ALPAC, 1966). Zaključak izvješća bio je da je strojno prevođenje skuplje, manje precizno i sporije od klasičnog ljudskog prevođenja te da nije izgledno da strojno prevođenje u skoroj budućnosti i s povećanjem financijske potpore dosegne razinu kvalitete čovjeka, tj. ljudskog prevoditelja. Izvješće je bilo intonirano vrlo negativno, što je zaustavilo razvoj strojnog prevođenja ne samo u Sjedinjenim Američkim Državama, već i u Sovjetskom Savezu te Velikoj Britaniji. Istraživanja su, međutim, nastavljena u Kanadi, Francuskoj i Njemačkoj.

Unatoč napuštanju strojnog prevođenja kao znanstvene discipline, u SAD-u s radom nastavljaju Peter Toma, osnivača tvrtke „Systran" (1968.) te Bernard Scott, osnivač tvrtke „Logos" (1970.). Od 1970. sustav Systran koristilo je američko ministarstvo obrane (eng. *United States Department of Defense*), a od 1976. Komisija Europske Zajednice (eng. *Commission of the European Community*). Iste godine se pojavljuje kanadski sustav „TAUM Météo" (franc. *Traduction Automatique à l'Université de Montréal*) razvijen na montrealskom sveučilištu, koji prevodi vremenske prognoze s engleskog na francuski (Hutchins, 2001b). Mogao je prevesti 80000 riječi dnevno, tj. cca. 30 milijuna riječi godišnje. Sustav TAUM Météo bio je u operativnoj upotrebi sve do 2011. 1980-ih godina izlaze i drugi komercijalni sustavi za strojno prevođenje, npr. „Logos" i „METAL" (Koehn, 2010). Razdoblje 1970-ih i ranih 1980-ih godina pripada strojnom prevođenju temeljenom na pravilima.

1980-ih i 1990-ih godina u središtu istraživanja je metoda međujezika koja nastoji formalno opisati i reprezentirati značenje neovisno o određenom jeziku. Sustavi temeljeni na metodi međujezika bili su „CATALYST", za prevođenje tehničkih priručnika tvrtke „Caterpillar" te sustav „Pangloss" (Koehn, 2010). U sklopu njemačkog projekta „Verbmobil" također su razvijeni sustavi temeljeni na metodi međujezika. U to vrijeme brojne japanske tvrtke razvijale su vlastite

sustave za strojno prevođenje: Brother, Fujitsu, Hitachi, Mitsubishi, NEC, Panasonic, Sanyo, Sharp, Toshiba (Chéragui, 2012).

S obzirom da je prevođenje jezika vrlo teško strogo formalizirati pravilima, 1980-ih godina se pojavila potreba za novim metodama strojnog prevođenja. Umjesto da se jezik striktno opiše pravilima, postavlja se pitanje kako „učiti" iz već prevedenih tekstova, tj. iz već viđenih primjera? Istraživanja na tom području dovela su do novih rješenja, uglavnom temeljenih na podatcima (eng. *data-driven methods*). Rani pokušaji da se iskoriste već prevedeni dijelovi rečenica rezultirali su strojnim prevođenjem temeljenim na primjerima (eng. *example-based machine translation*) (Hutchins, 2005). Iako ova metoda tada nije uspjela zainteresirati velik broj istraživača, danas se ovaj pristup itekako primjenjuje (Koehn, 2010). Naime, jezična tehnologija koja ima vrlo široku primjenu u računalno-potpomognutom prevođenju jest prijevodna memorija (eng. *translation memory*), koja pohranjuje segmente ili rečenice na jednom jeziku te već ranije (ljudski) prevedene semantičke prijevodne ekvivalente na drugom jeziku (Reinke, 2013; Baldwin, 2004). Kada prevoditelj prevodi novi tekst, sustav prijevodne memorije pretražit će postojeće prijevode u bazi podataka te prevoditelju ponuditi adekvatne prijevode, ukoliko se novi tekst do određene razine podudara s tekstom koji se već nalazi u prijevodnoj memoriji (Seljan i Pavuna, 2006). Poznati sustavi prijevodne memorije danas su SDL Trados, memoQ, MateCat, Wordfast, OmegaT, ATRIL Déjà Vu, STAR Transit, Memsource, Lionbridge Translation Workspace itd. (Skadiņš et al., 2014).

Kasnih 1980-ih godina dolazi do velikog skoka u razvoju strojnog prevođenja (Hutchins, 2001b). U sklopu istraživačkog projekta „CANDIDE" (1988.), usmjerenog na prepoznavanje govora (eng. *speech recognition*), postavljena je matematička osnova za daljnji razvoj strojnog prevođenja. Naime, IBM-ov istraživački tim prepoznavanje govora shvaća kao statistički problem kojemu se može pristupati promatrajući veliku količinu tekstualnih podataka, tj. korpusa. U sklopu projekta razvijen je i sustav za strojno prevođenje englesko-francuskog jezičnog para (Berger et al., 1994).

U isto vrijeme, broj digitalnih tekstualnih korpusa raste, prvenstveno zbog sve većeg broja stvaratelja digitalnih dokumenata (posebno na internetu), a dijelom i zbog provođenja postupka digitalizacije fizičke dokumentacije. 1990-ih godina cijena računala opada, a procesorska snaga i računalne performanse potrebne za statističku obradu podataka rastu. Nadalje, koncept središnjeg računala (eng. *mainframe computer*) zastarijeva, te se razvoj računala preusmjerava prema osobnim računalima i radnim stanicama (eng. *workstations*), a time potencijalno raste i broj korisnika strojnog prevođenja. 1990-ih se pojavljuje i besplatni internetski prevodilački servis „BabelFish" na mrežnoj stranici „AltaVista", koji se temelji na Systranovoj tehnologiji strojnog prevođenja temeljenog na pravilima. Sustav je dobio ime prema „Babel fishu", fiktivnom žutom biću koje

potječe iz bestselera Douglasa Adamsa „Vodič kroz galaksiju za autostopere“ (eng. *The Hitchhiker's Guide to the Galaxy*). Radi se o ribici koja se postavlja u uho, nakon čega za čovjeka simultano prevodi bilo koji jezik.

1999. sudionici radionice koja se održavala na Sveučilištu Johns Hopkins reimplementirali su IBM-ove metode, a novonastali programski alati postali su javno dostupni (Al-Onaizan et al., 1999; Koehn, 2010). Organizacija DARPA (eng. *Defense Advanced Research Projects Agency*) prepoznaje važnost ponovno implementiranih IBM-ovih metoda te potom financira istraživačke projekte TIDES (eng. *Translingual Information Detection Extraction and Summarization*) i GALE (eng. *Global Autonomous Language Exploitation*).

2000-e su godine velikog optimizma u području strojnog prevođenja. To je razdoblje sustava za statističko strojno prevođenje koje se temelji na empirijskim opažanjima (Koehn, 2010). Takvi sustavi se pretežno izgrađuju za određenu domenu, tj. karakteristično područje s ograničenim vokabularom i specifičnim rečenicama, s obzirom da za uže područje namjene generiraju kvalitetnije strojne prijevode (Haddow i Koehn, 2012). Prednosti takvog pristupa su relativno jeftina izgradnja sustava za statističko strojno prevođenje, jednostavno dodavanje novih jezika te automatizirano ugađanje sustava. Nadalje, takvi strojni prijevodi su vrlo tečni, tj. fluentni. S druge pak strane, statističko strojno prevođenje ima mnoge poteškoće s gramatičkim aspektima jezika (vrijeme, broj, padež, slaganje itd.), a samo ugađanje sustava ovisi o brojnim faktorima i stoga nije uvijek precizno (Dillinger i Marciano, 2012). Pored toga, statistički strojni prijevodi vrlo su nepredvidivi, a ispuštanje ili neprevođenje riječi je vrlo često. Za statističko strojno prevođenje potrebne su velike količine radne memorije i jake računalne performanse (Turchi et al., 2012; Turchi et al., 2008). Veliku zaslugu u razvoju empirijskih sustava imaju i automatske metrike za evaluaciju kvalitete strojnog prijevoda (Gonzàlez, 2014; Giménez, 2008).

Danas se statističkim strojnim prevođenjem bave brojne akademske institucije, istraživački centri i privatne organizacije, poput kompanija SDL, Systran, Asia Online, IBM, Google i Microsoft. Zadnjih godina se istražuju i hibridni pristupi strojnom prevođenju: npr. u sustave za statističko strojno prevođenje ugrađuju se drugi izvori jezičnog (sintaktičkog i morfološkog) znanja ili se kombiniraju s pristupom temeljenim na pravilima (Okpor, 2014; Costa-jussa` et al., 2013; Stein, 2013; Eisele et al., 2008). Današnji popularni sustavi temeljeni na pravilima su Apertium (Tyers et al., 2010) i Systran (Hutchins, 2003b). Takvi sustavi se dobro nose sa svim gramatičkim aspektima jezika te generiraju predvidive prijevode, a uz to ne zahtijevaju veliku računalnu snagu (Dillinger i Marciano, 2012). Nadalje, prednosti takvih sustava danas su i mogućnost vrlo preciznog ugađanja gramatike te malen broj ispuštanja riječi u strojnom prijevodu. Međutim, razvijanje takvih sustava je skupo i vremenski vrlo zahtjevno, a i dodavanje

novih jezika također je vrlo složeno. Ugađanje se vrši ručno, što povećava mogućnosti pogrešaka, a i strojni prijevodi su na kraju manje fluentni u odnosu na statističke prijevode (Dillinger i Marciano, 2012).

Danas se istražuju i mogućnosti strojnog prevođenja tipa govor-u-govor (eng. *speech-to-speech translation*) (Hutchins, 2009; Black et al., 2002) te mogućnosti integriranja sustava za strojno prevođenje u radni tok jednog poslovnog subjekta (eng. *workflow*) (Vilar et al., 2012; Sun et al., 2011). Integracija prijevodnih memorija u sustav za strojnog prevođenje također je od istraživačkog interesa (Wang et al., 2013; Dillinger i Marciano, 2012; Kanavos i Kartsaklis, 2010). Budući da je za izgradnju sustava za statističko strojno prevođenje neizbježna velika količina dvojezičnih tekstova, istražuju se mogućnosti prikupljanja i izrade paralelnih korpusa pomoću *crowdsourcinga* (Post et al., 2012). Komercijalne platforme *crowdsourcinga*, kao npr. mehanički Turčin (eng. *mechanical Turk*), koriste se i za potrebe evaluacije kvalitete strojnog prijevoda (Bojar et al., 2013; Callison-Burch et al., 2012).

Glavne kritike na račun strojnog prevođenja danas su usmjerene prema opasnostima zamjene ljudskih prevoditelja te kvaliteti i (domenskim) ograničenjima strojnog prevođenja (Dillinger i Marciano, 2012). Međutim, ideja strojnog prevođenja nije zamijeniti čovjeka, već asistirati mu pri prevođenju i omogućiti mu pristup informacijama na stranom jeziku. Ono što se mijenja su samo brzina te tijek i aktivnosti unutar procesa prevođenja. S obzirom da je danas količina teksta koju treba prevesti prevelika za čovjeka, računalo može generirati prijevode koje zatim čovjek samo dorađuje, umjesto da ih prevodi ispočetka. Kada skeptični ili neupućeni govore o kvaliteti strojnih prijevoda, tada se njihove prosudbe vrlo često donose na temelju opaženih performansi prevodilačkih servisa na internetu. No, milijuni ljudi svakodnevno koriste strojno prevođenje, najčešće kako bi dobili osnovnu informaciju o sadržaju koji ne razumiju (Koehn, 2010). Pored toga, takvi servisi su u pravilu besplatni. Svakako ovdje treba naglasiti da su sustavi za strojno prevođenje koji se po narudžbi izgrađuju za određeni poslovni subjekt znatno drugačiji i kvalitetniji. Naime, takvi sustavi su posebno prilagođeni radnom okruženju i potrebama poslovnog subjekta te stoga postižu daleko bolje i kvalitetnije strojne prijevode; npr. u Europskoj komisiji, u kompanijama automobilske industrije, u multinacionalnim uredima itd.

No, kritičari su u pravu kada tvrde da jedino čovjek može „ispravno“ razumjeti profinjenost i suptilnost jezika i kulture, te stoga ispravno prevesti razne oblike teksta (barem za sada). Kompleksnost i višeznačnost jezika izrazito otežava strojno prevođenje, međutim, ispostavilo se da postoji vrlo velik broj tipova uobičajenih rečenica i tekstova s kojim računalo ima manje poteškoća pri prevođenju (Dillinger i Marciano, 2012). Svakako se pritom ne misli na poeziju ili sofisticirane prijevode, već na tekstove s ustaljenim vokabularom, poretkom riječi i jasnim te

jednoznačnim jezikom bez metaforičkog značenja. To se primjerice može odnositi na brošure ili korisničke upute za korištenje raznih proizvoda (pogodno za lokalizaciju), prevođenje dinamičkog sadržaja na internetu, filmskih titlova (podslova), vremenskih prognoza, televizijskih ili radijskih vijesti, korisničkog sadržaja (komentari na mrežnim stranicama, SMS-ovi, poruke na chatu ili društvenim mrežama), na praćenje informacija iz stranih izvora ili na generiranje grubih prijevoda za osnovno razumijevanje informacije (Dillinger i Marciano, 2012).

3. MODEL STATISTIČKOG STROJNOG PREVOĐENJA

Jedan od pristupa automatskom strojnom prevođenju jest pristup upravljan podatcima. Takav pristup uključuje i model statističkog strojnog prevođenja (Brown et al., 1993) koji je neovisan o jeziku te stoga redovito ne zahtijeva posebno lingvističko znanje (Koehn, 2010). Istraživanja na području statističkog strojnog prevođenja započela su kasnih 1980-ih na IBM-ovom projektu „CANDIDE" (Koehn, 2010). Izvorna IBM-ova istraživanja statističkog strojnog prevođenja rezultirala su definiranjem tzv. IBM modela 1-5, koji pripadaju generativnoj metodi modeliranja. Naime, proces generiranja podataka dijele u manje korake koji se zatim zasebno modeliraju i statistički opisuju, a konačan rezultat nastaje kombiniranjem svih pojedinačnih koraka. IBM modeli polaze od uparivanja pojedinih riječi u izvornom i ciljnom jeziku te dozvoljavaju umetanje i ispuštanje riječi (Koehn, 2010). Statističko strojno prevođenje posljednjih je godina u središtu istraživačkih interesa, osobito zbog mogućnosti izgradnje sustava za automatsko strojno prevođenje primjenom velike količine jezičnih resursa u obliku paralelnih i jednojezičnih korpusa te jezično neovisnih alata. Paralelni korpusi su podatkovni skupovi koji se sastoje od tekstova na izvornom i ciljnom jeziku, odnosno ciljnom i izvornom jeziku (Klaper et al., 2013). Takvi korpusi predstavljaju osnovno sredstvo rada sustava za statističko strojno prevođenje (Wetzel i Bond, 2012).

Jedan od mogućih pristupa statističkom strojnom prevođenju jest primjena prijevodnog modela temeljenog na frazama (eng. *phrase-based translation model*) (Koehn, 2010). Osnova ideja statističkog strojnog prevođenja temeljenog na frazama jest segmentirati skup rečenica izvornog jezika (eng. *test set*) u fraze, odnosno nizove riječi koje se zatim prevode (tj. zamjenjuju) u ciljni jezik. Cjelovite rečenice u ciljnom jeziku nastaju sastavljanjem, tj. nizanjem prijevoda fraza, a nazivaju se prijevodni kandidati (eng. *candidate translations*). U modelu statističkog strojnog prevođenja riječi izvornog i ciljnog jezika se uparuju (sravnjuju), a same fraze ekstrahiraju se upravo iz sravnjenosti riječi u izvornom i ciljnom jeziku, pri čemu sve riječi iz fraznog para trebaju biti međusobno sravnjene (Koehn, 2010). Prema Koehn (2010), tri su temeljne faze u statističkom strojnom prevođenju temeljenom na frazama: sravnjivanje riječi, ekstrakcija fraznih

parova i izračun vjerojatnosti za svakih frazni par. Nadalje, tri komponente modela sustava za strojno prevođenje temeljeno na frazama izravno utječu na kvalitetu statističkog strojnog prijevoda (Koehn, 2010; Koehn, 2008):

- tablica prijevoda fraza, tj. fraznih struktura (eng. *phrase translation table*),
- model preslagivanja/premještanja redoslijeda, tj. poretka riječi (eng. *reordering model*) i
- jezični model (eng. *language model*).

Tim komponentama, tj. značajkama (eng. *features*) se u postupku učenja, tj. treniranja pojedinog modela pridružuju određene vrijednosti težina (eng. *weights*) koje utječu na logiku modela statističkog strojnog prevođenja temeljenog na frazama (Koehn, 2010). Značajke i težine su u modelu statističkog strojnog prevođenja temeljenog na frazama implementirani u obliku log-linearnog modela (Jurafsky i Martin, 2013). S obzirom da se izračuni vrijednosti težina pojedinih (pod)modela (prijevodni model, jezični model itd.) odvijaju u zasebnim koracima, tj. procesima, modeli nemaju optimalne parametre za dekodiranje, tj. strojno prevođenje. Stoga, primjenom višestrukog, ali ograničenog broja ponavljanja postupka ugađanja (eng. *tuning*), sustav postupno usklađuje vrijednosti pojedinih značajki modela za statističko strojno prevođenje (Koehn, 2010). Svakom prijevodnom kandidatu pridružen je skup pripadajućih vrijednosti težina značajki modela.

Podatkovni skup za treniranje n-gramskog jezičnog modela (eng. *monolingual training set*) čini jednojezični korpus ciljnog jezika (Koehn, 2015), pri čemu n-gram predstavlja niz uzastopnih riječi (Koehn, 2010). S obzirom da n-gramski jezični model ne može pokriti sve varijacije n-grama, postoji opasnost da se određenim nizovima riječi u postupku učenja jezičnog modela pridruže vjerojatnosti jednake nuli (Madnani, 2010). Stoga se pribjegava metodama izglađivanja (eng. *smoothing*) koje dotada neviđenim n-gramima pridružuju vjerojatnosti veće od nule, s obzirom da vjerojatnost nula izrazito loše utječe na procjenu vjerojatnosti niza riječi. Naime, distribucija vjerojatnosti tako postaje glađa (eng. *smoother*). Cilj izglađivanja je od viđenih n-grama oduzeti djelić vjerojatnosti te ga distribuirati među neviđenim n-gramima (Koehn, 2010). Nadalje, bolji jezični modeli kombiniraju procjene vjerojatnosti n-grama iz više različitih jezičnih modela (Madnani, 2010).

Ulazni podatkovni skup koji se koristi za treniranje prijevodnog modela sadrži rečenice, odnosno segmente na izvornom jeziku te semantičke prijevodne ekvivalente na ciljnom jeziku,

ekstrahirane iz nekog paralelnog korpusa (Koehn, 2005). Ukoliko je na raspolaganju ograničena količina kvalitetnog paralelnog korpusa, tada se sustav treba podesiti na način da veću težinu pridaje jezičnom modelu ciljnog jezika (Mauser et al., 2008). Upravo evaluacija automatskim metrikama može dati korisne smjernice na koji način podesiti težine modela (Gonzàlez, 2014).

Model sustava za statističko strojno prevođenje temeljeno na frazama posebno se obazire na leksičke i morfološke varijacije riječi, osobito važne kod morfološki bogatih jezika kao što je hrvatski. Učinkovito upravljanje leksičkim varijacijama riječi predstavlja snagu modela sustava za statističko strojno prevođenje temeljeno na frazama (Eisele et al., 2008). Takav model favorizira ili penalizira određene prijevode i šumove te se ne ograničava na fraze u lingvističkom smislu, kao sintaktički motivirane skupine riječi (Koehn, 2010). Čak naprotiv, ograničavanje na sintaktički motivirane fraze umanjuje kvalitetu statističkog strojnog prijevoda (Koehn et al., 2003). Naime, model funkcionira na statističkoj razini koristeći Bayesov teorem, Markovljev lanac i druge statističke paradigme za učenje iz paralelnih korpusa (Koehn, 2010; Manning i Schütze, 1999), a sam statistički strojni prijevod rezultat je niza ulančanih i uvjetovanih odluka koje se donose s određenom vjerojatnošću (Ueffing et al., 2007). Standardni model statističkog strojnog prevođenja sastoji se od tri ključne komponente: jezični model (eng. *language model*), prijevodni model (eng. *translation model*) i dekoder (eng. *decoder*) (Jurafsky i Martin, 2013). Bayesovim teoremom opisuje se vjerojatnost prevođenja segmenta iz izvornog $\boldsymbol{f}$ u ciljni jezik $\boldsymbol{e}$, što je prikazano jednadžbom (2.1) (Knight, 1999).

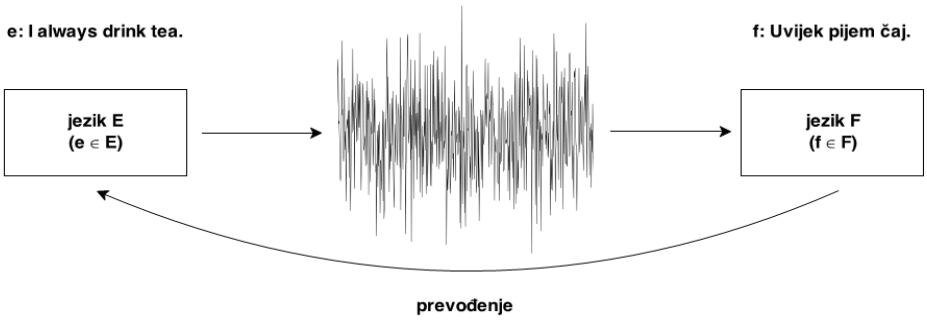
$$p(e|f) = \frac{p(e)p(f|e)}{p(f)} \tag{2.1}$$

S obzirom da je segment $\boldsymbol{f}$ konstantan u odnosu na sve moguće pripadajuće prijevode $\boldsymbol{e}$, $\boldsymbol{p(f)}$ se može zanemariti, što rezultira jednadžbom (2.2) (España-Bonet i Gonzàlez, 2014; Manning i Schütze, 1999).

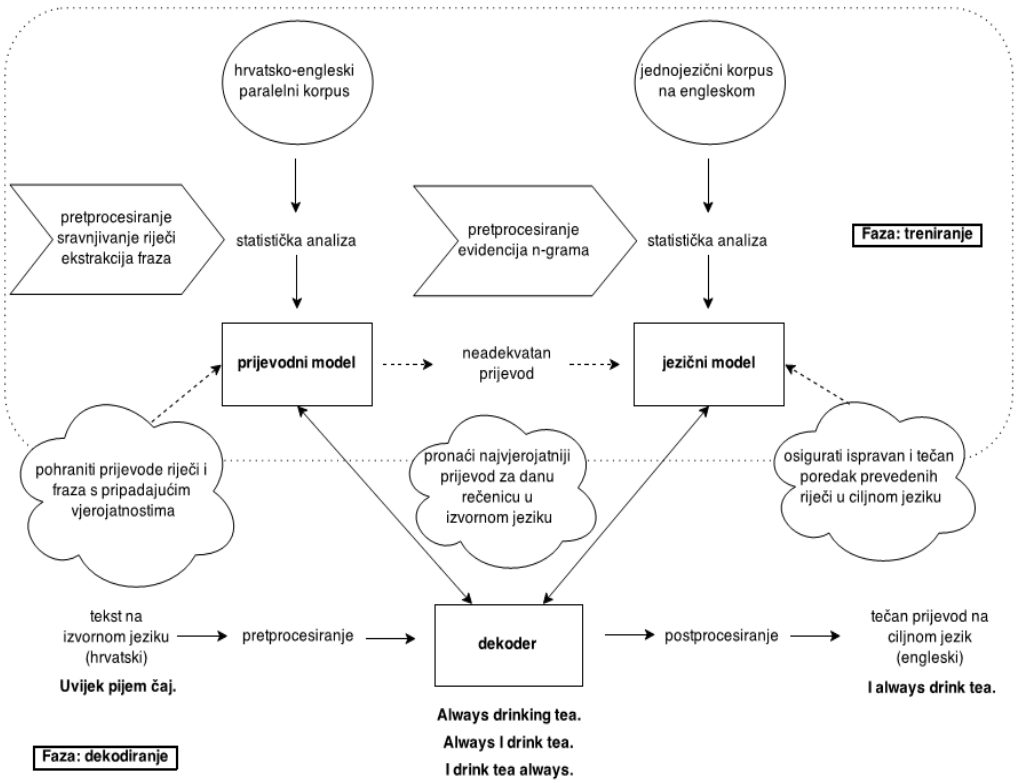
$$e = argmax_e p(e|f) = argmax_e p(e)p(f|e) \tag{2.2}$$

Takav pristup omogućuje razdvajanje i kombinaciju jezičnog modela $\boldsymbol{p(e)}$ i prijevodnog modela $\boldsymbol{p(f|e)}$, što ga svrstava u klasu modela kanala sa šumom (eng. *noisy-channel model*) (Koehn,

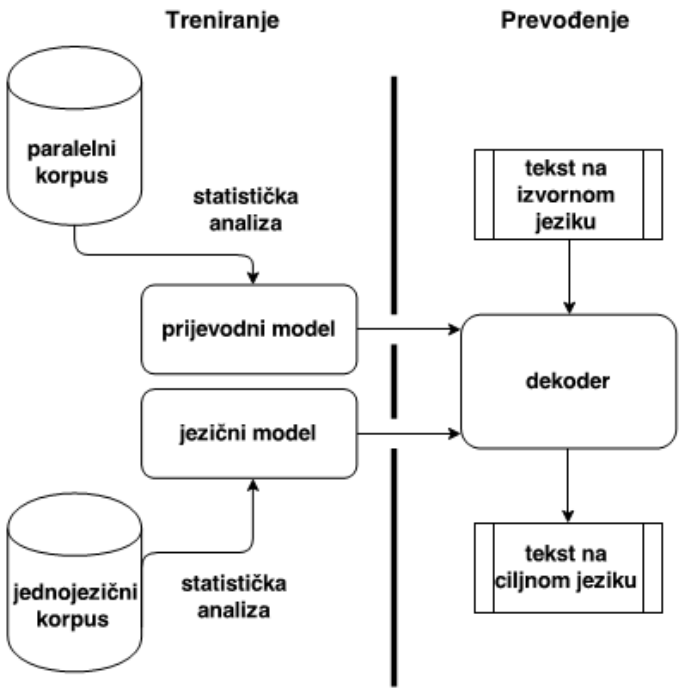
2010). U modelu kanala sa šumom, $\boldsymbol{f}$ predstavlja poruku, tj. rečenicu na izvornom jeziku koju treba prevesti u poruku, tj. rečenicu na ciljnom jeziku $\boldsymbol{e}$. Prijevodni model definira kako se ostvaruje prijevod poruke, tj. rečenice, dok jezični model procjenjuje koje su rečenice u ciljnom jeziku vjerojatne. Model kanala sa šumom prikazan je na jednom primjeru u nastavku (Slika 1) (adaptirano prema España-Bonet i Gonzàlez, 2014).



Slika 1. Prikaz modela kanala sa šumom.



Slika 2. Pojednostavljeni prikaz procesa u modelu sustava za statističko strojno prevođenje.



Slika 3. Model statističkog strojnog prevođenja s obzirom na fazu treniranja statističkih modela i prevođenja novog teksta.

Slike 2 i 3 prikazuju model sustava za statističko strojno prevođenje (adaptirano prema Koehn, 2004, odnosno Way i Hassan, 2009). Izvorni IBM-ovi modeli (Brown et al., 1993) nalažu da se proces strojnog prevođenja dekomponira u manje korake, pri čemu se svaki korak oslanja na riječi kao atomarne prijevodne jedinice (eng. *words as atomic translation units*) (Koehn, 2010).

Na kvalitetu strojnog prijevoda utječu i razlike između izvornoga i ciljnoga jezika. Naime, model statističkog strojnog prevođenja lakše prevodi s morfološki složenijeg jezika na manje složen jezik (Koehn, 2006).

Ispostavilo se, da se udvostručavanjem podatkovnih skupova, odnosno paralelnog korpusa za treniranje prijevodnog modela ili jednojezičnog korpusa za treniranje jezičnog modela, kontinuirano poboljšava i kvaliteta strojnog prijevoda, što se reflektira u rezultatima automatskih metrika (Turchi et al., 2012b).

Treba spomenuti da statistički pristup strojnom prevođenju ima i brojne poteškoće te izazove. Na ortografskoj/leksičkoj razini i najmanja pravopisna ili tipografska pogreška povećava vokabular i onemogućuje prevođenje riječi s obzirom da za vrijeme treniranja modela takva riječ nije statistički opisana. Zbog toga se prije izgradnje sustava za strojno prevođenje podatkovni skupovi trebaju adekvatno pripremiti, tj. pretprocesirati (eng. *data preprocessing*) (Koehn, 2015;

Buck et al., 2014). To znači da se riječi iz podatkovnih skupova prije treniranja modela pretvaraju u mala slova (eng. *lowercasing*) kako bi se izbjegle ortografske neusklađenosti među riječima.

Ipak, ispravan zapis riječi treba sačuvati. Proces pohranjivanja ispravnog zapisa početnog znaka riječi i postupak pretvaranja početnog znaka riječi u malo ili veliko slovo (eng. *truecasing*) može se primijeniti nakon dekodiranja novog podatkovnog skupa (eng. *recasing/capitalisation*). Nadalje, često se postupkom normalizacije uklanjaju riječi koje su semantički jednake, a treba izbjegavati i dvoznačne konstrukcije. Transliteracija predstavlja pretvaranje niza znakova iz jednog pravopisa u drugi, sačuvajući pri tome fonetiku u oba jezika (Manning i Schütze, 1999). Transliteracija odnosi se u pravilu na imena i brojeve (Koehn, 2010).

Postupkom tokenizacije (eng. *tokenisation*) se tekst dijeli u riječi, tj. umeću se razmaci između riječi i interpunkcije (u jezicima s latinskim alfabetom) (Manning i Schütze, 1999). Detokenizacija (eng. *detokenisation*) je suprotan proces kojim se tokenizirani tekst pretvara u prirodan oblik. Podatkovne skupove treba očistiti (eng. *corpus cleaning*), tj. treba ukloniti predugačke, prekratke, neuparene, nekompatibilne i prazne rečenice, odnosno segmente te suvišne razmake među riječima (Koehn, 2015).

Morfološki bogati jezici, poput hrvatskoga koji obiluje velikim brojem različitih morfema, također znatno povećavaju kompleksnost sustava za statističko strojno prevođenje. Sintaksa definira načela i pravila konstrukcije rečenica u prirodnom jeziku, međutim, ona nije integrirana u klasičnom modelu statističkog strojnog prevođenja temeljenog na frazama, s obzirom da su fraze obični nizovi riječi bez podataka o samoj strukturi niza riječi (Koehn, 2010). Zbog toga su problemi s redoslijedom riječi vrlo česti u statističkom strojnom prevođenju, pogotovo kada se prevodi s jezika s relativno slobodnim poretkom riječi u jezik s relativno fiksnim poretkom riječi kao što je engleski, ali i obrnuto.

Jednako tako, ukoliko izvorni i ciljni jezik imaju različite strukture (npr. subjekt-glagol-objekt i subjekt-objekt-glagol) problemi s ispravnim redoslijedom riječi su vrlo često neizbježni (Reddy i Hanumanthappa, 2013), naročito ukoliko se premještanje riječi unutar jedne rečenice treba izvršiti preko velike udaljenosti (eng. *long-range reordering*).

Na semantičkoj razini, značenje jedne rečenice izravno ovisi o dijelovima i odnosima unutar rečenice. Stoga se riječi koje čine jednu rečenicu, tj. neposredni kontekst, opisuju u jezičnom modelu ciljnoga jezika.

3.1. Jezični model

Jezični model $\boldsymbol{p(e)}$ osigurava fluentnost i prikladnost strojnog prijevoda u ciljnom jeziku, stoga se i trenira na korpusu ciljnoga jezika (eng. *monolingual training set*) (Koehn, 2015). Nadalje, jezični model pomaže pri odabiru točnih riječi i pri premještanju, tj. preslagivanju riječi (eng. *reordering*), što je prikazano na primjerima u nastavku (2.3) (Koehn, 2010).

$$p(danas\ je\ baš\ lijep\ dan) > p(danas\ je\ baš\ zgodan\ dan)$$
$$p(danas\ je\ baš\ lijep\ dan) > p(je\ danas\ dan\ lijep\ baš) \tag{2.3}$$

Jezični model može se zamisliti kao funkcija koja uzima rečenicu na ciljnom jeziku i vraća vjerojatnost da je ta rečenica u „duhu jezika“, tj. da odgovara stvarnoj upotrebi (Koehn, 2010). Drugim riječima, jezični model procjenjuje koliko je vjerojatan jedan segment rečenice u ciljnom jeziku. A koliko je vjerojatan jedan segment $\boldsymbol{e}$ u ciljnom jeziku može se procijeniti jednadžbom (2.4), pri čemu $\boldsymbol{N_e}$ predstavlja broj pojavljivanja određenog segmenta u korpusu ciljnoga jezika, a $\boldsymbol{N_n}$ predstavlja ukupan broj segmenata u korpusu ciljnog jezika (España-Bonet i Gonzàlez, 2014).

$$p(e) = \frac{N_e}{N_n} \tag{2.4}$$

Jezični modeli se u pravilu modeliraju pomoću uzastopnih riječi, tj. n-grama pa se takvi jezični modeli nazivaju i n-gramski jezični modeli (Manning i Schütze, 1999). Oni se oslanjaju na statistiku koja opisuje vjerojatnost da se određena riječ pojavi nakon neke druge riječi. Izračun uvjetne vjerojatnosti uzastopnih riječi (n-grama) može se pokazati na primjeru segmenata koji se sastoje od sekvencijalnog niza riječi $\boldsymbol{a}$, $\boldsymbol{b}$ te $\boldsymbol{a}$, $\boldsymbol{b}$ i $\boldsymbol{c}$ (jednadžbe 2.5 i 2.6). Radi se o bigramu i trigramu, tj. n-gramima koji se sastoje od dvije, odnosno tri uzastopne riječi. Takav pristup koji maksimizira vjerojatnost opisan je procjenom maksimalne vjerodostojnosti (eng. *maximum likelihood estimation, MLE*) kao što je prikazano u nastavku (Manning i Schütze, 1999).

$$p(b|a) = \frac{zbroj\_pojavljivanja\_niza\ ("a\ b")}{zbroj\_pojavljivanja\_riječi\ ("a\ ")} \tag{2.5}$$

$\boldsymbol{p(b|a)}$ je vjerojatnost da riječ $\boldsymbol{b}$ uslijedi nakon riječi $\boldsymbol{a}$, a $\boldsymbol{p(c|a\ b)}$ predstavlja vjerojatnost da se riječ $\boldsymbol{c}$ pojavi nakon bigrama, tj. uzastopnog niza riječi $\boldsymbol{a}$ i $\boldsymbol{b}$ (Knight, 1999).

$$p(c|a\ b) = \frac{zbroj\_pojavljivanja\_niza\ ("a\ b\ c")}{zbroj\_pojavljivanja\_niza\ ("a\ b")} \tag{2.6}$$

Procjena maksimalne vjerodostojnosti može se formalizirati kao u (2.7), pri čemu $\boldsymbol{c}$ (eng. *count*) predstavlja zbroj pojavljivanja (Manning i Schütze, 1999), tj. frekvenciju n-grama koji se sastoji od riječi $\boldsymbol{w_{i-1}}$ i $\boldsymbol{w_i}$.

$$p(w_i|w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})} \tag{2.7}$$

Izračuna maksimalne vjerodostojnosti na primjeru jednog trigrama dan je u nastavku .

uređaj za **... (ukupno 1397 pojava u korpusu)**		
riječ	**broj pojavljivanja**	**vjerojatnost**
mjerenje	625	0.447
ispitivanje	323	0.231
lakiranje	192	0.137
brušenje	154	0.110
održavanje	103	0.074

Iz gornjeg primjera proizlazi da se u jednom korpusu 1397 trigrama pojavilo počevši s riječima $\boldsymbol{uređaj\ za}$. Od toga, 625 trigrama završilo je riječima $\boldsymbol{mjerenje}$, što odgovara maksimalnoj vjerodostojnosti $\boldsymbol{p(mjerenje \mid uređaj\ za) = \frac{625}{1397} = 0.447}$.

Međutim, takav pristup ide na ruku kraćim segmentima za koje postoji izvjesna vjerojatnost da će se pojaviti u korpusu ciljnoga jezika. No, vrlo dugački segmenti se u pravilu rjeđe opažaju u korpusima, pa unatoč tome što su itekako mogući u jeziku i gramatički ispravni, to može rezultirati njihovom vjerojatnošću, $\boldsymbol{p(e) = 0}$ (ako je zbroj pojavljivanja, tj. frekvencija n-grama jednaka $\boldsymbol{0}$). S obzirom da veći $\boldsymbol{p(e)}$ upućuje na „bolji" i tečniji segment u ciljnom jeziku pribjegava se pristupu koji se temelji na kraćim dijelovima rečenica, tj. nizovima riječi (frazama).

Jezični model opisuje uzastopne nizove riječi pomoću n-grama i osigurava fluentan prijevod na ciljni jezik prebrojavanjem n-grama u jednojezičnom korpusu ciljnoga jezika (Latour, 2004). U jezičnom modelu pohranjuju se vjerojatnosti $\boldsymbol{p(e)}$ za razne kombinacije nizova riječi $\boldsymbol{w_i}$ u ciljnom jeziku, a mogu se opisati Markovljevim lancima, što je prikazano jednadžbom (2.8). Naime, n-gramski jezični modeli temelje se na Markovljevoj pretpostavci (eng. *Markov assumption*) da se vjerojatnost jedne rečenice može opisati kao produkt vjerojatnosti svake riječi u rečenici ako je dan određen broj prethodnih riječi (Koehn, 2010).

$$p(e) = p(w_1, w_2, w_3, \dots, w_n) = \\ p(w_1)p(w_2|w_1)p(w_3|w_1, w_2) \dots p(w_n|w_1, w_2, \dots, w_{n-1}) \tag{2.8}$$

U Markovljevim lancima na vjerojatnost jedne riječi utječe samo ograničen broj prethodnih stanja (eng. *history*), tj. prethodnih riječi (Koehn, 2010). Iz semantičke i lingvističke perspektive to i nije točno (utječe velik broj faktora), međutim, tehnički je najjednostavnije jezik opisati n-gramskim modelom.

Vjerojatnost jezičnog modela $\boldsymbol{p(w_1, w_2, w_3, \dots, w_n)}$ jednaka je umnošku vjerojatnosti riječi s obzirom na riječi koje su joj prethodile. Međutim, Markovljevi lanci kao matematički alat imaju ograničenje glede kapaciteta prethodnih riječi, tj. uzima se u obzir samo zadnjih $\boldsymbol{k}$ riječi. Stoga se za takav Markovljev model kaže se da je $\boldsymbol{k\text{-}tog}$ reda (Koehn, 2010). Formalno, vjerojatnost niza uzastopnih riječi jednaka je umnošku uvjetnih vjerojatnosti za svaku riječ $\boldsymbol{w_i}$ uzevši u obzir prethodna stanja, tj. riječi koje su se prethodno pojavile u nizu riječi. Vjerojatnost pojavljivanja riječi $\boldsymbol{w_i}$ uvjetovana je isključivo pojavljivanjem prethodnih $\boldsymbol{n}$, tj. $\boldsymbol{k}$ riječi.

Jezični model se u pravilu izgrađuje na temelju 3-grama ili 4-grama (Koehn, 2015), međutim, broj n-grama može biti i mnogo veći. Što je model složeniji (većeg reda) to je moguće opisati i dulje nizove uzastopnih riječi odgovarajućom vjerojatnošću. U nastavku je prikazana rečenica ***Danas je baš lijep dan***, opisana u 3-gramskom jezičnom modelu, koja za predviđanje treće riječi po redu, uzima u obzir samo prethodne dvije riječi (2.9) (España-Bonet i Gonzàlez, 2014).

$$p(e) = p(Danas|\emptyset,\emptyset)p(je|\emptyset,Danas)p(baš|Danas,je)\ldots \\ p(lijep|je,baš)p(dan|baš,lijep) \tag{2.9}$$

Svaki faktor u gornjoj jednadžbi računa se sljedećom analogijom (jednadžba 2.10) (España-Bonet i Gonzàlez, 2014):

$$p(je|\emptyset,Danas) = \frac{N_{(Danas\ je)}}{N_{(Danas)}} \tag{2.10}$$

$\boldsymbol{N_{(Danas\ je)}}$ i $\boldsymbol{N_{(Danas)}}$ označavaju ukupan broj pojavljivanja bigrama ***Danas je***, odnosno unigrama ***Danas*** u korpusu. Međutim, takvim pristupom se pomoću uvjetnih vjerojatnosti, tj. Markovljevog lanca, ne mogu opisati dulji nizovi riječi. Primjerice, 3-gram ***baš lijep dan*** ne obuhvaća riječi ***Danas je*** s početka rečenice te se na taj način može izgubiti važna informacija. Naime, da bi rečenica ***Danas je baš lijep dan*** sigurno bila opisana pomoću uvjetne vjerojatnosti, potrebno je izgraditi jezični model koji se temelji na 5-gramima i koji opisuje sve potrebne faktore u Markovljevom lancu.

No, s obzirom da s povećavanjem broja n-grama u jezičnom modelu raste i njegova kompleksnost, takav model bi bio nepraktičan, znatno složeniji i sporiji te bi zahtijevao više računalnih resursa (diskovnog prostora, radne memorije itd.) (Koehn, 2010).

Zadatak jezičnog modela jest procijeniti, odnosno predvidjeti (eng. *predict*) koja riječ slijedi nakon određenog niza riječi, kao što je prikazano na jednom primjeru u nastavku (2.11) (Jurafsky, 2015). Treba naglasiti da se istovremeno uvijek predviđa samo jedna riječ.

$$p(w_5|w_1, w_2, w_3, w_4) \tag{2.11}$$

Svaki n-gramski jezični model uzima u obzir samo prvih $\boldsymbol{n-1}$ prethodnih riječi koji čine prijašnji kontekst (eng. *history)* riječi $\boldsymbol{n}$ koju treba predvidjeti (Mooney, 2015). Mjera koja odgovara na pitanje koliko dobro jedan jezični model predviđa riječ (iz nekog testnog korpusa) s obzirom na dani kontekst naziva se perpleksnost (eng. *perplexity*) (Koehn, 2010). Perpleksnost je mjera neizvjesnosti u distribuciji vjerojatnosti, ukazuje na kvalitetu jezičnog modela te dozvoljava usporedbu različitih n-gramskih modela. Pripada kategoriji intrinzične evaluacije s obzirom da se ispituje neovisno o specifičnom okruženju i namjeni (Jurafsky, 2015).

S obzirom da vrijedi (2.12), pri čemu $\boldsymbol{b}$ predstavlja bazu logaritma, perpleksnost (2.14) se može izračunati pomoću ukrštene entropije (eng. *cross entropy*), kao u (2.13) (Sennrich, 2012b; Koehn, 2010; Manning i Schütze, 1999). Entropija odražava razinu neizvjesnosti jednog događaja. Ukoliko jedna distribucija ima vrlo sigurne ishode (događaje) tada je i entropija niska. Teorija informacije intuitivno reinterpretira entropiju kao broj bitova potrebnih po događaju za kodiranje poruke. Ukrštena entropija nastoji aproksimirati entropiju jezičnog modela (Koehn, 2010) s obzirom da nije poznata stvarna distribucija vjerojatnosti konteksta jednog jezika, već se koristi jezični model kao reprezentant jezika.

$$log_b(mn) = log_b(m) + log_b(n) \tag{2.12}$$

$$H(p) = -\frac{1}{n} log_2 p(w_1, w_2, \dots, w_n) = -\frac{1}{n} \sum_{i=1}^{n} log_2 p(w_i | w_1, \dots, w_{i-1}) \tag{2.13}$$

$$perpleksnost = 2^{H(p)} \tag{2.14}$$

Primjer izračuna perpleksnosti dan je u nastavku (Tablica 1).

Tablica 1. Izračun perpleksnosti na primjeru rečenice „**Danas je baš lijep dan.**“ u 3-gramskom jezičnom modelu.

predviđanje	$\boldsymbol{p(e)}$	$\boldsymbol{-log_2 p(e)}$
p(danas | </s><s>)	0.231	2.1140
p(je | <s>danas)	0.459	1.1234
p(baš | danas je)	0.144	2.7959
p(lijep | je baš)	0.203	2.3004
p(dan | baš lijep)	0.092	3.4422
p(. | lijep dan)	0.281	1.8313
p(</s> | dan .)	0.999	0.0014
$\sum_{i=1}^{n} \boldsymbol{log_2 p(w_i \mid w_1, \dots, w_{i-1})}$		13.6086
trigramski jezični model		$n = 3$
$\boldsymbol{H(p)}$		4.5362
perpleksnost		23.2023

Rečenica ***Danas je baš lijep dan.*** označena je oznakama za početak $(< \boldsymbol{s} >)$ i kraj rečenice $(</\boldsymbol{s} >)$ (Manning i Schütze, 1999). To omogućuje ispravno modeliranje vjerojatnosti za riječ koja se pojavljuje na prvom mjestu u rečenici (***Danas***). Jednako tako moguće je modelirati i kraj rečenice. Iz gornjeg primjera proizlazi da je vjerojatnost da rečenica započne s ***Danas*** $\mathbf{0.231}$. Ukrštena entropija jednaka je aritmetičkoj sredini logaritamskih vrijednosti vjerojatnosti pojavljivanja jedne riječi. Jezični modeli višeg reda (npr. $\boldsymbol{n = 4}$) mogu bolje predvidjeti sljedeću riječ u nizu s obzirom da su opisani većim kontekstom (npr. tri prethodne riječi), što znatno pojednostavljuje predviđanje. Shodno tome je i perpleksnost niža.

U pravilu, pri usporedbi dvaju jezičnih modela, model s nižom perpleksnošću je bolji jezični model. Drugim riječima, bolji jezični model je onaj model koji bolji predviđa novu riječ koja prije toga nije viđena u korpusu za treniranje jezičnog modela (Jurafsky, 2015).

Perpleksnost od $\mathbf{23.2023}$ iz gornjeg primjera znači da jezični model u prosjeku jednu riječ predviđa među 23 moguće (Chhatbar, 2010), tj. vjerojatnost da model pogodi ispravnu riječi je $\frac{\mathbf{1}}{\mathbf{23}}$. Manja perpleksnost modela ukazuje na to da model bira (predviđa) riječ iz manjeg skupa mogućih riječi (Wandmacher, 2008), i da je predviđanje stoga manje neizvjesno. Međutim, treba svakako uzeti u obzir i veličinu vokabulara. Niža perpleksnost može se postići sa specijaliziranim

korpusima s vrlo specifičnom terminologijom, s obzirom da je n-grame u takvom korpusu lakše predvidjeti. Ukoliko se perpleksnost ispituje na skupu izvan domene na kojoj je treniran jezični model, to će rezultirati većom perpleksnošću (Madnani, 2010).

Ključan problem u n-gramskim jezičnim modelima je upravljanje oskudnim podatcima (eng. *sparse data*). Naime, to što n-gramskim jezičnim modelom nisu pokrivene apsolutno sve kombinacije nizova riječi, ne znači da takve kombinacije nisu gramatički ispravne ili statistički moguće u nekom korpusu. Ako se npr. n-gram $\boldsymbol{baš\ lijep\ dan}$ pojavio u korpusu na kojemu je treniran jezični model, bit će mu dodijeljena i određena vjerojatnost $\boldsymbol{p(dan|baš\ lijep)}$. Ukoliko se zatim primjerice 3-gramski jezični model ispita na jednom korpusu, tj. podatkovnom skupu za testiranje (eng. *test set*), koji primjerice ne sadrži n-gram $\boldsymbol{baš\ lijep\ dan}$ već n-gram $\boldsymbol{baš\ lijep\ odmor}$, to će rezultirati njegovom vjerojatnošću $\mathbf{0}$, s obzirom da se takav n-gram ranije nije pojavio u jezičnom modelu, tj. njegov zbroj pojavljivanja (frekvencija) je nula (2.15).

$$p(odmor|baš\ lijep) = 0 \tag{2.15}$$

Nadalje, svakoj rečenici koja sadrži n-gram $\boldsymbol{baš\ lijep\ odmor}$ bit će također dodijeljena vjerojatnost $\mathbf{0}$, iako je ona gramatički ispravna i realno moguća u nekom korpusu. S obzirom da postoji velik broj takvih kombinacija riječi, procjena maksimalne vjerodostojnosti takvim će rijetkim, ali mogućim, kombinacijama nizova riječi uvijek pridružiti vjerojatnost $\mathbf{0}$. Tj., svaki put kad se pojavi kombinacija niza riječi koja se prethodno nije pojavila u korpusu za treniranje (zbroj pojavljivanja je $\mathbf{0}$), bit će joj dodijeljena vjerojatnost $\mathbf{0}$, što će rezultirati ukupnom vjerojatnošću $\mathbf{0}$ i za cijelu rečenicu koja sadrži takav niz riječi, a to će prouzročiti beskonačnu perpleksnost (Mooney, 2015). Samim dodavanjem nove količine korpusa za treniranje, u nadi da se pokrije veći broj kombinacija riječi, tj. n-grama, također se ne rješava problem. Taj fenomen opisan je Zipfovim empirijskim zakonom (Koehn, 2010) koji implicira da je frekvencija bilo koje riječi obrnuto proporcionalna sa svojim rangom u tablici frekvencija riječi (Manning i Schütze, 1999). Tj. Zipfova distribucija riječi ukazuje na to da se najfrekventnija riječ pojavljuje oko dva puta češće nego druga najfrekventnija riječ, tri puta češće nego treća najfrekventnija riječ itd.

Kako nije moguće opaziti sve teoretski moguće n-grame u ograničenom, tj. konačnom podatkovnom skupu za treniranje jezičnog modela, pribjegava se metodama izglađivanja (eng. *smoothing*) koje regulariziraju, tj. izglađuju nizove riječi s vjerojatnošću $\mathbf{0}$ (Koehn, 2010). Metode izglađivanja takvoj, u korpusu za treniranje jezičnog modela, neviđenoj kombinaciji riječi

pridružuju vjerojatnost na način da se preraspodijele vjerojatnosti koje su već pridružene drugim viđenim nizovima riječi (Koehn, 2010).

Odnosno, pridruživanje vjerojatnosti, takvim do tada neviđenim nizovima riječi, pretpostavlja da se od viđenih n-grama oduzima dio vjerojatnosti (eng. *discounting*) koji se zatim preraspodjeljuje do ukupne vjerojatnosti $\mathbf{1}$, s obzirom da zbroj vjerojatnosti u združenoj distribuciji, tj. razdiobi vjerojatnosti mora iznositi $\mathbf{1}$ (Koehn, 2010).

Najjednostavnija metoda izglađivanja je tzv. aditivna metoda. Jedna od takvih aditivnih metoda jest i Laplaceova metoda dodavanja 1 (eng. *add-one smoothing*) (Manning i Schütze, 1999), kao što je definirano u (2.16), pri čemu $\boldsymbol{c}$ predstavlja zbroj pojavljivanja n-grama u testnom podatkovnom skupu, koji se prethodno nisu pojavili u korpusu za treniranje jezičnog modela (tzv. neviđeni n-grami), $\boldsymbol{n}$ ukupan broj n-grama u korpusu, a $\boldsymbol{v}$ ukupan broj mogućih n-grama (sve kombinacije uzastopnih riječi pa makar bile i besmislene) (Koehn, 2010). Laplaceova metoda izvedena je iz procjene maksimalne vjerodostojnosti, a zamišlja korpus u kojemu se svaki mogući n-gram pojavio točno još jedanput više nego što se zaista pojavio, nakon čega se s obzirom na to preraspodjeljuju vjerojatnosti.

$$p = \frac{c}{n} \Longrightarrow p = \frac{c+1}{n+v} \tag{2.16}$$

Međutim, takva metoda pridružuje previše vjerojatnosti besmislenim kombinacijama riječi koje se nikada ne opažaju u stvarnom korpusu, što umanjuje vjerojatnost onih kombinacija koje se zasigurno pojavljuju. Zbog toga se Laplaceova metoda može modificirati (eng. *add-alpha smoothing*) tako da se umjesto $\mathbf{1}$ doda manji broj $\boldsymbol{\alpha}$, pri čemu vrijedi $\mathbf{0} < \boldsymbol{\alpha} < 1$ (2.17) (Koehn, 2010).

$$p = \frac{c+\alpha}{n+\alpha v} \tag{2.17}$$

Nije jednostavno odrediti vrijednost $\boldsymbol{\alpha}$, uglavnom se podešava eksperimentiranjem s vrijednošću i praćenjem rezultirajuće perpleksnosti. No, što je manji $\boldsymbol{\alpha}$, to se manje vjerojatnosti može preraspodijeliti, a s obzirom da pri $\boldsymbol{\alpha} = \mathbf{0}$ nema izglađivanja, ova metoda je također neefikasna (Madnani, 2010).

Zbog toga se primjenjuju naprednije metode izglađivanja, kao što su Good-Turing, interpolacija, metoda povlačenja (eng. *back-off*), Witten-Bell, Kneser-Ney itd. (MacCartney, 2005; Manning i Schütze, 1999; Chen i Goodman, 1996).

Jezični modeli višeg reda i nižeg reda (red je određen brojem prethodnih stanja) imaju različite prednosti i nedostatke. Modeli višeg reda pružaju veći kontekst, međutim, generiraju veći broj neviđenih n-grama. Modeli nižeg reda posjeduju ograničen kontekst, međutim, neviđeni n-grami su znatno rjeđi, što takav model čini robusnijim (Koehn, 2010).

Ideja interpolacije jest kombinirati n-gramske jezične modele višeg i nižeg reda te tako sačuvati n-grame koji se rijetko pojavljuju. Za interpolaciju potrebno je prvo izgraditi različite n-gramske jezične modele $\boldsymbol{p_n}$. Jezični modeli se prije same interpolacije mogu izgladiti, međutim, to nije nužno, s obzirom da sama interpolacija može ublažiti problem neviđenih n-grama (Koehn, 2010). Zatim se izgrađuje interpolirani jezični model $\boldsymbol{p_I}$ koji linearno kombinira jezične modele $\boldsymbol{p_n}$. Linearno interpolirani model može se formalizirati kao u (2.18), pri čemu $\boldsymbol{\lambda_i}$ predstavlja interpolacijsku težinu (eng. *weight*) svakog modela $\boldsymbol{p_i}$ (Sennrich, 2012b; Koehn, 2010).

$$p(x|y;\lambda) = \sum_{i=1}^{n} \lambda_i p_i(x|y) \tag{2.18}$$

Svaki jezični model doprinosi do određene razine, što se može opisati faktorom, tj. interpolacijskom težinom. Ukoliko se jezični model trenira na velikoj količini podataka, tj. na velikim korpusima, povjerenje u jezične modele višeg reda raste, a to omogućuje pridruživanje veće težine $\boldsymbol{\lambda}$. Da bi se osiguralo da interpolirani jezični model $\boldsymbol{p_I}$ ispravno distribuira vjerojatnosti, naredni uvjeti moraju biti zadovoljeni (2.19, 2.20) (Koehn, 2010; Manning i Schütze, 1999).

$$\forall \lambda_n : 0 \leq \lambda_n \leq 1 \tag{2.19}$$

$$\sum_{n} \lambda_n = 1 \tag{2.20}$$

Trigramski jezični model može se interpolirati s unigramskim i bigramskim modelom kao u (2.21), iz čega proizlazi da se linearna interpolacija temelji na procjeni maksimalne vjerodostojnosti.

$$p_I(w_3|w_1, w_2) = \lambda_1 p_1(w_3) + \lambda_2 p_2(w_3|w_2) + \lambda_3 p_3(w_3|w_1, w_2) \tag{2.21}$$

Primjer linearne interpolacije dan je u nastavku (2.22) (España-Bonet i Gonzàlez, 2014).

$$\begin{aligned} p_I(dan|baš, lijep) \\ = \lambda_1 \frac{N_{(dan)}}{N_{ukupno\_riječi}} + \lambda_2 \frac{N_{(lijep,dan)}}{N_{(lijep)}} + \lambda_3 \frac{N_{(baš,lijep,dan)}}{N_{(baš,lijep)}} \\ + \lambda_0 \end{aligned} \tag{2.22}$$

Vrijednosti težina $\boldsymbol{\lambda_1}$ unigramskog modela, odnosno težina $\boldsymbol{\lambda_2}$ bigramskog i $\boldsymbol{\lambda_3}$ trigramskog modela mogu se optimizirati pomoću posebnog korpusa (eng. *held-out data*) prema kojemu se žele podesiti vrijednosti težina, tj. za koji se želi maksimizirati učinak (vjerojatnost) (Manning i Schütze, 1999).

Jezični modeli višeg reda dozvoljavaju opis većeg konteksta i u pravilu predstavljaju bolje jezične modele (Koehn, 2010). No, s obzirom da je korpus za treniranje jezičnog modela ograničen, mnoštvo mogućih fluentnih n-grama višeg reda neće biti opaženo u korpusu za treniranje. Za svaki viđeni n-gram u korpusu za treniranje jezičnoga modela može se procijeniti vjerojatnost predviđanja riječi. No, za neviđene n-grama u korpusu to nije moguće.

Stoga se može primijeniti i metoda povlačenja prema jezičnom modelu nižeg reda (*back-off*), s obzirom da je ponekad poželjno koristiti manje konteksta (Jurafsky, 2015). Metoda povlačenja načelno uvijek koristi n-gramski jezični model najvišeg reda u kojemu je predviđena riječ opisana prethodnim riječima (eng. *history*). U suprotnom, *back-off* metoda nalaže povlačenje prema n-gramskim modelima nižeg reda opisanima s manjim kontekstom (tj. manjim brojem prethodnih riječi, eng. *shorter history*). Odnosno, ukoliko postoji dovoljno podataka za (vjerojatno) ispravnu konstrukciju n-grama u jezičnom modelu višeg reda, tada će se i on preferirati, dok u ostalim slučajevima pribjegava se jezičnim modelima nižeg reda koji mogu pružiti više statističkih podataka o n-gramu.

Model nižeg reda nije opisan velikim kontekstom, međutim, znatno je robusniji. Primjerice, trigrami ***hrvatski notorni primjer*** i ***hrvatski notorni primjerak*** možda neće biti opaženi, ukoliko se u korpusu za treniranje nije pojavio bigram ***hrvatski notorni***. Ovakav prijevod na ciljni jezik je moguć ukoliko se prevodi s izvornog jezika u kojemu jedna riječ (npr. ***example***) ima jednaku vjerojatnost da bude prevedena s dvije riječi (npr. ***primjer*** i ***primjerak***). Cilj jezičnog modela je ukazati na to da je ***hrvatski notorni primjer*** ispravan odabir n-grama.

Metode izglađivanja pridruživanje određene vjerojatnosti neviđenim n-gramima, kod kojih je zbroj pojavljivanja, tj. frekvencija jednaka nula, vrše preraspodjelom vjerojatnosti za viđene n-grame. Međutim, aditivne metode jednako tretiraju neviđene n-grame s jednakim brojem pojavljivanja i stoga ne mogu pomoći u razlučivanju trigrama ***hrvatski notorni primjer*** i ***hrvatski notorni primjerak***, i u odabiranju ispravnog n-grama u ciljnom jeziku, s obzirom da će metoda obojici trigrama pridružiti jednaku vjerojatnost. No, ideja *back-off* metode jest vratiti se u model nižeg reda, u ovom slučaju u bigramski model, i pretražiti bigrame ***notorni primjer*** i ***notorni primjerak***, jer se pukim dodavanjem riječi ***hrvatski*** i time proširivanjem konteksta u jezičnom modelu ne može riješiti problem. Jedna mogućnost je povlačenje prema bigramskom modelu koji možda bolje može razlikovati ta dva bigrama, s obzirom da je vjerojatnije da se za bigrame ***notorni primjer*** i ***notorni primjerak*** iz korpusa izvede više zaključaka o njihovim pojavnostima. Ukoliko povlačenje prema bigramskom modelu ne rezultira zadovoljavajućim odgovorima, povlačenje se može nastaviti i prema unigramskom modelu (Jurafsky, 2015). Metoda povlačenja može se formalizirati kao u (2.23) (Koehn, 2010).

$$p_n^{BO}(w_i|w_{i-n+1},\dots,w_{i-1}) = \begin{cases} d_n(w_{i-n+1},\dots,w_{i-1})p_n(w_i|w_{i-n+1},\dots,w_{i-1}) \; if \; zbroj_n(w_{i-n+1},\dots,w_i) > 0 \\ \alpha_n(w_{i-n+1},\dots,w_{i-1})p_{n-1}^{BO}(w_i|w_{i-n+2},\dots,w_{i-1}) \qquad else \end{cases} \quad (2.23)$$

Načelno se uvijek koristi najsloženiji (ili preferirani) model koji može opisati odgovarajući kontekst. Međutim, ukoliko to nije dovoljno (zbroj pojavljivanja određenog n-grama je i dalje **0**), metoda nalaže rekurzivno povlačenje prema modelu/-ima nižeg reda. Postupak povlačenja se ponavlja sve dok se ne dostigne model iz kojeg je moguće izvući podatke o n-gramu (Madnani, 2010). Tu se metoda povlačenja i ponajviše razlikuje od linearne interpolacije. Naime, metoda

povlačenja koristi određene modele ovisno o potrebi, za razliku od linearne interpolacije gdje se svi modeli različitog reda koriste istovremeno.

Metodom povlačenja se izgrađuje *back-off* jezični model $\boldsymbol{p_n^{BO}}$ na temelju jezičnih modela $\boldsymbol{p_n}$. Ako se pojavi n-gram koji je prethodno pronađen u skupu za treniranje jezičnog modela (tj. $\boldsymbol{zbroj_n(\dots) > 0}$), koristit će se vjerojatnost jezičnog modela $\boldsymbol{p_n}$. Za n-gram koji prethodno nije viđen koristit će se *back-off* jezični model nižeg reda $\boldsymbol{p_{n-1}^{BO}}$ (Koehn, 2010). Drugim riječima, u početku se koristi model najvišeg reda, a povlačenje prema modelu nižeg reda primjenjuje se samo ukoliko n-gram nije pronađen u modelu višeg reda. S obzirom da treba osigurati da je zbroj svih vjerojatnosti za niz prethodnih riječi $\boldsymbol{w_{i-n+1}, \dots, w_{i-1}}$ jednak $\boldsymbol{1}$, uvodi se funkcija diskontiranja (eng. *discounting function*) $\boldsymbol{d_n}$. Vjerojatnosti iz jezičnog modela $\boldsymbol{p_n}$ se diskontiraju, tj. preraspodjeljuju faktorom $\boldsymbol{0 \leq d \leq 1}$, a ostatak mase vjerojatnosti (eng. *probability mass*) se distribuira *back-off* jezičnom modelu (Koehn, 2010). Funkcija diskontiranja ovisit će o nizu prethodnih riječi n-grama (eng. *histories*), a *back-off* model o modelu predviđanja $\boldsymbol{\alpha_n}$. Ideja je grupirati prethodna stanja, tj. nizove prethodnih riječi (eng. *group histories*), na temelju njihovih frekvencija u korpusu. Ako se niz prethodnih riječi pojavio vrlo često, više će se vjerovati jezičnom modelu najvišeg reda koji sadrži n-gram te se stoga postavlja veća vjerojatnost za $\boldsymbol{d_n(w_1, \dots, w_{n-1})}$. S druge pak strane, za rjeđe nizove prethodnih riječi (eng. *rare histories*) pretpostavlja se da je samo djelić mogućih predviđenih riječi $\boldsymbol{w_n}$ viđen, te se zbog toga *back-off* modelu pridaje veća „težina“, što će uzrokovati nižu vrijednost za $\boldsymbol{d}$ (Koehn, 2010).

Kneser-Ney metoda jedna je od najkorištenijih metoda izglađivanja u strojnom prevođenju (Koehn, 2010). Temelji se na frekvencijama i raznolikostima prethodnog niza riječi (eng. *history diversity*) u odnosu na jednu riječ. Primjerice, riječ $\boldsymbol{Brod}$ se u jednom korpusu može pojaviti jednako često kao i bilo koja druga riječ, npr. $\boldsymbol{trči}$. Stoga u unigramskom jezičnom modelu može imati vrlo visoku vjerojatnost pojavljivanja. Međutim, uočeno je da se vrlo često riječ $\boldsymbol{Brod}$ pojavljuje nakon riječi $\boldsymbol{Slavonski}$, čineći time naziv grada $\boldsymbol{Slavonski\ Brod}$. Kao što je već ranije rečeno, unigramski jezični model može se koristiti kao *back-off* model ukoliko bigramski model ne daje zadovoljavajuće odgovore. Iz gornjeg primjera proizlazi da je vrlo vjerojatno da se riječ $\boldsymbol{Brod}$ pojavi na drugoj poziciji u bigramu, ukoliko je prije toga viđen $\boldsymbol{Slavonski}$. Iako se možda riječ $\boldsymbol{Brod}$ često pojavljuje zasebno u korpusu, u *back-off* unigramskom modelu toj riječi treba smanjiti vjerojatnost bez obzira na njenu frekvenciju pojavljivanja. To znači da se veća težina pridaje raznolikosti prethodnog niza riječi. Zbroj različitih prethodnih nizova riječi za jednu riječ (eng. *count of histories for a word*) može se izračunati prema (2.24) (Koehn, 2010).

$$N_{1+}(\bullet\, w) = |\{w_i : c(w_i, w) > 0\}| \tag{2.24}$$

Umjesto primjene klasičnog zbroja pojavljivanja riječi u korpusu (kao u metodi procjene maksimalne vjerodostojnosti), u Kneser-Ney metodi izglađivanja primjenjuje se zbroj, tj. frekvencija prethodnih nizova riječi ($\boldsymbol{N_{1+}(\bullet\, w)}$), kao što je prikazano u (2.25) (Koehn, 2010).

$$p_{KN}(w) = \frac{N_{1+}(\bullet\, w)}{\sum_{w_i} N_{1+}(\bullet\, w_i)} \tag{2.25}$$

Ukoliko su primjerice za riječ $\boldsymbol{Brod}$ identificirana samo dva prethodna niza riječi (Slavonski Brod, Tvrđava Brod), a za riječ $\boldsymbol{trči}$ šest prethodnih niza riječi (on trči, ona trči, Željko trči, zašto trči, ne trči, najbrže trči), tada će smanjivanje vjerojatnosti (eng. *basing probabilities*) pomoću zbroja prethodnog niza riječi rezultirati znatno većom vjerojatnošću povlačenja (eng. *back-off probability*) prema unigramskom modelu za riječ $\boldsymbol{trči}$ nego za riječ $\boldsymbol{Brod}$.

Modificirana inačica Kneser-Ney metode izglađivanja temelji se na kombinaciji interpolacije i metode *back-off*, a može se formalizirati kao u (2.26) (Koehn, 2010).

$$p_I(w_n|w_1, \dots, w_{n-1}) = \begin{cases} \alpha(w_n|w_1, \dots, w_{n-1}) & if\ c(w_1, \dots, w_n) > 0 \\ \gamma(w_1, \dots, w_{n-1}) p_I(w_n|w_2, \dots, w_{n-1}) & else \end{cases} \tag{2.26}$$

U gornjoj jednadžbi implementirane su dvije funkcije. Funkcija $\boldsymbol{\alpha}$ povezuje svaki n-gram u korpusu s pripadajućom vjerojatnošću. Za svaku instancu prethodnog niza riječi postoji funkcija $\boldsymbol{\gamma}$ koja se odnosi na masu vjerojatnosti rezerviranu za neviđene riječi koje slijede niz prethodnih riječi (Koehn, 2010). Modificirana inačica Kneser-Ney metode izglađivanja (eng. *modified Kneser-Ney smoothing*) koristi interpolirano povlačenje (eng. *interpolated back-off*) te relativno diskontiranje, što efikasno umanjuje masu vjerojatnosti koje su namijenjene viđenim n-gramima (Jurafsky i Martin, 2013; Chen i Goodman, 1999). Funkcija $\boldsymbol{\alpha}$ za n-gramski model najvišeg reda (eng. *highest order n-gram model*) može se izračunati kao u (2.27) (Koehn, 2010).

$$\alpha(w_n|w_1,\dots,w_{n-1}) = \frac{c(w_1,\dots,w_n) - D}{\sum_w c(w_1,\dots,w_{n-1},w)} \tag{2.27}$$

Za n-gramski jezični model najvišeg reda, vrijednost diskontiranja (eng. *discounting value*) $\boldsymbol{D}$ oduzima se od svakog zbroja n-grama. Vrijednost $\boldsymbol{D}$ nije fiksna, već ovisi o frekvencijama n-grama $\boldsymbol{c}$ (Koehn, 2010; Chen i Goodman, 1998) (2.28, 2.29).

$$D(c) = \begin{cases} 0 & ako\ c = 0 \\ D_1 & ako\ c = 1 \\ D_2 & ako\ c = 2 \\ D_{3+} & ako\ c \geq 3 \end{cases} \tag{2.28}$$

Optimalne vrijednosti $\boldsymbol{D_1}$, $\boldsymbol{D_2}$ i $\boldsymbol{D_{3+}}$ mogu se izračunati kao što je dano u (2.29), pri čemu $\boldsymbol{N_c}$ predstavlja frekvenciju n-grama koji su se pojavili točno $\boldsymbol{c}$ puta (Koehn, 2010).

$$\begin{aligned} Y &= \frac{N_1}{N_1 + 2N_2} \\ D_1 &= 1 - 2Y\frac{N_2}{N_1} \\ D_2 &= 2 - 3Y\frac{N_3}{N_2} \\ D_{3+} &= 3 - 4Y\frac{N_4}{N_3} \end{aligned} \tag{2.29}$$

Funkcija $\boldsymbol{\gamma}$ može se definirati kao u (2.30) (Koehn, 2010).

$$\gamma(w_1,\dots,w_{n-1}) = \frac{\sum_{i\in\{1,2,3+\}} D_i\, N_i(w_1,\dots,w_{n-1}\,\bullet)}{\sum_{w_n} c(w_1,\dots,w_n)} \tag{2.30}$$

$\boldsymbol{N_i}$ za $\boldsymbol{i} \in \{\mathbf{1}, \mathbf{2}, \mathbf{3+}\}$ mogu se izračunati na temelju zbroja proširenja jednog niza prethodnih riječi $\boldsymbol{w_1, \ldots, w_{n-1}}$ sa zbrojem $\mathbf{1}$, odnosno $\mathbf{2}$, $\mathbf{3}$ ili $\boldsymbol{više}$. $\boldsymbol{D_1}$ se oduzima od svakog n-grama sa zbrojem $\mathbf{1}$, tako da je za $\boldsymbol{\gamma}$ raspoloživo $\boldsymbol{D_1}$ puta broj n-grama za zbrojem $\mathbf{1}$ itd. (Koehn, 2010). Izračun $\boldsymbol{\alpha}$ za n-gramske modele nižeg reda (eng. *lower order n-gram models*) dan je u (2.31). $\boldsymbol{D}$ se također određuje s obzirom na frekvenciju niza prethodnih riječi $\boldsymbol{w_1, \ldots, w_{n-1}}$.

$$\alpha(w_n|w_1, \ldots, w_{n-1}) = \frac{N_{1+}(\bullet\, w_1, \ldots, w_n) - D}{\sum_w N_{1+}(\bullet\, w_1, \ldots, w_{n-1}, w)} \tag{2.31}$$

Model povlačenja (eng. *back-off model*) oslanja se na n-gram najvišeg reda koji se uparuje s nizom prethodnih riječi i s predviđenim riječima (eng. *predicted words*). Međutim, ukoliko su podatci potrebni za izgradnju modela povlačenja nepotpuni ili oskudni (eng. *sparse*), tj. nije moguće sigurno utvrditi niz prethodnih riječi niti predvidjeti riječi koje slijede, tada takvi modeli povlačenja mogu biti vrlo nepouzdani (Koehn, 2010). Naime, ukoliko se dva različita n-grama s identičnim nizom prethodnih riječi pojave samo jednom u korpusu za treniranje, njihovim predviđenim riječima bit će dodijeljena jednaka vjerojatnost. No, takav jedan n-gram možda značajno odstupa od prosječnih vrijednosti (eng. *statistical outlier*), dok drugi n-gram može biti nereprezentativan uzorak nekog učestalog oblika niza riječi. Stoga se uvijek predlaže uporaba *back-off* modela nižeg reda, čak i ako su n-grami možda prethodno viđeni. To se može ostvariti prilagodbom $\boldsymbol{\alpha}$ funkcije u interpoliranu $\boldsymbol{\alpha_I}$ funkciju (2.32) dodavanjem *back-off* vjerojatnosti, pri čemu vrijednost $\boldsymbol{\gamma}$ također treba odgovarajuće umanjiti.

$$\begin{aligned} \alpha_I(w_n|w_1, \ldots, w_{n-1}) & \\ &= \alpha(w_n|w_1, \ldots, w_{n-1}) \\ &+ \gamma(w_1, \ldots, w_{n-1}) p_I(w_n|w_2, \ldots, w_{n-1}) \end{aligned} \tag{2.32}$$

Eksperimentalno je utvrđeno da interpolirana modificirana inačica Kneser-Ney metode izglađivanja efikasnije izglađuje vjerojatnosti u odnosu na ostale metode izglađivanja (Koehn, 2010) koje su ranije spomenute.

Još jedan vrlo čest problem u modelu strojnog prevođenja su nepoznate riječi (eng. *unknown words*) (Koehn, 2010). To su riječi izvan vokabulara (eng. *out-of-vocabulary words, OOV*) o kojima sustav za strojno prevođenje nema nikakve podatke i dekoder ne može generirati prijevod u postupku dekodiranja, s obzirom da ih prijevodni model ne opisuje (Manning i Schütze, 1999). Stoga se za svaku nepoznatu riječ u postupku normalizacije korpusa za treniranje modela predlaže uporaba posebnog $< \boldsymbol{UNK} >$ tokena. Tokeni ili pojavnice predstavljaju nizove znakova omeđene znakom razdvajanja (eng. *delimiter*).

Ideja je svaku nepoznatu riječ, tj. riječ izvan vokabulara, zamijeniti tim posebnim tokenom. Sam postupak treniranja modela odvija se jednako kao i za bilo koju drugu riječ. Pri dekodiranju, tj. generiranju strojnog prijevoda sve nepoznate riječi također treba zamijeniti posebnim tokenom $< \boldsymbol{UNK} >$, nakon čega se koristi vjerojatnost kao za bilo koju riječ koja nije uključena u podatkovni skup za modeliranje jezičnog modela (Jurafsky, 2015). Takav postupak negativno utječe na izglađivanje, međutim, to je vrlo često neizbježno.

Imena, brojevi, šum u obliku pravopisnih grešaka itd. generiraju velik broj novih tokena i time povećavaju vokabular u modelu. Posebno velik broj novih tokena generiraju brojevi, s obzirom da ih je beskonačno mnogo. Stoga se primjenjuje poseban simbol @ koji zamjenjuje sve znamenke (Koehn, 2010).

Budući da su jezični modeli relativno veliki (veličina se kreće od nekoliko stotina megabajta pa do nekoliko gigabajta ili terabajta), oni zahtijevaju veliku količinu računalnih resursa, a ponajviše radne memorije (RAM), prilikom dekodiranja nekog podatkovnog skupa. Kako prilikom dekodiranja novog teksta nije potrebno učitati jezični model u cijelosti, iz jezičnog modela se pripremaju samo oni segmenti koji su zaista i potrebni prilikom dekodiranja (Koehn, 2010). Naime, prijevodni model, o čemu će biti riječ u sljedećem poglavlju, generirat će samo malen broj riječi u ciljnom jeziku pomoću riječi izvornog jezika i tablice prijevoda fraza, tj. fraznih struktura. Zbog toga je potrebno u radnu memoriju učitati samo one n-grame iz jezičnog modela koji se pojavljuju u tekstu koji treba prevesti na ciljni jezik. Taj postupak naziva se i filtriranje jezičnog modela (eng. *language model filtering*) i standardno se koristi u sustavima za statističko strojno prevođenje (Koehn, 2010).

3.2. Prijevodni model

Prijevodni model $\boldsymbol{p(f|e)}$ uparuje riječi, odnosno nizove riječi, iz izvornog jezika u ciljni jezik, tj. procjenjuje podudarnost leksičkih jedinica u izvornom i ciljnom jeziku (Koehn, 2010). Stoga se razvija pomoću dvojezičnog sravnjenog paralelnog korpusa koji se sastoji od segmenata na izvornom jeziku te ljudski prevedenih prijevoda na ciljnom jeziku (Koehn, 2015). Prijevodni model opisuje vjerojatnost prijevoda na temelju paralelnog korpusa, s obzirom da bez korpusa nije moguće izravno procijeniti vjerojatnost prijevodnih parova (Latour, 2004). Načelno vrijedi, što je veća vjerojatnost $\boldsymbol{p(f|e)}$ to je izglednije da se radi o dobrom prijevodu.

Statističko modeliranje prijevodnog modela zasniva se na IBM-ovim modelima koji dozvoljavaju dijeljenje složenih procesa obrade paralelnog korpusa u manje korake, koji se zatim zasebno modeliraju te omogućuju izračun parametara potrebnih za izgradnju modela (España-Bonet i Gonzàlez, 2014; Koehn, 2010). IBM-ov model 1 najjednostavniji je model, a dodavanjem parametara raste i složenost modela, pri čemu je IBM model 5 najkompleksniji model (Och i Ney, 2000). IBM-ov model 1 opisuje leksičku vjerojatnost prevođenja riječi, IBM-ov model 2 dodaje model preslagivanja/premještanja redoslijeda riječi, dok IBM-ov model 3 dodaje model fertiliteta, a pretražuje najvjerojatniju sravnjenost riječi i susjednih riječi, tzv. Viterbijevu sravnjenost riječi (eng. *Viterbi alignment*) (Tiedemann, 2009) dobivenu IBM modelom 2 pomoću algoritma penjanja uzbrdo (eng. *hill climbing*) (Koehn, 2010). IBM-ov model 4 dodaje relativnu distorziju redoslijeda, tj. poretka prevedenih riječi (eng. *relative distortion*) s obzirom na prethodno prevedene riječi iz izvornog jezika s fertilitetom većim od $\mathbf{0}$ (eng. *cepts*) (Koehn, 2010).

Međutim, IBM-ovi modeli 1-4 imaju određene manjkavosti (eng. *deficiency*), s obzirom da netočne sravnjenosti mogu u modelu imati vjerojatnost veću od $\mathbf{0}$. Pored toga, više riječi u ciljnom jeziku može biti mapirano u istu poziciju, što umanjuje efikasnost distribucije vjerojatnosti (Koehn, 2010). IBM-ov model 5 nastoji riješiti nedostatke prethodnih modela dozvoljavajući popunjavanje samo prikladnih upražnjenih pozicija riječi u ciljnom jeziku (eng. *vacant positions*). U modelima 3, 4 i 5 generiranje prijevoda u ciljnom jeziku moguće je izvršiti uzevši u obzir i veći broj riječi u ciljnom jeziku (Clark, 2010). IBM modeli mogu se dodatno proširiti tzv. skrivenim Markovljevim modelom, tj. HMM-om (eng. *Hidden Markov Model*) (Jurafsky i Martin, 2013). U HMM-u vjerojatnost sravnjenosti jednog para riječi izravno ovisi o poziciji sravnjenosti prethodne riječi, s obzirom da je okolina riječi (eng. *local neighbourhood*) prilikom sravnjivanja često sačuvana (Vogel et al., 1996). IBM-ovi modeli i danas imaju široku upotrebu te

su često implementirani u brojne alate za sravnjivanje riječi, kao što su npr. fast_align ili GIZA++ (Och i Ney, 2003).

Za uspješnu izgradnju prijevodnog modela potrebno je za svaku riječ iz izvornog jezika modelirati odgovarajući prijevod, broj potrebnih riječi u ciljnom jeziku, poziciju unutar rečenice u prijevodu te broj riječi koje treba generirati iz posebnog tokena $\boldsymbol{NULL}$ (Koehn, 2010). Na taj način prijevodni model doprinosi izboru semantički ispravnih riječi u ciljnome jeziku. U nastavku je dan primjer uparivanja riječi za izvornu rečenicu $\boldsymbol{Uvijek\ ga\ rado\ vidim.}$ i prijevoda $\boldsymbol{I\ always\ look\ forward\ to\ seeing\ him.}$ (Slika 4).

Slika 4. Prikaz uparivanja riječi u izvornom i ciljnom jeziku.

Prema IBM-ovim modelima, prijevodni model izvodi četiri ključna procesa: prevođenje, upravljanje fertilitetom, permutiranje riječi te umetanje, odnosno ispuštanje riječi (Watanabe i Sumita, 2002). Shodno tome, IBM-ov prijevodni model može se definirati pomoću četiri parametra (Bonet i Gonzàlez, 2014; Koehn, 2010; Koehn, 2008):

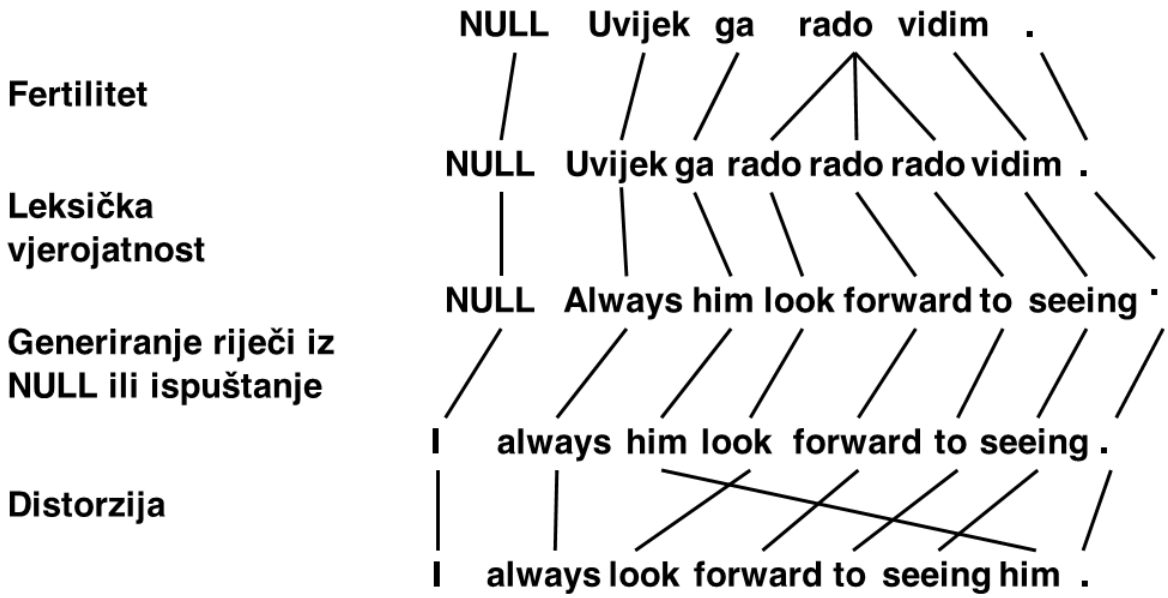
1. leksička vjerojatnost (eng. *lexical probability*), $\boldsymbol{t}$
2. fertilitet (eng. *fertility*), $\boldsymbol{n}$
3. distorzija (eng. *distortion/reordering*), $\boldsymbol{d}$
4. vjerojatnost umetanja ili ispuštanja riječi (eng. *insertion/dropping probability*), $\boldsymbol{p_{NULL}}$

Leksička vjerojatnost opisuje vjerojatnost da se primjerice riječ $\boldsymbol{vidim}$ prevede sa $\boldsymbol{seeing}$. S obzirom da se jedan koncept ne može uvijek doslovno prevesti jednom riječju, fertilitet $\boldsymbol{n(\phi|f)}$

opisuje vjerojatnost da se jedna riječ izvornog jezika prevede s više riječi u ciljnom jeziku, odnosno da $\boldsymbol{rado}$ generira riječi $\boldsymbol{look\ forward\ to}$. U tom slučaju je $\boldsymbol{n(3|rado)}$. Distorzija je vjerojatnost da riječ na poziciji $\boldsymbol{j}$ generira riječ na $\boldsymbol{i\text{-}toj}$ poziciji, tj. $\boldsymbol{d(j|i,m,n)}$, pri čemu $\boldsymbol{m}$ i $\boldsymbol{n}$ označavaju duljinu segmenta u izvornom, odnosno ciljnom jeziku (Bonet i Gonzàlez, 2014).

Distorzija se opisuje permutacijskim modelom, tj. modelom preslagivanja/premještanja redoslijeda riječi (eng. *reordering model*). Razlike u sintaksnim strukturama jezika koji se prevode zahtijevaju umetanje ili ispuštanje (funkcijskih) riječi prilikom prevođenja. $\boldsymbol{NULL}$ token se koristi kada riječ u ciljnom jeziku nema para u izvornom jeziku (Koehn, 2010).

Iz narednog primjera (Slika 5) proizlazi, da parametar vjerojatnosti umetanja riječi (eng. *insertion*) opisuje vjerojatnost da se umetnuta riječ $\boldsymbol{I}$ generira iz $\boldsymbol{NULL}$, tj. $\boldsymbol{p_{NULL}(I|NULL)}$. Jednako tako se riječi mogu ispuštati u prijevodu; to su riječi s fertilitetom $\boldsymbol{0}$ (Al-Onaizan et al., 1999).



Slika 5. Prikaz primjene četiri parametra u prijevodnom modelu.

Prevođenje zasebnih riječi vrlo je jednostavno, a primjer izračuna leksičke vjerojatnosti na temelju frekvencije pojavljivanja riječi $\boldsymbol{računalo}$ u jednom korpusu dan je u nastavku (Tablica 2).

Tablica 2. Frekvencije pojavljivanja riječi u korpusu.

Prijevod riječi „računalo“	**Broj pojavljivanja u korpusu**
computer	9000
personal computer	6800
PC	4400
desktop computer	3500
laptop	1700
notebook	600
ukupno	26000

Leksička vjerojatnost prevođenja riječi opisana je procjenom maksimalne vjerodostojnosti kao što je prikazano u nastavku (2.33).

$$p(e) = \begin{cases} 0.346 & ako\ e = computer \\ 0.262 & ako\ e = personal\ computer \\ 0.169 & ako\ e = PC \\ 0.135 & ako\ e = desktop\ computer \\ 0.066 & ako\ e = laptop \\ 0.022 & ako\ e = notebook \end{cases} \tag{2.33}$$

IBM model 1 isključivo koristi leksičku vjerojatnost te spada u kategoriju generativnih modela, s obzirom da postupak prevođenja rečenice dijeli u više manjih koraka (prevođenje riječi). Prema tom modelu vjerojatnost prevođenja izvorne rečenice $\boldsymbol{f} = (\boldsymbol{f_1}, \ldots, \boldsymbol{f_{l_f}})$ duljine $\boldsymbol{l_f}$ u ciljnu rečenicu $\boldsymbol{e} = (\boldsymbol{e_1}, \ldots, \boldsymbol{e_{l_e}})$ duljine $\boldsymbol{l_e}$ može se opisati jednadžbom (2.34), pri čemu se sravnjenost svake riječi u ciljnom jeziku $\boldsymbol{e_j}$ prema riječi $\boldsymbol{f_i}$ u izvornom jeziku može formalizirati funkcijom sravnjivanja $\boldsymbol{a}$ (eng. *alignment function*) (Koehn, 2010).

$$p(e, a|f) = p(e, a|f) = \frac{\epsilon}{(l_f + 1)^{l_e}} \prod_{j=1}^{l_e} t(e_j|f_{a(j)}) \tag{2.34}$$

U središtu IBM modela 1 je produkt leksičkih vjerojatnosti za sve $\boldsymbol{l_e}$ generirane riječi $\boldsymbol{e_j}$ u ciljnom jeziku (Koehn, 2010). S obzirom da je uključen i poseban ***NULL token***, riječi u izvornom jeziku zapravo je $\boldsymbol{l_f + 1}$. Stoga je i $\boldsymbol{(l_f + 1)^{l_e}}$ različitih sravnjenosti koje mapiraju $\boldsymbol{l_f + 1}$ riječi izvornog jezika s $\boldsymbol{l_e}$ riječi ciljnog jezika. Parametar $\boldsymbol{\epsilon}$ predstavlja normalizacijsku konstantu, koji vodi računa o tome da je distribucija vjerojatnosti zadovoljena (suma vjerojatnosti je jednaka $\boldsymbol{1}$) (Koehn, 2010).

Primjer prevođenja rečenice IBM modelom 1 prikazan je u nastavku (2.35):

danas

e	$t(e\|f)$
today	**0.9**
this	0.05
now	0.05

je

e	$t(e\|f)$
is	**0.8**
are	0.1
has	0.1

lijep

e	$t(e\|f)$
beautiful	**0.7**
pretty	0.15
nice	0.15

dan

e	$t(e\|f)$
day	**0.6**
daylight	0.3
daytime	0.1

$$p(e,a|f) = \frac{\epsilon}{5^4} \cdot t(today|danas) * t(is|je) * t(beautiful|lijep) * t(day|dan) = \frac{\epsilon}{5^4} * 0.9 * 0.8 * 0.7 * 0.6 = 0.000484\epsilon \quad (2.35)$$

Prema (2.36), funkcija $\boldsymbol{a}$ mapira riječ iz izvornog jezika na poziciji $\boldsymbol{j}$ u riječ ciljnoga jezika na poziciji $\boldsymbol{i}$ (Koehn, 2010).

$$a: j \rightarrow i \quad (2.36)$$

Sravnjenost se jednako tako može definirati i kao u (2.37), pri čemu se riječ u izvornom jeziku na poziciji $\boldsymbol{poz_f}$ prevodi s nula ili više riječi na poziciji, odnosno pozicijama, $\boldsymbol{poz_e}$ u ciljnom jeziku (Sima'an, 2013).

$$a: \{poz_f\} \rightarrow (\{poz_e \cup \{0\}) \tag{2.37}$$

$\boldsymbol{a_i}$ predstavlja poziciju riječi u $\boldsymbol{e}$ s kojom je riječ $\boldsymbol{f_i}$ sravnjena. Na primjeru rečenice $\boldsymbol{NULL\ Uvijek\ ga\ rado\ vidim}$ ($\boldsymbol{NULL}$ na nultoj poziciji, $\boldsymbol{vidim}$ na četvrtoj poziciji) i prijevoda $\boldsymbol{I\ always\ look\ forward\ to\ seeing\ him}$ ($\boldsymbol{I}$ na prvoj poziciji, a $\boldsymbol{him}$ na sedmoj poziciji), funkcija sravnjivanja $\boldsymbol{a}$ rezultira sljedećim mapiranjima (2.38):

$$a: \{1 \rightarrow 0, 2 \rightarrow 1, 3 \rightarrow 3, 4 \rightarrow 3, 5 \rightarrow 3, 6 \rightarrow 4, 7 \rightarrow 2\} \tag{2.38}$$

IBM-ov model 2 pretpostavlja da vjerojatnost sravnjenosti dviju riječi ovisi o pozicijama riječi koje povezuje i o duljini nizova (Clark, 2010). Stoga dozvoljava prevođenje riječi izvornog jezika na poziciji $\boldsymbol{i}$ u riječ ciljnog jezika na poziciji $\boldsymbol{j}$. Drugi IBM-ov model strojno prevođenje definira kao u (2.39), pri čemu $\boldsymbol{a(i|j, l_e, l_f)}$ predstavlja model preslagivanja/premještanja redoslijeda riječi (Koehn, 2010).

$$p(e, a|f) = \epsilon \prod_{j=1}^{l_e} t\left(e_j|f_{a(j)}\right) a\left(i|j, l_e, l_f\right) \tag{2.39}$$

Načelno se prijevodni model $\boldsymbol{p(f|e)}$ može formalizirati s obzirom na sravnjenost $\boldsymbol{a}$ kao u (2.40) (Sima'an, 2013).

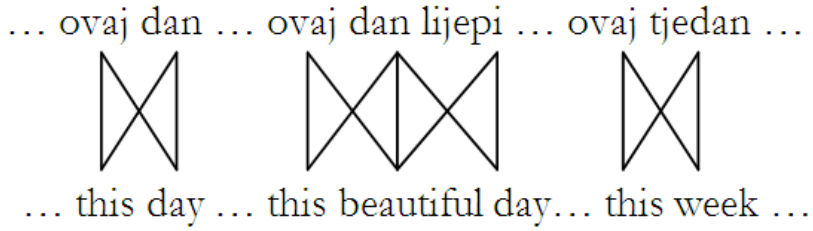
$$p(f|e) = \sum_a p(a, f|e) = \sum_a p(a|e) p(f|a, e) \tag{2.40}$$

Navedena četiri parametra (leksička vjerojatnost, fertilitet, distorzija, vjerojatnost umetanja/ispuštanja riječi) u statističkom modelu procjenjuju se na temelju frekvencija i pozicija riječi iz velikih skupova podataka, tj. rečenično-sravnjenih paralelnih korpusa koji, međutim, ne sadrže specifične podatke o sravnjenosti riječi te o „vezama" između izvornog i ciljnog jezika

(Koehn, 2010). Da bi se otkrile „veze“ između izvornog i ciljnog jezika u korpusu, potrebno je pomoću podataka iz paralelnog korpusa istrenirati prijevodni model. Takav paralelni korpus naziva se i podatkovni skup za treniranje (eng. *bilingual training set*). Treniranje prijevodnog modela je proces procjenjivanja sravnjenosti riječi (eng. *word alignment*) u jednom podatkovnom skupu za treniranje, zajedno s pripadajućim parametrima (Koehn, 2010). Treniranje se vrši pomoću algoritma maksimizacije očekivanja (eng. *expectation-maximisation algorithm, EM*) koji otkriva strukturu sravnjenosti riječi te veza unutar korpusa (eng. *hidden variable*).

Algoritam je opisan u nastavku (Jurafsky i Martin, 2013; Koehn, 2010; Koehn, 2008):

1. prvo se inicijaliziraju parametri (Slika 6)
    - npr. svi parametri imaju jednaku vrijednost (uniformna distribucija, $\boldsymbol{t = 0.25}$)
    - sve veze među riječima izvornog i ciljnog jezika imaju jednaku vjerojatnost, tj. jednako je vjerojatno da se svaka riječ izvornog jezika $\boldsymbol{f}$ prevede s bilo kojom riječi u ciljnom jeziku $\boldsymbol{e}$.

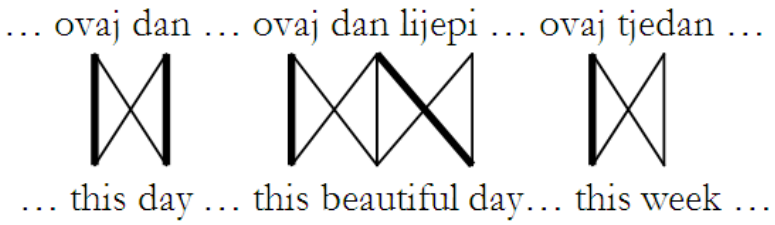


Slika 6. Sve veze među riječima imaju jednaku vjerojatnost.

2. potom se izračunavaju vjerojatnosti sravnjenosti (povezanosti) riječi u zadanom ulaznom skupu podataka za treniranje
    - $\boldsymbol{p(a|e,f)}$ je vjerojatnost jedne sravnjenosti s obzirom na dane rečenice u izvornom i ciljnom jeziku (2.41)

$$p(a|e,f) = \prod_{j=1}^{l_e} \frac{t(e_j|f_{a(j)})}{\sum_{i=0}^{l_f} t(e_j|f_i)} \tag{2.41}$$

- model uči da je npr. riječ ***ovaj*** često sravnjena s ***this***
- nakon prve iteracije, sravnjenosti riječi ***ovaj*** i ***this*** te ***dan*** i ***day*** sve su vjerojatnije (Slika 7)



Slika 7. Neke veze među riječima su vjerojatnije.

3. korak (2) se iterira, tj. ponovno se procjenjuju vrijednosti parametara modela na temelju podataka iz koraka (2)
    - nakon druge iteracije, sravnjenosti riječi ***lijepi*** i ***beautiful*** te ***tjedan*** i ***week*** sve su vjerojatnije

4. učenje iz modela: postupak se ponavlja koracima (2) i (3) za sve riječi u skupu za treniranje te se izračunavaju vjerojatnosti sravnjenosti riječi pomoću procjene maksimalne vjerodostojnosti
    - prebrojavanjem različitih sravnjenosti (2.42) u postupku treniranja može se uočiti da su određene veze među riječima vrlo česte

$$c(e|f;e,f) = \sum_{a} p(a|e,f) \sum_{j=1}^{l_e} \delta(e,e_j)\delta(f,f_{a(j)}) \qquad (2.42)$$

pri čemu Kronecker delta funkcija $\boldsymbol{\delta(x,y)}$ je $\mathbf{1}$ ako $\boldsymbol{x = y}$, inače $\mathbf{0}$.

- model na temelju prebrojavanja sravnjenosti određenih riječi uči da su određene kombinacije riječi vjerojatnije te procjenjuje vjerojatnost njihovih veza (2.43)

$$t(e|f) = \frac{\sum_{(e,f)} c(e|f;e,f)}{\sum_e \sum_{(e,f)} c(e|f;e,f)} \tag{2.43}$$

5. ukoliko vjerojatnosti sravnjenosti riječi konvergiraju prema određenoj vrijednosti, postupak se zaustavlja te se konačne vrijednosti uzimaju kao finalne vrijednosti parametara, a finalno stanje sravnjenosti riječi se pohranjuje
    - u suprotnom, postupak se vraća u fazu ponovne procjene vrijednosti parametara modela, korak (3), sve dok vrijednosti parametara ne konvergiraju (Slika 8)

Slika 8. Vjerojatnosti sravnjenosti riječi konvergiraju.

Rezultat algoritma maksimizacije očekivanja dan je na jednom primjeru leksičke vjerojatnosti $\boldsymbol{t}$ u nastavku (2.44):

$$\begin{gathered} t(ovaj|this) = 0.692 \\ t(ovaj|that) = 0.308 \\ t(dan|day) = 0.992 \\ t(lijepi|beautiful) = 0.614 \\ t(lijepi|pretty) = 0.209 \\ t(lijepi|nice) = 0.171 \\ t(tjedan|week) = 1 \\ \ldots \end{gathered} \tag{2.44}$$

Perpleksnost je mjera koja ukazuje na to koliko dobro jedan model opisuje podatke pomoću kojih je treniran. Perpleksnost garantirano opada (ili teoretski ostaje ista) sa svakom novom iteracijom algoritma maksimizacije očekivanja, što omogućuje postizanje konvergencije vjerojatnosti sravnjenosti riječi.

Perpleksnost se računa kao u (2.45), a primjer izračuna dan je u nastavku (za $\epsilon = 1$) (Tablica 3).

$$log_2 perpleksnost = -\sum_s log_2 p(e_s|f_s) \tag{2.45}$$

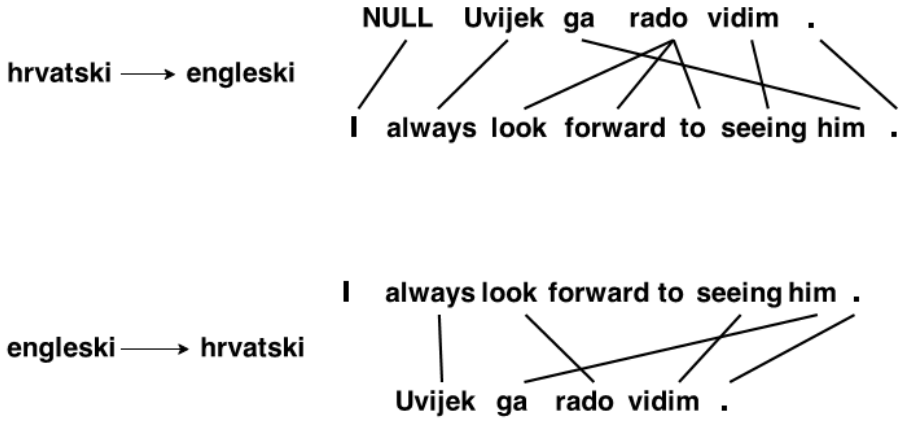
Tablica 3. Primjer izračuna perpleksnosti.

	inicijalno	**1. iteracija**	**2. iteracija**	**3. iteracija**	**…**	**konačno**
p(ovaj dan | this day)	0.0525	0.1795	0.1824	0.1831	…	0.1822
p(dan lijepi | beautiful day)	0.0525	0.1476	0.1771	0.1924	…	0.23
p(ovaj tjedan | this week)	0.0525	0.1873	0.1914	0.1919	…	0.1809
perpleksnost	6902	201	157	148	…	**132**

Podatci o sravnjenosti riječi vrlo su važni za model statističkog strojnog prevođenja (Koehn, 2010). Čovjeku je uočavanje uzoraka (eng. *pattern recognition*) i logičkih veza među riječima relativno jednostavan zadatak. Računalu, međutim, to može predstavljati problem.

Treba ponoviti da parametar „fertilitet" uzrokuje asimetričnost pri sravnjivanju riječi (Slika 5), s obzirom na IBM-ovu definiciju modela da se svaka riječ ciljnog jezika uvijek uparuje s najviše jednom riječi u izvornom jeziku (uključujući i ***NULL token***), tj. sravnjuje se pomoću funkcije sravnjivanja u vezu tipa **1-1** (jedan prema jedan) s riječi u izvornom jeziku (Koehn, 2010).

Primjer sravnjivanja riječi IBM modelom 1 prikazan je u nastavku (Slika 9). No, uobičajena je pojava u prirodnom jeziku da redoslijed riječi u izvornom jeziku ne odgovara redoslijedu riječi u ciljnom jeziku (Manning i Schütze, 1999). Naime, u prirodnom jeziku pojavljuje se potreba sravnjivanja riječi s vezom tipa **1-*n*** (jedan prema više pri čemu obrat nužno ne vrijedi) ili ***n*-*n*** (više prema više).



Slika 9. Asimetričnost uzrokovana fertilitetom.

U nastavku je vizualizirana asimetričnost uzrokovana parametrom fertiliteta (Tablice 4 i 5).

Tablica 4. Grafički prikaz sravnjenosti riječi, hrvatsko-engleski.

	NULL	Uvijek	ga	rado	vidim	.
I	■					
always		■				
look				■		
forward				■		
to				■		
seeing					■	
him			■			
.						■

Tablica 5. Grafički prikaz sravnjenosti riječi, englesko-hrvatski.

	NULL	Uvijek	ga	rado	vidim	.
I						
always		■				
look				■		
forward						
to						
seeing					■	
him			■			
.						■

Metoda koja heuristikom upotpunjuje sravnjivanje riječi prema IBM-ovim modelima (Koehn, 2010; Koehn et al., 2003) rješava problem asimetričnosti riječi uzrokovanih fertilitetom. Tim pristupom rješava se problematika sravnjivanja tipa $\boldsymbol{n}$-$\boldsymbol{n}$. Glavna ideja ove metode jest korpus sravniti u oba smjera te sravnjenosti riječi zatim spojiti pomoću presjeka ili unije. Postupak spajanja sravnjenosti riječi iz oba smjera naziva se simetrizacija sravnjenosti riječi (eng. *symmetrisation of word alignments*) (Jurafsky i Martin, 2013; Koehn, 2010).

Nakon što se korpusi sravne u oba smjera, traže se presjeci sravnjenih riječi, što odgovara visokoj preciznosti (eng. *precision*) i pouzdanosti, međutim, niskom odzivu (eng. *recall*) (España-Bonet i Gonzàlez, 2014). Zatim se pomoću operacije „unija" traže dodatne zajedničke točke preklapanja sravnjenih riječi (eng. *grow additional alignment points*) (Koehn et al., 2005), što povećava odziv pri čemu se ne utječe na preciznost, što je prikazano i u nastavku (Tablice 6 i 7).

Postoje brojne heuristike koje se razlikuju s obzirom na metode traženja dodatnih zajedničkih točaka (tj. riječi) preklapanja (Koehn, 2015). Određene heuristike uzet će u obzir jezični smjer sravnjenosti ili, primjerice, samo one riječi u okolini koje nisu sravnjene. Heuristike se mogu oslanjati i na to da li potencijalna dodatna točka preklapanja već graniči s jednom drugom točkom preklapanja ili ne. Ako graniči, da li izravno graniči sa susjednom točkom preklapanja sravnjenosti (nalazi li se pored ili graniči dijagonalno) ili neizravno? Može se bazirati i na frekvencijama te leksičkoj vjerojatnosti potencijalne točke preklapanja.

Pseudo-kod popularne simetrizacijske heuristike *grow-diag-final-and* dan je u nastavku (Haddow, 2009):

GROW-DIAG-FINAL-AND(e2f,f2e)

susjedno = {(-1,0),(0,-1),(1,0),(0,1),(-1,-1),(-1,1),(1,-1),(1,1)}

sravnjenost $\mathcal{A}$ = presjek(e2f,f2e);

GROW-DIAG();

FINAL-AND(e2f);

FINAL-AND(f2e);

GROW-DIAG()

while nove točke dodane **do**

for all riječ ciljnog jezika $e \in [1 \dots e_n]$, riječ izvornog jezika $f \in [1 \dots f_n]$, $(e, f) \in \mathcal{A}$ **do**

for all susjednu točku sravnjenosti (e_{nova}, f_{nova}) **do**

if (e_{nova} nesravnjena **OR** f_{nova} nesravnjena) **AND** $(e_{nova}, f_{nova}) \in$ unija (e2f,f2e) **then** dodaj točku (e_{nova}, f_{nova}) u $\mathcal{A}$

end if

end for

end for

end while

FINAL-AND()

for all nova riječ ciljnog jezika $e_{nova} \in [1 \dots e_n]$, nova riječ izvornog jezika $f_{nova} \in [1 \dots f_n]$ **do**

if (e_{nova} nesravnjena **AND** f_{nova} nesravnjena) **AND** $(e_{nova}, f_{nova}) \in$ sravnjenost $\mathcal{A}$ **then** dodaj točku (e_{nova}, f_{nova}) u $\mathcal{A}$

end if

end for

Nema jasnih odrednica kako odabrati optimalnu heuristiku (Wu et al., 2007). Naime, izbor heuristike ovisi i o veličini korpusa te jezičnom paru. Matrica simetrizacije sravnjenosti riječi (eng. *alignment symmetrisation matrix*) vizualizira sravnjivanje riječi te primjenjuje dvije glavne strategije simetrizacije riječi: presjek te uniju (Koehn, 2010). Primjer presjeka sravnjenih riječi za oba jezična smjera (eng. *intersection of bidirectional alignments*) prikazan je u Tablici 6.

Tablica 6. Simetrizacija sravnjenih riječi primjenom presjeka, za oba smjera: (hrvatsko-engleski) ∩ (englesko-hrvatski).

	NULL	Uvijek	ga	rado	vidim	.
I						
always		■				
look				■		
forward						
to						
seeing					■	
him			■			
.						■

Pogodnost unije je visok odziv, međutim, takav postupak ima manju pouzdanost (España-Bonet i Gonzàlez, 2014). Primjer unije sravnjenih riječi iz oba jezična smjera (eng. *union of bidirectional alignments*) prikazan je u Tablici 7.

Tablica 7. Simetrizacija sravnjenih riječi primjenom unije, za oba smjera: (hrvatsko-engleski) ∪ (englesko-hrvatski).

	NULL	Uvijek	ga	rado	vidim	.
I	■					
always		■				
look				■		
forward				■		
to				■		
seeing					■	
him			■			
.						■

U gornjem primjeru niz riječi ***look forward to*** sravnjen je s ***rado***, međutim, izvan tog konteksta riječ ***rado*** zasigurno ima i bolje prijevode. S obzirom da se u korpusima vrlo često pojavljuju nizovi riječi s idiomatskim značenjem i kolokacije (Sinhal i Chandak, 2011; Manning i Schütze, 1999), očigledno je da puko sravnjivanje riječi nije dovoljno.

Jedno od mogućih rješenja jest izgradnja prijevodnog modela koji se ne temelji na riječima, već frazama (Koehn, 2010). To nisu fraze u lingvističkom smislu riječi, gdje se definiraju kao sintaktički motivirane skupine riječi (npr. ***NP*** (imenska skupina), ***VP*** (glagolska skupina), ***PP*** (prijedložna skupina) itd.). Naime, ispostavilo se da ograničavanje fraza na fraze u lingvističkom smislu umanjuje kvalitetu statističkog strojnog prijevoda (Knight i Koehn, 2003; Koehn et al., 2003).

Zapravo, fraza je skup proizvoljnih riječi koje se ekstrahiraju iz sravnjenosti riječi izvornog i ciljnog jezika, pri čemu se uzima u obzir vjerojatnost pojavljivanja niza riječi koje sastavljaju, tj. čine frazu (Koehn, 2010). Naime, određeni nizovi riječi se uobičajeno prevode zajedno, kao u ovom primjeru niz riječi ***looking forward to***. Stoga se takvi skupovi riječi, tj. fraze u statističkom strojnom prevođenju temeljenom na frazama uzimaju kao atomarne prijevodne jedinice (eng. *phrases as atomic translation units*). Takav model je kontekstno osjetljiv te omogućuje prijevode tipa ***n-n***, odnosno prijevode s više riječi u više riječi. Stoga je tim pristupom lakše upravljati složenijim idiomatskim konstrukcijama te dvoznačnošću fraza. Treba naglasiti da načelno vrijedi, što su skupovi za treniranje prijevodnog modela veći to su i prijevodi kvalitetniji,

s obzirom da se iz većih korpusa mogu naučiti dulje fraze, ponekad čak i cijele rečenice (Koehn, 2010).

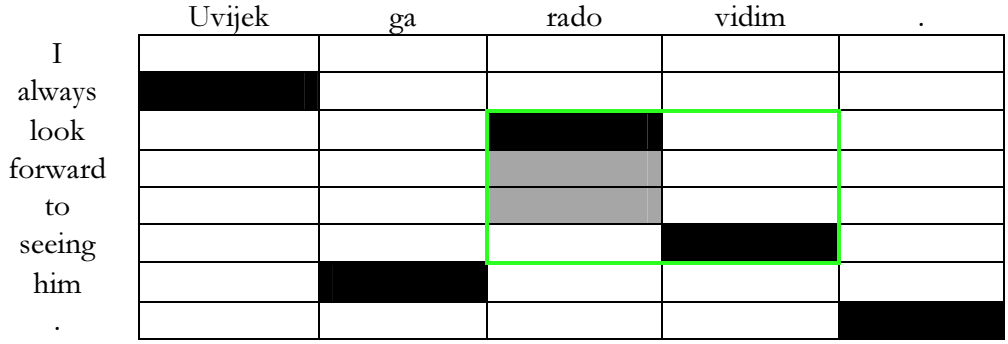
Prijevodni model $\boldsymbol{p(f|e)}$ koji se temelji na frazama (eng. *phrase-based model*) moguće je ostvariti modifikacijom standardnog IBM-ovog statističkog modela koji se temelji na riječima (Koehn, 2010). Pristup statističkom strojnom prevođenju koji koristi takav tip proširenog prijevodnog modela naziva se statističko strojno prevođenje temeljeno na frazama (eng. *phrase-based statistical machine translation*) (Koehn, 2010). Ideja je rečenice u izvornom jeziku podijeliti, tj. segmentirati u fraze te zatim svaku frazu zasebno prevesti u ciljni jezik, a eventualne permutacije fraza izvršiti naknadno.

Međutim, da bi takav pristup bio moguć, i samo sravnjivanje treba biti na razini fraza, a ne riječi. Pa se postavlja pitanje, kako na temelju sravnjenih riječi ekstrahirati fraze (eng. *word alignment induced phrases*) te dohvatiti vjerojatnosti sravnjenosti fraza?

Odgovor na to pitanje daje metoda koja nalaže da sve točke preklapanja sravnjenih riječi koje čine frazni par (eng. *phrase pair*) trebaju, unutar matrice simetrizacije sravnjenosti riječi, biti u okviru sravnjenosti fraze (eng. *phrase alignment box*), pri čemu nije važno nalaze li se u okviru sravnjenosti fraze i neke nesravnjene riječi, pa makar se one pojavile i u rubnom dijelu okvira (Koehn, 2010), što je prikazano Tablicama 8 i 9 (adaptirano prema Dyer, 2011).

Drugim riječima, za uspješnu ekstrakciju fraza potrebno je prikupiti sve frazne parove koji su konzistentni sa sravnjenošću svojih sastavnih riječi (eng. *consistent with a word alignment*), tj. svaka sravnjena fraza treba sadržavati sve točke sravnjenosti, tj. riječi koje je sačinjavaju (Koehn, 2010).

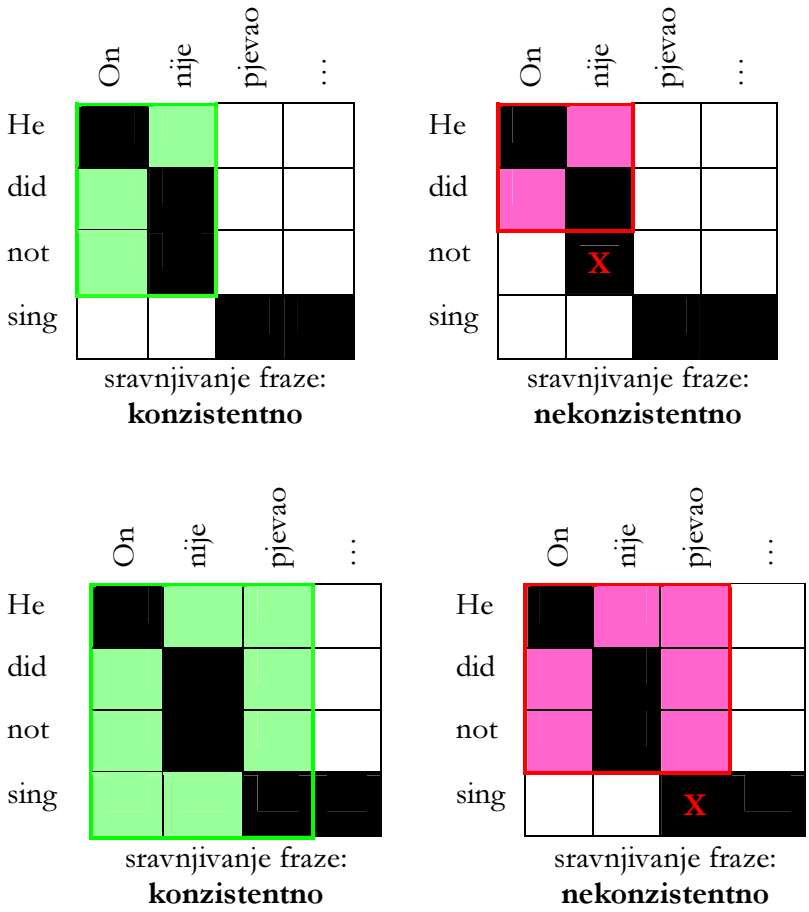
Tablica 8. Primjer ekstrakcije jedne fraze iz sravnjenosti riječi.



(**look forward to seeing**, **rado vidim**)

Primjeri ekstrahiranih fraza su $\boldsymbol{uvijek - always}$, $\boldsymbol{rado - look\ forward\ to}$, $\boldsymbol{vidim - seeing}$, $\boldsymbol{rado\ vidim - look\ forward\ to\ seeing}$ itd.

Tablica 9. Prikaz konzistentnog i nekonzistentnog sravnjivanja fraze.



Frazni par $\langle \bar{f} | \bar{e} \rangle$ je konzistentan sa sravnjenošću $\boldsymbol{\mathcal{A}}$ akko (Sima'an, 2013):

- se najmanje jedan par sravnjenih riječi, tj. jedna točka sravnjenosti nalazi u $\langle \bar{f} | \bar{e} \rangle$,
- niti jedna pozicija u izvornom jeziku unutar $\langle \bar{f} | \bar{e} \rangle$ nije sravnjena s pozicijama izvan $\langle \bar{f} | \bar{e} \rangle$,
- niti jedna pozicija u ciljnom jeziku unutar $\langle \bar{f} | \bar{e} \rangle$ nije sravnjena s pozicijama izvan $\langle \bar{f} | \bar{e} \rangle$.

Nekonzistentno sravnjivanje fraza je sravnjivanje pri kojem fraza ne sadrži sve riječi koje je sačinjavaju. Stanje konzistentnosti se može formalizirati kao u (2.46) (Koehn, 2010).

$$(\bar{e}, \bar{f})\ konzistentan\ s\ \mathcal{A} \Leftrightarrow \forall e_i \in \bar{e} : (e_i, f_j) \in \mathcal{A} \rightarrow f_j \in \bar{f}\ AND\ \forall f_j \in \bar{f} : (e_i, f_j) \in \mathcal{A} \rightarrow e_i \in \bar{e}\ AND\ \exists e_i \in \bar{e}, f_j \in \bar{f} : (e_i, f_j) \in \mathcal{A} \quad (2.46)$$

Frazni par $(\bar{e}, \bar{f})$ konzistentan je sa sravnjenošću $\mathcal{A}$ ako sve riječi $f_1, \ldots, f_n$ u $\bar{f}$ koje imaju točke sravnjenosti u $\mathcal{A}$, te iste točke sravnjenosti imaju s riječima $e_1, \ldots, e_n$ u $\bar{e}$ i obrnuto. Ideja je pronaći sve točke sravnjenosti za frazu u ciljnom jeziku te najkraću frazu u izvornom jeziku koja sadrži sve potrebne riječi za izgradnju fraze u ciljnom jeziku (Koehn, 2010). Točke sravnjenosti time ograničavaju broj mogućih ekstrahiranih fraza. Primjer ekstrakcije konzistentnih fraza na temelju sravnjenih riječi dan je u nastavku (Slika 11) (Koehn, 2004).

Slika 10. Primjer ekstrakcije konzistentnih fraza na temelju sravnjenih riječi.

Popis ekstrahiranih konzistentnih fraza prema gornjem primjeru dan je u nastavku:

(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green), (Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the), (bruja verde, green witch), (Maria no daba una bofetada, Mary did not slap), (no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch), (Maria no daba una bofetada a la, Mary did not slap the), (daba una bofetada a la bruja verde, slap the green witch), (no daba una bofetada a la bruja verde, did not slap the green witch), (Maria no daba una bofetada a la bruja verde, Mary did not slap the greenwitch)

Broj ekstrahiranih fraza ovisit će o simetrizacijskoj metodi (presjek i/ili unija) i odabranoj heuristici. Međutim, kako obje metode imaju svoje prednosti i nedostatke, predlaže se kombinacija metoda radi uspješnije ekstrakcije fraza, što izravno utječe na kvalitetu strojnog prijevoda (Koehn, 2015; España-Bonet i Gonzàlez, 2014).

Za svaki rečenični par ekstrahiraju se konzistentni frazni parovi (Jurafsky i Martin, 2013). Zatim se prebrojava u koliko rečeničnih parova se određeni frazni par ekstrahira, a broj pojavljivanja se pohranjuje u $\boldsymbol{zbroj\_pojavljivanja}(\bar{\boldsymbol{e}}, \bar{\boldsymbol{f}})$. Za svaki prikupljeni frazni par procjenjuje se distribucija vjerojatnosti prevođenja fraze (eng. *phrase translation probability distribution*) $\boldsymbol{\phi}(\bar{\boldsymbol{f}}|\, \bar{\boldsymbol{e}})$ na temelju relativne frekvencije fraznog para (eng. *relative frequency*) u paralelnom korpusu, što je prikazano jednadžbom (2.47), pri čemu $\bar{\boldsymbol{e}}$ i $\bar{\boldsymbol{f}}$ predstavljaju fraze na ciljnom, odnosno izvornom jeziku (Koehn, 2010). Postupak se naziva i bodovanje fraznih prijevoda (eng. *scoring phrase translations*).

$$\phi(\bar{f}|\, \bar{e}) = \frac{zbroj\_pojavljivanja(\bar{e}, \bar{f})}{\sum_{\bar{f}_i} zbroj\_pojavljivanja\ (\bar{e}, \bar{f}_i)} \tag{2.47}$$

Svi ekstrahirani konzistentni frazni parovi zajedno s pripadajućim vjerojatnostima pohranjeni su u tablici prijevoda fraza, tj. fraznih struktura (eng. *phrase translation table*). Tj. ona nastaje sravnjivanjem riječi pomoću IBM-ovih modela, ekstrakcijom fraznih parova te izračunom vjerojatnosti svakog fraznog para (eng. *phrase pair score*) (Koehn, 2010). Skup mogućih fraznih parova čini skup mogućnih prijevoda – opcije prijevoda (eng. *translation options*). Takva tablica prijevoda fraza, tj. fraznih struktura sadržavat će sve frazne parove koji su konzistentni sa sravnjenošću riječi koje sačinjavaju frazu. Primjer tablice prijevoda fraza, tj. fraznih struktura dan

je u nastavku (Tablica 10). Takva tablica može se koristiti i za ekstrakciju terminologije te izradu terminoloških baza i leksikona (Thurmair i Aleksić, 2012).

Tablica 10. Primjer tablice prijevoda fraza, tj. fraznih struktura na primjeru fraze „**od ove godine**“.

Engleski	$\boldsymbol{p(e\|f)}$
from this year	0.7122
this year on	0.1052
this year's	0.0963
of this year	0.0427
this year	0.0396
this year on ,	0.0273
annual	0.0219
those	0.0031
…	…

Gornja tablica fraznih struktura sadrži leksičke varijacije riječi (***year*** i ***annual***), morfološke varijacije (***this year's*** i ***this year***) te uključuje funkcijske riječi sa sporednom ulogom u rečenici (***from***, ***this*** itd.) i različite oblike šuma (***those***).

3.3. Dekoder

Statistički modeli svakom mogućem prijevodu pridružuju određene vrijednosti, tj. vjerojatnosti (eng. *score*) s obzirom na dani niz riječi u izvornom jeziku (Koehn, 2010). Vjerojatnosti prevođenja fraze i preslagivanja, tj. premještanja redoslijeda riječi te vjerojatnost jezičnog modela se objedinjuju kako bi se svakom mogućem prijevodu pridružio jedan združeni rezultat, tj. konačna vjerojatnost (eng. *final score*). Dekoder zatim treba pronaći najbolji prijevod iz skupa mogućih prijevoda (España-Bonet i Gonzàlez, 2014).

Može se reći da je zadaća dekodera u modelu statističkog strojnog prevođenja za izvornu rečenicu $\boldsymbol{f}$ pronaći prijevod $\boldsymbol{e}$ u ciljnom jeziku s maksimalnom vjerojatnošću (2.48) (España-Bonet i Gonzàlez, 2014).

$$e = argmax_e p(e|f) \tag{2.48}$$

Tj., dekoder u prostoru mogućih prijevoda treba pronaći segment, tj. rečenicu koja maksimizira ($\boldsymbol{argmax}$) vjerojatnost prijevodnog modela $\boldsymbol{p(f|e)}$, kao što je prikazano u (2.49) (Koehn, 2010).

$$e = argmax_e p(e|f) = argmax_e p(e)p(f|e) \tag{2.49}$$

Drugim riječima, uz dani jezični model $\boldsymbol{p(e)}$ i prijevodni model $\boldsymbol{p(f|e)}$, dekoder za rečenicu u ciljnom jeziku $\boldsymbol{e}$, u prostoru mogućih prijevoda, traži onaj prijevod koji maksimizira $\boldsymbol{p(e)p(f|e)}$.

U postupku dekodiranja prijevod se izgrađuje postupno, riječ po riječ, od početka do kraja rečenice. No, prije nego li se prevede rečenica s izvornog jezika na ciljni, dekoder pretražuje tablicu prijevoda fraza, tj. fraznih struktura i pronalazi prikladan skup njenih mogućih prijevoda. Za vrijeme dekodiranja djelomični prijevodi (eng. *partial translations*), koji se postupno izgrađuju, pohranjuju se u strukturi podataka koja se naziva hipoteza (eng. *hypothesis*) (Koehn, 2010). Dekodiranjem se hipoteze nadograđuju (proširuju) ovisno o sljedećem prijevodu fraze. S obzirom

da se nadogradnjom hipoteza povećava računalna kompleksnost dekodiranja, potrebno je minimizirati broj mogućih prijevoda.

Postoji nekoliko algoritama i metoda za dekodiranje, primjerice A* algoritam (eng. *A* search*) (Manning i Schütze, 1999) koji predstavlja standardnu tehniku pretraživanja u umjetnoj inteligenciji, ili pak algoritam penjanja uzbrdo (eng. *greedy hill-climbing*) koji prvo generira grubi prijevod, a eventualne korekcije vrši naknadno (Koehn, 2010). Za dekodiranje uspješno su primijenjeni i konačni pretvarači (eng. *finite state transducers*) (Liu i Gildea, 2008; Lopez, 2008).

Također, jedan od mogućih algoritama za dekodiranje, tj. za pretraživanje i pronalaženje prijevoda jest algoritam za zrakasto pretraživanje (eng. *beam search*) (Manning i Schütze, 1999), koji s obzirom na trošak prevođenja (eng. *cost*) odabire „najjeftiniji" manji skup hipoteza (djelomičnih prijevoda), tj. skup s najboljim karakteristikama (eng. *beam*). Neformalni pojam „trošak" analogan je pojmu „vjerojatnost": visok trošak predstavlja nisku vjerojatnost.

Trenutni trošak nove (proširene) hipoteze odnosi se na trošak originalne (prethodne) hipoteze multipliciran s troškovima jezičnog modela, prijevodnog modela te modela preslagivanja, tj. premještanja riječi za dodani (novi), tj. prošireni prijevod fraze. Algoritam zrakastog pretraživanja uvijek kreće s lijeve strane i postupno izgrađuje prijevod ekspandiranjem hipoteze, pritom napredujući udesno (Koehn, 2010; Koehn i Callison-Burch, 2008). Algoritam zrakastog pretraživanja opisan je u nastavku (Koehn, 2010; Koehn, 2008):

1. iz tablice prijevoda fraza, tj. fraznih struktura za sve riječi i nizove riječi prikupiti moguće prijevode s obzirom na leksičku vjerojatnost, odnosno distribuciju vjerojatnosti prevođenja fraze

2. inicijalizirati stanje
    - prazan djelomičan prijevod, tj. hipoteza bez prijevoda (eng. *empty hypothesis*)
    - niti jedna riječ izvornog jezika nije pokrivena, tj. prevedena
    - hipoteza ne sadrži niti jednu riječ u ciljnom jeziku

    2.1. odabrati riječ ili niz riječi u izvornom jeziku koju treba prevesti u ciljni jezik (s lijeve strane prema desno)

    2.2. pomoću tablice prijevoda fraza, tj. fraznih struktura pronaći i sačuvati prijevod za riječ ili niz riječi u ciljnom jeziku (veze $\mathbf{1\text{-}1}$, $\mathbf{1}\text{-}\boldsymbol{n}$, $\boldsymbol{n}\text{-}\mathbf{1}$, $\boldsymbol{n}\text{-}\boldsymbol{n}$)

    2.3. izvršiti preslagivanje/premještanje redoslijeda riječi

    2.4. prevedenu riječ ili niz riječi označiti kao prevedenu

2.5. evidentirati sve moguće prijevode riječi i nizova riječi (eng. *translation options*)

- mogući prijevodi sadržavaju sljedeće informacije: prva riječ u izvornom jeziku koja je pokrivena, zadnja riječ u izvornom jeziku koja je pokrivena, prijevod fraze na ciljni jezik, vjerojatnost prevođenja fraze

3. djelomičan prijevod, tj. hipotezu proširiti s mogućim prijevodima (eng. *hypothesis expansion with translation options*)
    - hipoteza se izgrađuje tako da se proširuje prijevodima riječi ili nizova riječi iz izvornog jezika, s obzirom na leksičku vjerojatnost ili distribuciju vjerojatnosti prevođenja fraze
    - za svaki mogući prijevod izgraditi novu hipotezu

    3.1. odabrati prvi mogući prijevod te kreirati prvu hipotezu
    - prva riječ ili niz riječi izvornog jezika pokrivena/pokriven
    - hipoteza sadrži prvu prevedenu riječ ili niz riječi na ciljnom jeziku
    - leksička vjerojatnost riječi evidentirana, distribucija vjerojatnosti prevođenja fraza evidentirana, trenutni trošak evidentiran

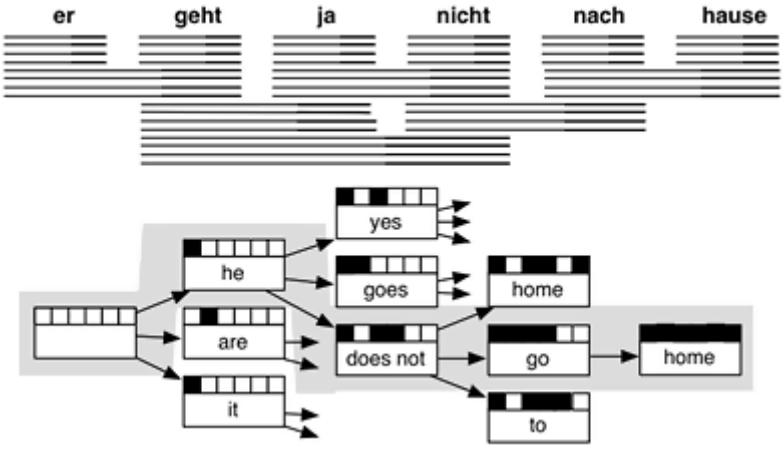
    3.2. dodaj još jednu hipotezu te ponovi korak (3.1)
    - pohranjena veza s najboljom prethodnom proširenom frazom
    - evidentirane riječi u izvornom jeziku koje su već pokrivene
    - broj hipoteza naglo raste, tj. eksponencijalno s duljinom rečenice koja se prevodi (Koehn et al., 2003)

4. procijeniti trošak (eng. *cost*) svakog mogućeg prijevoda, tj. trošak proširenja hipoteze
    - ako su sve riječi iz izvornog jezika pokrivene hipotezom, prijeći u korak (5)
    - ako hipotezom nisu pokrivene sve riječi iz izvornog jezika, vratiti u korak (3)

5. vratiti „najjeftiniji“ prijevod, tj. hipotezu s najmanjim troškom metodom unatražnog praćenja (eng. *backtracking*), pomoću koje se može rekonstruirati put (eng. *path*) do najpovoljnije hipoteze među mogućim prijevodima

Vizualan prikaz tijeka dekodiranja prikazan je u nastavku (Slika 11) (Koehn, 2010). Naime, dekoder *a priori* ne zna ispravan sastav prijevoda, već se u procesu odabiranja i preslagivanja/premještanja redoslijeda riječi i fraza hipoteza postupno izgrađuje. Prvo se

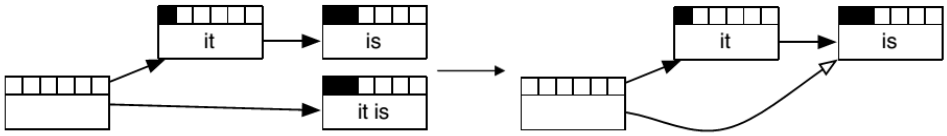
inicijalizira stanje hipoteze, nakon čega se najizglednija hipoteza proširuje dodavanjem mogućih prijevoda (označeno strelicama), a rezultirajući poredak riječi i nizova riječi pohranjuje se kao nova hipoteza, tj. djelomični prijevod u grafu pretraživanja (eng. *search graph*). Zacrnjeni kvadratići predstavljaju riječi koje su pokrivene u izvornom jeziku, a analizom strelica mogu se identificirati hipoteze-roditelji. Metodom unatražnog praćenja može se pronaći put do cjelovitog prijevoda koji se sastoji od više djelomičnih prijevoda, tj. hipoteza (Koehn, 2010). Kvaliteta strojnog prijevoda i vrijeme pretraživanja cjelovitog prijevoda ovisit će i o broju djelomičnih prijevoda (Koehn i Hoang, 2013).



Slika 11. Vizualizacija procesa dekodiranja metodom *beam search* i pronalazak prijevoda metodom unatražnog praćenja.

Ovaj algoritam strogo napreduje s lijeve prema desnoj strani i postupno izgrađuje prijevod. S obzirom da se radi o vremenski vrlo zahtjevnom i skupom algoritmu koji za svaku hipotezu izračunava trošak nepokrivenih riječi, predlaže se smanjivanje prostora pretraživanja mogućih prijevoda (eng. *search space reduction*) (Koehn, 2010).

Smanjivanje prostora pretraživanja može se izvršiti tzv. metodom rekombinacije hipoteza (eng. *hypothesis recombination method*) koja je bez rizika (eng. *risk free*) (Koehn, 2010). Radi se o tehnici dinamičkog programiranja koja odbacuje hipoteze koje statistički ne mogu biti dio najboljeg prijevoda. Naime, ukoliko dvije hipoteze rezultiraju dvjema jednakim hipotezama, no s različitim vjerojatnostima, pri čemu je jednak broj izvornih riječi preveden te je jednak broj ciljnih riječi generiran, ona manje vjerojatna, tj. skuplja hipoteza se odbacuje (Slika 12).

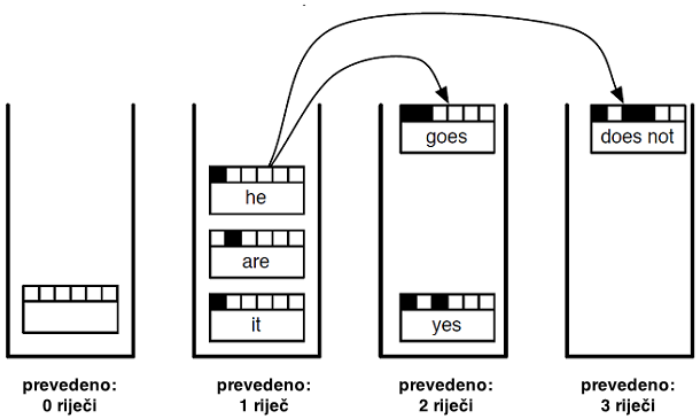


Slika 12. Rekombinacija fraza.

No, kako se rekombinacijom fraza efikasno ne smanjuje prostor pretraživanja mogućih prijevoda, tj. hipoteza, predlaže se primjena zrakastog pretraživanja s odbacivanjem (eng. *beam search with pruning*) (Koehn, 2010). Ideja je što ranije odbaciti „loše" hipoteze, slaganjem usporedivih hipoteza u stogove (eng. *hypothesis stack*) koji su ograničenog kapaciteta, tj. koji limitiraju broj mogućih hipoteza u svakom stogu, te ranim uklanjanjem iz stoga. Usporedive hipoteze su one hipoteze koje su prevele jednak broj riječi iz izvornog jezika (Koehn, 2010).

Stogovi se u pravilu čiste ograničavanjem kapaciteta hipoteza (eng. *histogram pruning*), međutim, ova metoda može se proširiti na način da se odbacuju samo one hipoteze koje su u odnosu na najbolju hipotezu u stogu lošije za određeni faktor (eng. *threshold pruning*). Metoda zrakastog pretraživanja s odbacivanjem međusobno uspoređuje hipoteze s jednakim brojem prevedenih riječi iz izvornoga jezika i odbacuje sve one manje kvalitete.

Slika 13 prikazuje primjer stogova koji sadrže usporedive (eng. *comparable*) hipoteze (Koehn i Hoang, 2013). Prvi stog sadrži samo jednu hipotezu. To je prazna hipoteza inicijalizirana s početka algoritma za zrakasto pretraživanje.



Slika 13. Primjer stogova s hipotezama.

Kvaliteta hipoteze opisana je ostvarenim troškom prevođenja te procjenom budućeg troška prevođenja ostatka rečenice (eng. *future cost estimation*) (Jurafsky i Martin, 2013). Budući trošak odnosi se na dijelove rečenice koji još nisu prevedeni te se računa za svaki mogući prijevod (eng. *translation option*), pritom uzevši u obzir trošak prijevodnog modela te procjenjivši trošak jezičnog modela (Koehn, 2010). Jezični model može se samo procijeniti, jer unatoč tome što su riječi iz ciljnog jezika možda poznate jezičnom modelu, kontekst neprevedenog dijela rečenice je nepoznat jezičnom modelu. Ukoliko se proširenjem hipoteze generira samo jedna riječ u ciljnom jeziku, trošak jezičnog modela aproksimira se računajući vjerojatnost jezičnog modela samo za tu generiranu riječ u ciljnom jeziku (eng. *unigram probability*). Ukoliko se generiraju dvije riječi, računa se unigramska vjerojatnost prve riječi te bigramska vjerojatnost druge generirane riječi itd. Također, prilikom procjene budućeg troška ne uzima se u obzir model preslagivanja/premještanja redoslijeda riječi (Koehn, 2010). S obzirom da se radi samo o procjeni, tj. o aproksimaciji, ova metoda nije bez rizika (España-Bonet i Gonzàlez, 2014).

No, u postupku prevođenja nove rečenice, dekodiranje se može ograničiti na monotono dekodiranje (eng. *monotone decoding*) u kojem se riječi nove rečenice prevode isključivo sekvencijalno (bez premještanja), pri čemu je dozvoljeno preslagivanje, tj. premještanje redoslijeda riječi unutar same fraze, međutim, ne dozvoljavaju se promjene u poretku fraza (Way i Hassan, 2009). Takav pristup može se kombinirati s dekodiranjem koje dozvoljava premještanje riječi samo do određene udaljenosti, što također umanjuje prostor pretraživanja mogućih prijevoda (Koehn, 2015).

Primjer procjene budućeg troška prikazanog u obliku log vjerojatnosti dan je u nastavku (adaptirano prema Koehn, 2010), a s obzirom da je logaritam broja iz skupa $[\mathbf{0}, \mathbf{1}]$ negativan, log vjerojatnosti u primjeru također su negativne (Tablica 11).

Tablica 11. Primjer procjene budućeg troška rečenice za **n** riječi (s obzirom na prvu riječ).

prva riječ	**1**	**2**	**3**	**4**	**5**	**6**
you	-1.6	-3.4	**-5.2**	-6.5	-7.5	-8.4
should	-1.8	-3.6	-4.9	-5.9	-6.6	
save	-1.8	-3.1	-4.1	**-5.0**		
yourself	-1.3	-2.3	-3.2			
the	-1.0	-1.9				
trouble	-2.0					

Funkcijske riječi (eng. *function words*) su „jeftinije“ (npr. $\boldsymbol{the}: \boldsymbol{-1.0}$) u odnosu na sadržajne riječi (eng. *content words*), npr. $\boldsymbol{trouble}: \boldsymbol{-2.0}$. Analogno tome, fraze koje se češće pojavljuju u korpusima (npr. $\boldsymbol{save\ yourself\ the\ trouble}: \boldsymbol{-5.0}$) „jeftinije“ su od onih rjeđih fraza (npr. $\boldsymbol{you\ should\ save}: \boldsymbol{-5.2}$), čak iako su dulje.

Sadržajne riječi su one riječi koje daju značenje, dok funkcijske riječi opisuju odnose između sadržajnih riječi. Sadržajne riječi su primjerice imenice, glagoli, pridjevi i prilozi, dok funkcijske riječi obuhvaćaju prijedloge, zamjenice, članove, veznike, negacije itd.

Treba ponoviti, da kako se radi samo o procjeni budućeg troška postoji opasnost da se, među mogućim prijevodima, u postupku odbacivanja hipoteza, uklone i oni dobri prijevodi (Koehn, 2010). Naime, nakon što se definira veličina skupa hipoteza (eng. *beam size*), novonastale hipoteze se raspoređuju, tj. slažu u stogove s obzirom na jednak broj prevedenih riječi iz izvornog jezika, tj. one riječi koje su pokrivene u hipotezi (Koehn, 2010). Zatim se stogovi pune odgovarajućim hipotezama. Iz svake hipoteze generira se nova hipoteza (proširenje hipoteze) koja se opet potom sprema u odgovarajući stog. Stogovi se s vremenom pune, a s obzirom da je stog ograničenog kapaciteta potrebno ga je očistiti (eng. *prune*). Tj. iz stogova se odbacuju one hipoteze koje se ne nalaze u definiranom skupu hipoteza, odnosno koje su opisane većim (ostvarenim i budućim) troškom (Koehn, 2006). Na kraju se kao prijevod uzima ona proširena hipoteza koja je „najjeftinija“ i koja pokriva sve riječi iz izvornog jezika.

Algoritam zrakastog pretraživanja dobio je naziv po tome što konačni prijevod nastaje kao rezultat prolaska grafom pretraživanja i sekvencijalne izgradnje hipoteza. Sama sekvencijalna izgradnja hipoteze može se zamisliti kao zraka svijetlosti koja „osvjetljava“ prostor pretraživanja, pri tome osvjetljavajući i ostale moguće prijevode koji se ne razlikuju mnogo od najboljeg, tj. konačnog izgrađenog prijevoda (Koehn, 2010).

Naravno, cilj algoritma je generirati i sačuvati „najjeftiniji“ skup proširenih hipoteza, tj. „najjeftiniji“ prijevod rečenice. Međutim, od velike koristi može biti i evidentiranje „osvijetljenih“ proširenih hipoteza koje po kvaliteti slijede odmah nakon najboljeg prijevoda (eng. *top n-best candidate translations*) (Koehn, 2010). Tako se u postupku dekodiranja može sačuvati n-najboljih proširenih hipoteza (npr. $\boldsymbol{n = 100}$), s obzirom da se vrlo često njihove vjerojatnosti prevođenja malo razlikuju (Koehn, 2015).

Izračun vjerojatnosti prevođenja dekoder vrši se za svaki djelomičan prijevod (2.50) (Koehn, 2010).

$$e = argmax_e \prod_{i=1}^{sve\ fraze} \phi(\bar{f}_i|\bar{e}_i) * d(start_i - end_{i-1} - 1) * \prod_{i=1}^{|e|} p_{JM}(e) \quad (2.50)$$

Naime, odabire se fraza $\bar{\boldsymbol{f}}_{\boldsymbol{i}}$ koja se treba prevesti u frazu $\bar{\boldsymbol{e}}_{\boldsymbol{i}}$ te se zatim u tablici fraznih struktura pronalazi vjerojatnost $\boldsymbol{\phi}(\bar{\boldsymbol{f}}_{\boldsymbol{i}}|\ \bar{\boldsymbol{e}}_{\boldsymbol{i}})$. S obzirom da prethodna fraza završava u $\boldsymbol{end_{i-1}}$, a trenutna fraza započinje u $\boldsymbol{start_i}$, treba izračunati $\boldsymbol{d(start_i - end_{i-1} - 1)}$ što predstavlja model preslagivanja, tj. premještanja mogućih prijevoda. U n-gramskom jezičnom modelu potrebno je pratiti zadnjih $\boldsymbol{n-1}$ riječi kako bi se izračuo $\boldsymbol{p_{JM}(e_i|e_{i-(n-1)}, \dots, e_{i-1})}$ za riječi $\boldsymbol{e_i}$ (Koehn, 2010).

3.4. Proširenje standardnog modela statističkog strojnog prevođenja

U prethodnim poglavljima opisan je standardni model statističkog strojnog prevođenja temeljenog na IBM-ovim modelima, koji su zatim proširen na fraze. Strojno prevođenje temeljeno na frazama generira složenije i kvalitetnije strojne prijevode u odnosu na statističke IBM-ove modele temeljene na riječima (Koehn, 2010). Tri su ključna faktora koja izravno utječu na kvalitetu standardnog modela, a time i na model proširen frazama (Koehn, 2010):

- jezični model, koji osigurava tečan strojni prijevod u ciljnom jeziku,
- prijevodni model zajedno s tablicom prijevoda fraza, tj. fraznih struktura, koji vodi računa o tome da su riječi ili nizovi riječi iz izvornog jezika ispravno uparene s riječima, odnosno nizovima u ciljnome jeziku,
- te permutacijski model: model preslagivanja, tj. premještanja redoslijeda riječi, koji osigurava ispravno preslagivanje/premještanje riječi, a time i ispravan poredak riječi u rečenici.

Model statističkog strojnog prevođenja temeljenog na frazama uključuje jezični model, prijevodni model temeljen na frazama te permutacijski model, a može se opisati kao u (2.51), pri čemu $\boldsymbol{\phi}(\bar{\boldsymbol{f}}_{\boldsymbol{i}}|\bar{\boldsymbol{e}}_{\boldsymbol{i}})$ predstavlja prijevodni model temeljen na frazama iz kojeg proizlazi tablica prijevoda fraza, tj. fraznih struktura, $\boldsymbol{p_{JM}}(\boldsymbol{e_i}|\boldsymbol{e_i} \dots \boldsymbol{e_{i-1}})$ predstavlja jezični model, a $\boldsymbol{d}(\boldsymbol{start_i} - \boldsymbol{end_{i-1}} - \boldsymbol{1})$ predstavlja model preslagivanja/premještanja fraza s obzirom na pozicije riječi (eng. *distance-based reordering model*). Takav model u pravilu destimulira premještanje ili dozvoljava premještanje samo preko određenog broja riječi (do maksimalne udaljenosti) (Koehn, 2010).

Rečenica u izvornom jeziku $\boldsymbol{f}$ podijeljena je u niz fraza $\bar{\boldsymbol{f}}_{\boldsymbol{i}}$ koje se zatim prevode u niz fraza na ciljnom jeziku $\bar{\boldsymbol{e}}_{\boldsymbol{i}}$. Model preslagivanja/premještanja fraza kažnjava pomake fraza preko velikih udaljenosti (eng. *large distances*) (Koehn, 2010). $\boldsymbol{start_i}$ i $\boldsymbol{end_i}$ predstavljaju pozicije prvih, odnosno zadnjih riječi fraze $\bar{\boldsymbol{f}}_{\boldsymbol{i}}$ koja se prevodi u frazu $\bar{\boldsymbol{e}}_{\boldsymbol{i}}$.

$$e = argmax_e \prod_{i=1}^{sve\ fraze} \phi(\bar{f}_i|\bar{e}_i) * d(start_i - end_{i-1} - 1) * \prod_{i=1}^{|e|} p_{JM}(e_i|e_i \dots e_{i-1}) \tag{2.51}$$

Ukoliko se dvije fraze prevode sekvencijalno, onda $\boldsymbol{start_i = end_{i-1} + 1}$, tj. pozicija prve riječi u frazi $\boldsymbol{i}$ jednaka je poziciji zadnje riječi u prethodnoj frazi $\boldsymbol{+1}$ (Koehn, 2010).

Standardni model statističkog strojnog prevođenja pripada kategoriji modela temeljenih na procjeni maksimalne vjerodostojnosti. Međutim, takav pristup ne dozvoljava da se primjerice jezičnom modelu dodijeli veća „težina" u odnosu na prijevodni model, tj. svi se podmodeli tretiraju jednako važnima (Koehn, 2010).

No, standardni model može se proširiti tako da se određenim faktorima u modelu statističkog strojnog prevođenja, tj. značajkama $\boldsymbol{h}$ (eng. *feature*) pridruži određena težina $\boldsymbol{\lambda}$, što je prikazano u (2.52), čime se može utjecati na poboljšanje kvalitete strojnog prijevoda (Koehn, 2010).

$$e = argmax_e \prod_{i=1}^{sve\ fraze} \phi(\bar{f}_i|\bar{e}_i)^{\lambda_\phi} * d(start_i - end_{i-1} - 1)^{\lambda_d} * \prod_{i=1}^{|e|} p_{JM}(e_i|e_i \dots e_{i-1})^{\lambda_{JM}} \tag{2.52}$$

Na taj način moguće je definirati utjecaj svakog podmodela, tj. značajke pomoću težina $\boldsymbol{\lambda_\phi}$, $\boldsymbol{\lambda_d}$ i $\boldsymbol{\lambda_{JM}}$ (Koehn, 2010). Model koji dozvoljava pridruživanje težina značajkama i koji se vrlo često koristi u strojnom prevođenju i strojnom učenju ima oblik log-linearnog modela (eng. *log-linear model*).

Log-linearni modeli općenito imaju oblik kao u (2.53) (Koehn, 2010; Koehn et al., 2005).

$$p(x) = exp \sum_{i=1}^{n} \lambda_i h_i(x) \tag{2.53}$$

Općenita jednadžba za opis log-linearnog modela može se primijeniti i na modelu statističkog strojnog prevođenja temeljenog na frazama, pri čemu se svaki podmodel (npr. prijevodni model, jezični model itd.) može opisati funkcijom značajke (eng. *feature function*) (Koehn, 2010). Ako se zada sljedeće (2.54):

$$\begin{gathered} \text{broj funkcija značajki } \boldsymbol{n = 3}, \\ \text{nasumična varijabla } \boldsymbol{x = (e, f, start, end)}, \\ \text{funkcija značajke } \boldsymbol{h_1 = log\phi}, \\ \text{funkcija značajke } \boldsymbol{h_2 = logd}, \\ \text{funkcija značajke } \boldsymbol{h_3 = logp_{JM}}, \end{gathered} \tag{2.54}$$

pri čemu $\boldsymbol{a}$ predstavlja sravnjenost (eng. *alignment*), a $\boldsymbol{d(a_i - b_{i-1} - 1)}$ predstavlja model distorzije s obzirom na danu sravnjenost, funkcija $\boldsymbol{p(e, a|f)}$ za izračun vjerojatnosti prevođenja rečenice $\boldsymbol{f}$ u rečenicu $\boldsymbol{e}$ sa sravnjenošću $\boldsymbol{a}$, modelira se kao log-linearna kombinacija funkcija značajki $\boldsymbol{h_i}$ i pripadajućih težina $\boldsymbol{\lambda_i}$ (2.55) (Koehn, 2010).

$$\begin{aligned} p(e, a|f) = \exp\Bigg[& \lambda_\phi \sum_{i=1}^{sve\ fraze} log\phi\left(\bar{f}_i|\bar{e}_i\right) \\ & + \lambda_d \sum_{i=1}^{sve\ fraze} logd(a_i - b_{i-1} - 1) \\ & + \lambda_{JM} \sum_{i=1}^{sve\ fraze} logp_{JM}(e_i|e_i \dots e_{i-1}) \Bigg] \end{aligned} \tag{2.55}$$

Prevođenje rečenice $\boldsymbol{f}$ na izvornom jeziku u rečenicu $\boldsymbol{e}$ na ciljnom jezikom pomoću statističkog strojnog prevođenja temeljenog na frazama stoga se svodi na (2.56), pri čemu $\boldsymbol{E(f)}$ predstavlja skup mogućih prijevoda rečenice $\boldsymbol{f}$ (Koehn, 2010).

$$e = argmax_{e \in E(f)} p(e, a|f) \tag{2.56}$$

Log-linearni model, za razliku od modela temeljenog na procjeni maksimalne vjerodostojnosti, omogućuje jednostavno uključivanje dodatnih značajki $\boldsymbol{h_m}$ u model, međutim, potrebno je svakoj značajci odrediti pripadajuću težinu $\boldsymbol{\lambda_m}$. Standardno se u log-linearnom modelu statističkog strojnog prevođenja upotrebljava osam značajki (España-Bonet i Gonzàlez, 2014; Koehn, 2010):

1. Jezični model, $\boldsymbol{p_{JM}(e)}$
2. Prijevodni model temeljen na frazama, $\boldsymbol{p(f|e)}$, odnosno tablica prijevoda fraza, tj. fraznih struktura $\boldsymbol{\phi(\bar{f}, \bar{e})}$
3. Prijevodni model temeljen na frazama za obrnuti jezični smjer, $\boldsymbol{p(e|f)}$, odnosno tablica prijevoda fraza, tj. fraznih struktura $\boldsymbol{\phi(\bar{e}, \bar{f})}$
4. Model leksičkih težina za prevođenje riječi (eng. *lexical weighting model*), $\boldsymbol{lex(\bar{e}|\bar{f}, a)}$
5. Model leksičkih težina za prevođenje riječi za obrnuti jezični smjer, $\boldsymbol{lex(\bar{f}|\bar{e}, a)}$
6. Penaliziranje prekratkih ili predugačkih generiranih riječi (eng. *word penalty*), $\boldsymbol{wc(e)}$
7. Penaliziranje prekratkih ili predugačkih generiranih fraza (eng. *phrase penalty*), $\boldsymbol{pc(e)}$
8. Model leksičkog preslagivanja/premještanja fraza, tj. distorzije fraza (eng. *lexicalised reordering/ distortion model*), $\boldsymbol{p_o}$

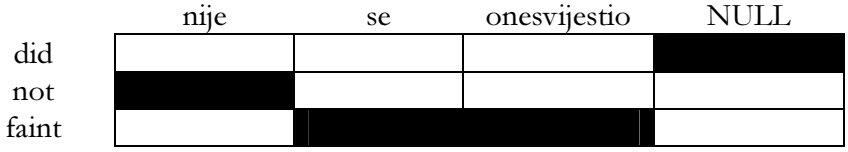
Uporaba prijevodnog modela temeljenog na frazama za obrnuti jezični smjer može uz odgovarajuću težinu znatno povećati kvalitetu strojnog prijevoda (Koehn, 2010). Nadalje, vrlo rijetki frazni parovi mogu izazvati problem u modelu strojnog prevođenja, pogotovo ukoliko proizlaze iz ulaznih podatkovnih skupova koji sadrže izvjesnu količinu šuma. Ukoliko se obje fraze $\bar{e}$, $\bar{f}$ pojave samo jednom tada je i $\boldsymbol{\phi(\bar{e}, \bar{f}) = \phi(\bar{f}, \bar{e}) = 1}$ što vrlo često precjenjuje pouzdanost fraznog para (Koehn, 2010).

Drugim riječima, rijetki frazni parovi imaju vrlo nepouzdane distribucije vjerojatnosti prevođenja fraze. Stoga je ideja frazne parove dekomponirati u prijevode riječi i analizirati međusobnu sravnjenost riječi koje sačinjavaju frazu, s obzirom da su riječi opsežnije statistički opisani, a time i pouzdanije (eng. *lexical weighting*) (Koehn, 2010). Upravo model leksičkih težina $\boldsymbol{lex(e|f)}$ procjenjuje koliko je pouzdan jedan frazni par koji se vrlo rijetko pojavljuje u korpusu na kojemu je treniran model sustava za strojno prevođenje. Radi se zapravo o metodi izglađivanja koja uzima u obzir leksičku vjerojatnost riječi. Ovaj model frazu dekomponira u prijevode riječi i analizira distribuciju vjerojatnosti riječi (jednadžba 2.57), pri čemu $\bar{\boldsymbol{e}}$ i $\bar{\boldsymbol{f}}$ predstavljaju fraze na ciljnom, odnosno izvornom jeziku, dok $\boldsymbol{a}$ predstavlja njihovu sravnjenost riječi (Koehn, 2010).

$$lex(\bar{e}|\bar{f}, a) = \prod_{i=1}^{duljina(\bar{e})} \frac{1}{|\{j|(i,j) \in a\}|} \sum_{\forall(i,j)\in a} w(e_i|f_j) \tag{2.57}$$

Svaki frazni par proizlazi iz postupka esktrahiranja fraza, tj. iz sravnjenosti riječi pa stoga za svaku frazu postoji i podatak o sravnjenosti riječi u frazi. Na temelju te sravnjenosti, moguće je izračunati vjerojatnost leksičkog prevođenja fraze $\bar{\boldsymbol{e}}$ pomoću fraze $\bar{\boldsymbol{f}}$. Prema gornjoj jednadžbi (2.57), svaka riječ u ciljnom jeziku $\boldsymbol{e_i}$ bit će generirana pomoću sravnjenih riječi iz izvornog jezika $\boldsymbol{f_i}$ s distribucijom vjerojatnosti prevođenja riječi $\boldsymbol{w(e_i|f_j)}$. Ukoliko je jedna riječ ciljnog jezika sravnjena s više riječi u izvornom jeziku, uzima se srednja vrijednost pripadajuće leksičke vjerojatnosti prevođenja (Koehn, 2010). Ukoliko pak jedna riječ u ciljnom jeziku nije sravnjena s riječi u izvornom jeziku, za nju se kaže da je sravnjena s tokenom $\boldsymbol{NULL}$, koja se također uzima u obzir pri izračunu vjerojatnosti prevođenja riječi. Primjer izračuna leksičke težine fraznog para $\langle\bar{\boldsymbol{e}}|\bar{\boldsymbol{f}}\rangle$ s danom sravnjenošću riječi $\boldsymbol{a}$ i distribucijom vjerojatnosti leksičkog prevođenja riječi $\boldsymbol{w}$ prikazan je u nastavku (Tablica 12, jednadžba 2.58)

Tablica 12. Izračun leksičke težine fraznog para.



(did not faint, nije se onesvijestio)

$$lex(\bar{e}|\bar{f}, a) = w(did|NULL) * w(not|nije) * \frac{1}{2}\big(w(faint|se) + w(faint|onesvijestio)\big) \quad (2.58)$$

Iz gornjeg primjera vidljivo je da je leksička težina fraznog para jednaka produktu triju faktora, jednog za svaku riječ u ciljnom jeziku. Implementacija modela leksičkih težina za prevođenje riječi za obrnuti jezični smjer može također znatno poboljšati kvalitetu strojnog prijevoda (Koehn, 2010).

U modelu strojnog prevođenja moguće je podesiti željenu duljinu prijevoda u ciljnom jeziku i time favorizirati ili penalizirati duljinu (broj riječi) strojnog prijevoda pomoću parametra $\boldsymbol{wc(e)}$ (eng. *word count*). To je vrlo efikasna metoda koja se upotrebljava za vrijeme ugađanja duljine segmenata u ciljnom jeziku te značajno poboljšava kvalitetu strojnog prijevoda (Koehn, 2010).

Naime, jezični model preferira kraće segmente koje je jednostavnije statistički opisati. Kako bi se osiguralo da segmenti nisu predugački ili prekratki, svakoj generiranoj riječi u ciljnom jeziku pridružuje se faktor $\boldsymbol{\omega}$ (2.59) (Boyd-Graber, 2014). Ako je $\boldsymbol{\omega} < 1$ model će preferirati kraće prijevode, a ukoliko je $\boldsymbol{\omega} > 1$ model će preferirati dulje prijevode (Koehn, 2010).

$$wc(e) = log|e|^{\omega} \quad (2.59)$$

Prije nego li se izvrši strojni prijevod neke nove proizvoljne rečenice, ona se segmentira u fraze u izvornom jeziku. Međutim, u modelu statističkog strojnog prevođenja temeljenog na frazama svi segmenti u izvornom jeziku jednako su vjerojatni, a samo odabrani frazni prijevodi zajedno s pripadajućim vjerojatnostima prevođenja i premještanja te vjerojatnostima jezičnog modela izravno utječu na novu segmentiranu rečenicu koju treba prevesti na ciljni jezik (Koehn, 2010).

U procesu odabiranja prijevoda za svaku frazu, faktorom penaliziranja fraze $\boldsymbol{\rho}$ moguće je favorizirati više kratkih fraza ($\boldsymbol{\rho > 1}$) ili manje dugačkih fraza ($\boldsymbol{\rho < 1}$) (eng. *phrase penalty*). Favoriziranje ili penaliziranje duljine fraza $\boldsymbol{pc(e)}$ opisano je u (2.60), pri čemu $\boldsymbol{I}$ predstavlja ukupnost fraza u izvornom jeziku koje treba prevesti na ciljni jezik (Koehn, 2010).

$$pc(e) = log|I|^{\rho} \tag{2.60}$$

Iako su dulje fraze rjeđe te stoga i statistički manje pouzdane, u pravilu se ipak one favoriziraju jer sadrže veći kontekst i rezultiraju kvalitetnijim strojnim prijevodom, a ionako model leksičkih težina vodi računa o nepouzdanim fraznim parovima (Koehn, 2010).

U standardnom modelu statističkog strojnog prevođenja temeljenog na frazama primjenjuje se model preslagivanja/premještanja fraza s obzirom na pozicije riječi (eng. *distance-based reordering model*) (Koehn, 2010). Takav jednostavan model distorzija fraza u ciljnom jeziku može se opisati kao u (2.61) pomoću odgovarajućeg parametra $\boldsymbol{\alpha}$, pri čemu $\boldsymbol{start_{fraza_i}}$ označava početnu poziciju fraze koja se iz izvornog jezika prevodi u $\boldsymbol{i}$**-tu** frazu ciljnog jezika, a $\boldsymbol{end_{fraza_{i-1}}}$ predstavlja završnu poziciju fraze koja se iz izvornog jezika prevodi u $(\boldsymbol{i-1})$**-tu** frazu u ciljnom jeziku (Koehn, 2015; España-Bonet i Gonzàlez, 2014; Koehn, 2010; Ohashi et al., 2005).

$$d\left(start_{fraza_i}, end_{fraza_{i-1}}\right) = \alpha^{|start_{fraza_i} - end_{fraza_{i-1}} - 1|} \tag{2.61}$$

Takav jednostavan model distorzije kažnjava pomake fraza preko velikih udaljenosti, iako je to vrlo često potrebno i ovisi o jezičnom paru. Štoviše, neke fraze se statistički gledano češće premještaju od drugih (Koehn, 2010), kao npr. u slučaju francusko-hrvatskog jezičnog para. U francuskom jeziku, pridjev i imenica koja mu prethodi zamjenjuju pozicije kada se prevode na hrvatski jezik. Očigledno je da je potrebno implementirati poseban model koji uči različite tipove ponašanja prilikom premještanja specifičnih fraznih parova u određenim jezicima.

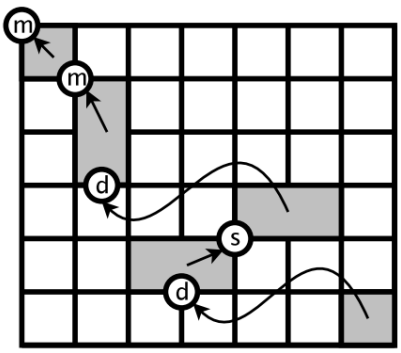
Stoga se predlaže proširenje standardnog modela preslagivanja/premještanja fraza s obzirom na pozicije riječi, na način da se uključe dodatni podatci o leksičkom preslagivanju, tj. premještanju fraza (eng. *lexicalised reordering*) (Auli et al., 2014).

Naime, za svaku ekstrahiranu frazu može se evidentirati tip orijentacije sravnjivanja iz matrice sravnjenosti riječi (Koehn, 2010). Tip orijentacije može se identificirati promatranjem zajedničkih točaka preklapanja sravnjenih riječi (eng. *alignment points*) s gornje lijeve ili gornje desne strane ekstrahiranog fraznog para. Točka preklapanja sravnjenih riječi koja upućuje na orijentaciju prema „gore lijevo“ označava da je prethodna riječ u ciljnom jeziku sravnjena s prethodnom riječi izvornog jezika. Točka preklapanja koja upućuje na orijentaciju prema „gore desno“ označava da

je prethodna riječ u ciljnom jeziku sravnjena s narednom (tj. sljedećom) riječi u izvornom jeziku (Koehn, 2010).

Za razliku od standardnog modela preslagivanja/premještanja fraza s obzirom na pozicije riječi, leksička distorzija fraza, tj. preslagivanje/premještanje fraza manifestira se u tri oblika, tj. tri su moguća tipa orijentacije (Slika 14) (Koehn, 2010; Koehn, 2008):

- monotona distorzija (eng. *monotone*), $\boldsymbol{m}$
    - ako je zajednička točka preklapanja sravnjenih riječi orijentirana „gore-lijevo" prema ekstrahiranoj frazi
    - nema preslagivanja/premještanja, već se fraza samo pomiče za $\boldsymbol{N}$ riječi
    - prethodna riječ u ciljnom jeziku sravnjena je s prethodnom riječi u izvornom jeziku
- zamjena s prethodnom frazom (eng. *swap*), $\boldsymbol{s}$
    - ako je zajednička točka preklapanja sravnjenih riječi orijentirana „gore-desno" prema ekstrahiranoj frazi
    - prethodna riječ u ciljnom jeziku sravnjena je s narednom riječi u izvornom jeziku
- diskontinuirana distorzija (eng. *discontinuous*), $\boldsymbol{d}$
    - ako zajednička točka preklapanja sravnjenih riječi ne upućuje na orijentaciju „gore-desno" ili „gore-lijevo" prema ekstrahiranoj frazi



Slika 14. Matrica sravnjenosti riječi, s prikazom tri tipa leksičke distorzije fraza (**m, s, d**).

Model leksičkog preslagivanja/premještanja fraza $\boldsymbol{p_o}$ (eng. *lexicalised reordering/distortion model*) predviđa orijentaciju tipa $\boldsymbol{m}$, $\boldsymbol{s}$, $\boldsymbol{d}$ uzevši u obzir trenutni frazni par. U takvom modelu zbraja se koliko puta se svaki ekstrahirani frazni par pojavljuje s odgovarajućim tipom orijentacije, npr. $\boldsymbol{p_o(zamjena \mid \bar{e}, \bar{f})}$ (Koehn, 2010). Drugim riječima, svaki put kada se ekstrahira frazni par, tip orijentacije u tom slučaju se također evidentira. Zbraja se koliko puta se svaki ekstrahirani frazni par pojavio sa svakim od moguća tri tipa orijentacije. Distribucija vjerojatnosti $\boldsymbol{p_o}$ se zatim procjenjuje na temelju tog zbroja pomoću procjene maksimalne vjerodostojnosti (Koehn, 2010). To rezultira trima novim značajkama, koje se procjenjuju na temelju relativne frekvencije pojavljivanja: $\boldsymbol{p_m}$, $\boldsymbol{p_s}$, $\boldsymbol{p_d}$ (jednadžba 2.62), pri čemu je $\boldsymbol{orijentacija \in \{m, s, d\}}$. Model se može opisati i s 6 značajki ukoliko se zbrajanje orijentacija vrši za oba smjera (eng. *bidirectional*) (Koehn, 2010).

$$p_o(orijentacija|\bar{f}, \bar{e}) = \frac{\sum_{\bar{f}} \sum_{\bar{e}} zbroj(orijentacija, \bar{e}, \bar{f})}{\sum_o \sum_{\bar{f}} \sum_{\bar{e}} zbroj(orijentacija, \bar{e}, \bar{f})} \tag{2.62}$$

U proširenom log-linearnom modelu statističkog strojnog prevođenja temeljenog na frazama 13 je standardnih značajki (España-Bonet i Gonzàlez, 2014):

- $\boldsymbol{p(e)}$, tj. $\boldsymbol{p_{JM}(e)}$
- $\boldsymbol{p(f|e)}$, tj. $\boldsymbol{\phi(\bar{f}, \bar{e})}$
- $\boldsymbol{p(e|f)}$, tj. $\boldsymbol{\phi(\bar{e}, \bar{f})}$
- $\boldsymbol{lex(\bar{e}|\bar{f}, a)}$, $\boldsymbol{lex(\bar{f}|\bar{e}, a)}$,
- $\boldsymbol{wc(e)}$, $\boldsymbol{pc(e)}$,
- $\boldsymbol{p_m(orijentacija|\bar{f}, \bar{e})}$, $\boldsymbol{p_s(orijentacija|\bar{f}, \bar{e})}$, $\boldsymbol{p_d(orijentacija|\bar{f}, \bar{e})}$.
- $\boldsymbol{p_m(orijentacija|\bar{e}, \bar{f})}$, $\boldsymbol{p_s(orijentacija|\bar{e}, \bar{f})}$, $\boldsymbol{p_d(orijentacija|\bar{e}, \bar{f})}$,

Međutim, moguće je dodavati i brojne druge funkcije značajki, poput dodatnih jezičnih modela, prijevodnih modela, ostalih izvora znanja itd. Težinu svake značajke treba optimizirati, tj.

treba odrediti vrijednosti koje maksimiziraju učinak modela (España-Bonet i Gonzàlez, 2014). Svaka težina značajke upućuje i na relativnu važnost značajke u modelu strojnog prevođenja.

Za određivanje optimalnih vrijednosti, tj. ugađanje (eng. *tuning*) težina značajki, primjenjuje se postupak nadziranog učenja (eng. *supervised learning*) pomoću manjeg sravnjenog paralelnog korpus koji se naziva i skup za ugađanje (eng. *tuning/development set*) (Koehn, 2010; Manning i Schütze, 1999). Sastoji se od sravnjenih rečenica (segmenata) na izvornom i ciljnom jeziku, u pravilu njih oko tisuću, te je namijenjen ugađanju komponenti modela sustava za statističko strojno prevođenje temeljeno na frazama. Odnosno, namijenjen je podešavanju parametara, tj. ugađanju težina značajki u sustavu za strojno prevođenje (Koehn, 2015).

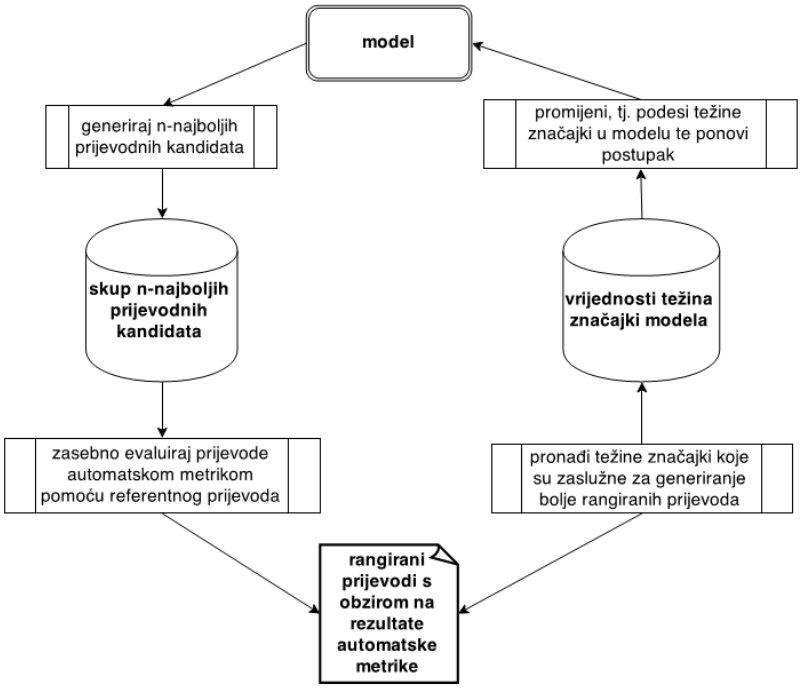
Jedna od strategija nadziranog učenja jest i diskriminativno učenje/treniranje (eng. *discriminative training*). Diskriminativno učenje koristi se za optimiziranje težina značajki (eng. *feature weights*) na skupu za ugađanje, i u svrhu razlikovanja (eng. *discriminate*) „dobrih" strojnih prijevoda od „loših". Time se izravno utječe na performanse sustava za statističko strojno prevođenje. Skup za ugađanje i skup za treniranje sustava za statističko strojno prevođenje su disjunktni skupovi, tj. ne smije biti preklapanja rečenica, tj. segmenata (Koehn, 2015).

Jedna od metoda diskriminativnog učenja i podešavanja parametara (eng. *parameter tuning*), tj. težina značajki, je treniranje s minimalnom stopom pogreške (eng. *minimum error-rate training*, *MERT*) (Dyer, 2012; Koehn, 2010; Och, 2003). Radi se o metodi koja umanjuje broj pogrešaka u strojnom prijevodu i u pravilu se uvijek izvodi u odnosu na odabranu metriku kvalitete za koju se želi postići maksimalni mogući rezultat (Slike 15 i 16, adaptirano prema Koehn, 2004b; Koehn, 2010). S obzirom da se sustav optimizira u odnosu na određenu evaluacijsku metriku na podatkovnom skupu za ugađanje (Servan i Schwenk, 2011), postoji opasnost pretjeranog ugađanja (eng. *overfitting*) sustava za strojno prevođenje (Giménez, 2008). Takav pretjerano ugođeni sustav vrlo dobro prevodi podatkovni skup za ugađanje, međutim, vrlo loše prevodi ostale podatkovne skupove, tj. nove skupove za testiranje (ispitivanje). Podatkovni skupovi koji se koriste za ugađanje značajki podupiru adaptaciju modela sustava specifičnim domenama ili određenim tipovima teksta (Koehn, 2010).

Naime, podešene, tj. ugođene težine nemaju znatnu moć poopćavanja (eng. *generalisation*), stoga se predlaže proširenje skupa za ugađanje i/ili primjena metode izgađivanja (Koehn, 2010). Općenito, model statističkog strojnog prevođenja nema veliku moć generalizacije, s obzirom da sve fraze promatra zasebno i ne prepoznaje varijante iste fraze (Sima'an, 2013). No, unatoč tome, prednosti modela temeljenog na frazama su u konačnici manja leksička dvoznačnost i efikasno permutiranje riječi.

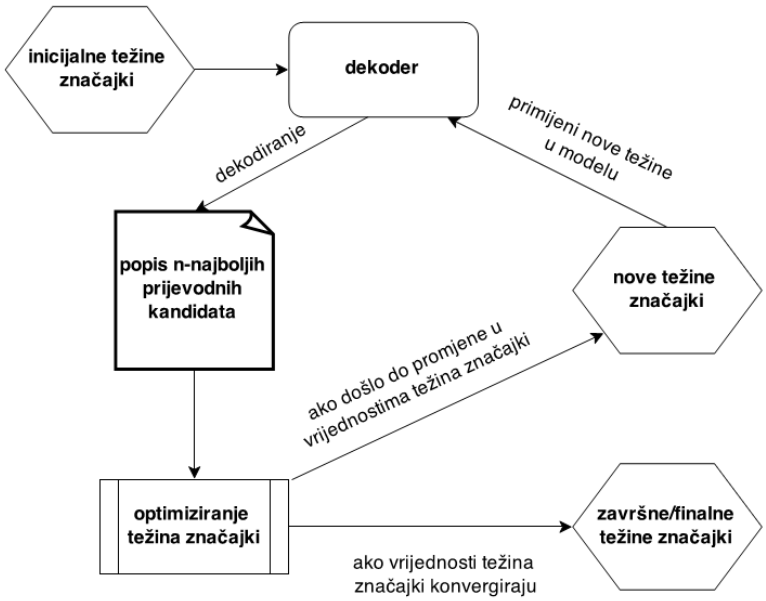
Dekoder u modelu statističkog strojnog prevođenja na temelju skupa za ugađanje generira popis n-najboljih prijevodnih kandidata, nakon čega se prijevodni kandidati (eng. *candidate translations*) evaluiraju s obzirom na odabranu automatsku metriku kvalitete. Svaki prijevodni kandidat reprezentiran je vektorom vrijednosti značajki (eng. *feature vector*), tj. uređenom n-torkom realnih brojeva (najčešće njih 10-15) (Koehn, 2010; Koehn, 2004). Nakon dekodiranja svi prijevodni kandidati iz popisa n-najboljih prijevodnih kandidata (eng. *n-best list*) rangiraju se prema rezultatima evaluacijske metrike. Metoda treniranja s minimalnom stopom pogreške zatim pronalazi težine značajki koje generiraju bolje rangirane prijevodne kandidate. Tj. istražuje se utjecaj različitih težina parametara na n-najboljih prijevodnih kandidata radi otkrivanja optimalnih postavki. Ideja je otkriti koje promjene u vrijednostima težina značajki rezultiraju najvećim poboljšanjem rezultata evaluacijske metrike (Koehn, 2004b).

Pronađene optimalne vrijednosti težina značajki reimplementiraju se u modelu statističkog strojnog prevođenja, a cijeli postupak generiranja n-najboljih prijevodnih kandidata i podešavanja parametara može se ponoviti $\boldsymbol{n}$ puta, sve do konvergencije parametara (najčešće potrebno **10-20** iteracija) (Koehn, 2004b).



Slika 15. Treniranje s minimalnom stopom pogreške.

Ono što se percipira pogreškom u strojnom prevođenju često se može identificirati metrikom kvalitete. No, takve mjere/metrike kvalitete ne koreliraju uvijek s ljudskom evaluacijom kvalitete strojnog prijevoda koja se uzima kao „zlatno“ referentno polazište za ocjenjivanje strojnog prijevoda (eng. *gold standard*). Optimizacija sustava, odnosno ugađanje težina značajki pomoću treniranja s minimalnom stopom pogreške najčešće se vrši u odnosu na metriku BLEU (Dyer, 2012; Koehn, 2004b). Naime, BLEU je vrlo brza metrika te se pokazala kao najefikasnija metrika za ugađanje statističkih modela. Tj. modeli koji su ugađani pomoću BLEU metrike daju najbolje rezultate ljudske evaluacije, a to često potvrđuju i druge metrike (Chen i Cherry, 2014).



Slika 16. Iterativno ugađanje težina značajki pomoću treniranja s minimalnom stopom pogreške (MERT).

Budući da se broj pogrešaka odražava na vrijednost evaluacijske metrike, ugađanje parametara može se zamisliti kao proces traženja vrijednosti za skup parametara, tj. vrijednosti težina $\boldsymbol{\lambda_1, \ldots, \lambda_n}$ za značajke $\boldsymbol{h_1, \ldots, h_n}$, pri kojima je broj pogrešaka u strojnom prijevodu minimalan.

S obzirom da je prostor pretraživanja mogućih vrijednosti težina značajki višedimenzionalan (13 značajki odgovara 13 dimenzija pretraživanja), tj. velik, za pretraživanje prostora predlaže se Powellova metoda pretraživanja (eng. *Powell's search method*) (Foster i Kuhn, 2009). Ideja ove metode je u prostoru vrijednosti parametara pronaći vrijednosti (točke u prostoru) koje, što je moguće bolje, aproksimiraju pravac u prostoru (Koehn, 2010).

Drugim riječima, prvo se na temelju podatkovnog skupa za ugađanje generira popis n-najboljih prijevodnih kandidata s inicijalnim vrijednostima težina značajki (eng. *basic parameter setting*) (Koehn, 2015). Zatim se za prvi popis n-najboljih prijevodnih kandidata pronalaze optimalne vrijednosti težina. Pretraživanje Powellovom metodom ide sekvencijalno, parametar po parametar i po svakoj osi, tj. dimenziji. Ukoliko se nova optimalna vrijednost razlikuje od prethodne vrijednosti, stara vrijednost težine zamjenjuje se novom u modelu strojnog prevođenja. Potom se nanovo generira popis n-najboljih prijevodnih kandidata pomoću novih težina značajki. Ukoliko se prijevodi razlikuju, popis se može samo nadopuniti novim prijevodima ili se stari prijevodi zamjenjuju novima. Tj., postupak pretraživanja nastavlja se iz pronađenog optimuma, a cjelokupni postupak se iterira desetak do dvadesetak puta, sve dok vrijednosti konačno ne konvergiraju prema jednoj vrijednosti težine (Foster i Kuhn, 2009).

Bitan nedostatak MERT metode jest što se može koristiti samo za treniranje, tj. ugađanje modela s manjim brojem značajki, najčešće desetak značajki (Arun i Koehn, 2007), ali ne više od dvadesetak (Dyer, 2012). Nadalje, značajke koje imaju potencijal statistički značajno doprinijeti u modelu strojnog prevođenja mogu u postupku ugađanja biti odbačene ukoliko MERT metoda za njih ne uspije pronaći zadovoljavajuće težine značajki (Foster i Kuhn, 2009).

Osim MERT-a, za optimizaciju parametara mogu se koristiti „PRO“ (eng. *Pairwise ranked optimization*) (Hopkins i May, 2011) i „MIRA“ (eng. *Margin Infused Relaxed Algorithm*) (Hasler et al., 2011; Crammer i Singer, 2003).

Danas se istražuju i metode poput perceptrona (Koehn, 2010; Manning i Schütze, 1999) koje omogućuju kompleksno treniranje modela (eng. *large-scale discriminative training*) s oko milijun značajki (Liang et al., 2006), te brojne druge metode koje dozvoljavaju ugađanje do čak 35 milijuna značajki (Tillmann i Zhang, 2006) koristeći izravnu optimizaciju težina pomoću metrike BLEU. Treba spomenuti da MIRA dozvoljava treniranje modela s nekoliko tisuća značajki (Chiang et al., 2009).

Za ekstrakciju najboljih prijevodnih kandidata, graf pretraživanja koji je nastao u procesu dekodiranja pomoću metode zrakastog pretraživanja, može se pretvoriti i u rešetku riječi (eng. *word lattice*) koja se u pravilu prikazuje u obliku konačnog automata (eng. *finite state machine*). Takav automat određen je brojem stanja, početnim i završnim stanjima te prijelazima između stanja (Manning i Schütze, 1999). Prijelazi između stanja modelirani su prema distribucijama vjerojatnosti (eng. *probabilistic finite state machine*) (Koehn, 2010).

3.5. Mogućnosti implementacije jezičnog znanja u model sustava za statističko strojno prevođenje

Standardni log-linearni model sustava za statističko strojno prevođenje temeljeno na frazama ne uključuje jezično znanje, tj. svi procesi unutar modela odvijaju se na čisto statističkom principu (España-Bonet i Gonzàlez, 2014; Koehn, 2010). Riječi se tretiraju kao obični nizovi znakova, a rečenice kao skup nizova znakova odvojenih bjelinom. Takav sustav potpuno različito tretira primjerice riječi ***prijevod*** i ***prijevodi***, iako im je korijen riječi jednak.

Morfološki bogati jezici generiraju velik broj oblika riječi, a time proširuju i vokabular u standardnom sustavu za statističko strojno prevođenje. To uzrokuje mnogobrojne probleme koji proizlaze uglavnom iz nedostatka potrebne količine podataka (eng. *sparse data*).

Jezik je živo „biće". Radi se o vrlo kompleksom fenomenu i teško ga je matematički opisati i procesirati. Stoga se pojavljuje potreba za implementacijom dodatnog jezičnog znanja u model sustava za statističko strojno prevođenje. Integracijom jezičnog znanja mogu se bolje opisati uloge sadržajnih i funkcijskih riječi, imena i brojevi (primjerice XML oznakama, eng. *XML markup*) (Manning i Schütze, 1999), pravila preslagivanja/premještanja riječi u rečenici te razriješiti anafore (eng. *anaphora resolution*).

Nadalje, prirodni jezik ima specifično svojstvo – rekurziju, koja omogućuje izgradnju vrlo složenih ugniježđenih rečeničnih konstrukcija sa sintaksno povezanim riječima (Koehn, 2010). Pored toga, izgradnja složenica (eng. *compound words*) također uzrokuje probleme u standardnom modelu statističkog strojnog prevođenja, s obzirom da se složenicama može generirati velik broj riječi izvan vokabulara (Fishel i Sennrich, 2014; Stymne, 2011; Knight i Koehn, 2003).

Implementacija jezičnog znanja, međutim, ima jedan vrlo važan nedostatak – zahtijeva posebno pripremljene, tj. označene (tzv. anotirane) korpuse koji u pravilu nisu dostupni te ih vrlo često treba mukotrpno ručno izgraditi. Nadalje, dodavanje sintaksnih anotacija povećava kompleksnost u modelu statističkog strojnog prevođenja.

Postoji nekoliko varijacija modela sustava za statističko strojno prevođenje koji dozvoljavaju implementaciju jezičnoga znanja (España-Bonet i Gonzàlez, 2014; Koehn, 2010):

- faktorirani prijevodni model (eng. *factored translation model*), pri čemu faktor predstavlja svaki oblik dodatne informacije pridružene jednoj (ali svakoj) riječi

- sintaksno-temeljeni prijevodni model (eng. *syntactic translation model*), koji sadrži sintaksu izvornog i/ili ciljnog jezika

Faktorirani prijevodni model (eng. *factored translation model*) je proširenje modela statističkog strojnog prevođenja temeljenog na frazama. Radi se o tome da je svaka riječ zamijenjena vektorom faktora: $\boldsymbol{riječ} \Longrightarrow (\boldsymbol{riječ, lema, PoS\ oznaka, morfologija,} \ldots)$, pri čemu lema predstavlja osnovni oblik riječi, a POS oznaka (eng. *part of speech tags*) (Manning i Schütze, 1999) definira vrstu riječi. Načelno se proces prevođenja može podijeliti u tri koraka (Koehn, 2010), što je prikazano na primjeru prijevoda riječi $\boldsymbol{osobni\ automobil}$ na engleski jezik:

1. prijevod leme ili korijena riječi iz izvornog jezika u ciljni (korjenovanje riječi, eng. *stemming*, vrši se najčešće odbacivanjem sufiksa, tj. morfema; Manning i Schütze, 1999)

$$osobni\ automobil \rightarrow car, automobile, motorcar$$

2. prijevod morfološkog faktora i faktora POS oznaka, pri čemu $\boldsymbol{NN}$ predstavlja imenicu

$$NN|množina - nominativ - muški\_rod \rightarrow NN|množina, NN|jednina$$

3. generiranje oblika riječi (eng. *surface forms*) u ciljnom jeziku na temelju prevedene leme ili korijena riječi, prevedenih POS oznaka i morfološkog faktora

$$car|NN|množina \rightarrow cars$$
$$car|NN|jednina \rightarrow car$$
$$automobile|NN|množina \rightarrow automobiles$$
$$automobile|NN|jednina \rightarrow automobile$$
$$motorcar|NN|množina \rightarrow motorcars$$
$$motorcar|NN|jednina \rightarrow motorcar$$

Primjena gornja tri koraka na frazu u izvornom jeziku naziva se proširenje (eng. *expansion*) (Koehn, 2010). U nastavku je prikazan primjer proširenja:

$$osobni\ automobili|osobni\ automobil|NN|množina$$

1. prijevod: mapiranje lema

$$\{?|car|?|?,\quad ?|automobile|?|?,\quad ?|motorcar|?|?\}$$

2. prijevod: mapiranje morfologije

$$\{?|car|NN|množina,\quad ?|automobile|NN|množina,\quad ?|motorcar|NN|množina,$$
$$?|car|NN|jednina,\quad ?|automobile|NN|jednina,\quad ?|motorcar|NN|jednina\}$$

3. generiranje oblika riječi

$$\{cars|car|NN|množina,\quad automobiles|automobile|NN|množina,$$
$$motorcars|motorcar|NN|množina,$$
$$car|car|NN|jednina,\quad automobile|automobile|NN|jednina,$$
$$motorcar|motorcar|NN|jednina\}$$

Faktorirani prijevodni model zahtijeva anotirani paralelni korpus, u kojemu su jezične jedinice gramatički obilježene (Koehn i Haddow, 2012b). Za svaki tip faktora može se izraditi poseban jezični model, a samo strojno prevođenje odvija se na istom principu kao i kod standardnog log-linearnog modela (España-Bonet i Gonzàlez, 2014). Treći korak, tj. korak generiranja oblika riječi, implicira da je treniranje potrebno samo na strani ciljnoga jezika te predstavlja jednu vrstu jezičnog modela. Faktorirani prijevodni model pokazao se vrlo efikasnim u prevođenju morfološki bogatih jezika s velikom slobodom premještanja/preslagivanja redoslijeda riječi, a u log-linearni model sustava za statističko strojno prevođenje se može integrirati u obliku funkcije značajke (Koehn, 2010).

Sintaksno-temeljeni prijevodni model pogodan je za jezike s različitom strukturom rečenice (Yamada i Knight, 2001), a može se realizirati na tri načina (España-Bonet i Gonzàlez, 2014; Koehn, 2010):

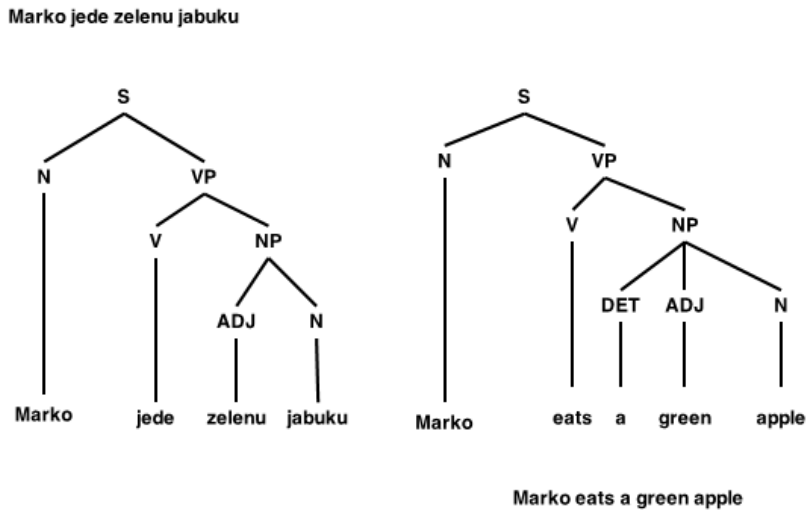
- stablo-u-niz (eng. *tree-to-string*), pri čemu se sintaksno stablo (eng. *parse tree*) u ciljnom jeziku mapira u niz znakova u izvornom jeziku,
- niz-u-stablo (eng. *string-to-tree*), pri čemu se niz znakova u ciljnom jeziku mapira u stablo u izvornom jeziku,
- stablo-u-stablo (eng. *tree-to-tree*), pri čemu se sintaksno stablo mapira u sintaksno stablo.

Ovaj pristup temelji se na ideji prevođenja cjelovitih sintaksnih jedinica, umjesto prevođenja zasebnih riječi ili fraza. Primjerice, sintaksno stablo može prikazati cjelokupnu rečenicu ili njen podskup. Sintaksno-temeljeno prevođenje također je pogodno za jezike s velikom slobodom u poretku riječi, a pogotovo kada prijevod glagola ovisi o subjektu ili objektu u rečenici. Također, ovaj pristup strojnom prevođenju jasnije definira ulogu prijedloga ili članova u rečenici. No, postupak dekodiranja znatno se razlikuje u odnosu na statističko strojno prevođenje temeljeno na riječima ili frazama, s obzirom da se prijevod ne izgrađuje s lijeva na desno, već se izgrađuje pomoću gramatičkih pravila (eng. *grammar rules*) i parsanjem (eng. *parsing*) (Koehn, 2010).

Hijerarhijsko strojno prevođenje temeljeno na frazama kombinira prednosti statističkog strojnog prevođenja temeljenog na frazama i sintaksna pravila pomoću beskontekstnih gramatika (eng. *hierarchical phrase-based machine translation*) (Chiang, 2005).

Beskontekstne gramatike definiraju strukturu mogućih rečenica i pripadajuća sintaksna stabla te se sastoji od neterminala (oznaka), terminala (riječi), startnog simbola i pravila koja mapiraju jedan neterminal s lijeve strane (eng. *left-hand side*) u najmanje jedan ili niz terminala i neterminala s desne strane (eng. *right-hand side*) (Jurafsky i Martin, 2013; Koehn, 2010; Manning i Schütze, 1999).

Slika 17 prikazuje sintaksna stabla u izvornom i ciljnom jeziku na primjeru rečenice ***Marko jede zelenu jabuku***, pri čemu ***S*** označava početak rečenice, ***VP*** glagolsku skupinu, ***NP*** imensku skupinu, ***V*** glagol, ***N*** imenicu, ***ADJ*** pridjev, a ***DET*** član. Npr., pravilo beskontekstne gramatike $\boldsymbol{NP \rightarrow DET\ ADJ\ N}$ dozvoljava izgradnju imenske fraze koja se sastoji od člana, pridjeva i imenice.



Slika 17. Prikaz sintaksnih stabala s primjerom.

U hijerarhijskom strojnom prevođenju koriste se sinkronijske gramatike (eng. *synchronous grammars*) koje istovremeno izgrađuju dva stabla – jedno za izvorni te jedno za ciljni jezik (Koehn, 2010). Takve sinkronijske gramatike dozvoljavaju modificiranje stabala i time rješavaju problem preslagivanja/premještanja riječi. Gramatička pravila pohranjuju se u tzv. prefiksnom stablu (eng. *prefix tree*), a to omogućuje da se za dane riječi određenog raspona u izvornom jeziku pronađu adekvatna pravila (Koehn, 2010). Hijerarhijsko strojno prevođenje koristi velik broj istih metoda i statističkih paradigmi za učenje iz anotiranih korpusa kao i statističko strojno prevođenje temeljeno na frazama. Međutim, vrlo bitna razlika je u postupku dekodiranja, s obzirom da se prijevodi više ne mogu izgrađivati od lijevo prema desno, već se koristi metoda odozdo-gore (eng. *bottom-up*) i tzv. *chart* parsanje (eng. *chart parsing*).

U ovom doktorskom radu jezično znanje nije implementirano u sustave za statističko strojno prevođenje.

3.6. Adaptacija domene u modelu statističkog strojnog prevođenja

Sustavi za statističko strojno prevođenje u pravilu se izgrađuju za određenu domenu (Vertan i Duma, 2013; Chatzitheodorou i Poulis, 2012), s obzirom da je svako područje specifično (različit vokabular, terminologija, sintaksa, diskurs itd.). Domene se mogu zamisliti kao područja interesa, poput politike, sporta, gospodarstva, prava, tehnike, prognoze vremena itd., za koja se želi optimizirati ishod strojnog prevođenja. Primjerice, riječ ***miš*** je leksički dvoznačna: može se odnositi na životinju (domena biologije) ili računalnu ulaznu jedinicu (domena računalna tehnologije), a sam kontekst jednog relevantnog korpusa odredit će prikladno značenje. U nastavku je na primjeru povratnog glagola ***smijati se*** prikazano na koji način domena utječe na odabir riječi i rečeničnu konstrukciju.

- Općeniti korpus: ***Smijem se.***
- Medicina: ***Osmijeh je važan za proizvodnju serotonina.***
- Patentna dokumentacija: ***Senzor za detekciju facijalnih izraza i osmijeha***
- Računalni softver: ***Umetnite smješka u polje za unos.***
- Filmski titlovi (podslova): ***hahaha***
- Društvene mreže: ***LOL*** (akronim za eng. *laughing out loud*)
- Parlamentarne rasprave: ***Nakon glasovanja, sabornicom se prolomio smijeh.***
- Vijesti: ***Smijehom protiv stresa***

Statističko strojno prevođenje temelji se na pretpostavci da se distribucija podataka iz podatkovnog skupa za treniranje modela strojnog prevođenja podudara s distribucijom podataka iz podatkovnog skupa za testiranje sustava za strojno prevođenje, tako da se „pravila prevođenja" naučena iz paralelnih korpusa mogu primijeniti i na nove rečenice koje tek treba prevesti iz testnog podatkovnog skupa (Carpuat et al., 2012). Ipak, takva pretpostavka u praksi nije održiva.

Naime, sustavi za statističko strojno prevođenje razvijaju se vrlo često pomoću velikih općenitih paralelnih korpusa (eng. *general domain/out-of-domain corpora*) koji uključuju raznorazne domene, s obzirom da je na raspolaganju mnogo više općenitih korpusa nego specifičnih i visokokvalitetnih korpusa iz određene domene (eng. *in-domain corpora*).

No, veliki općeniti korpusi koji se koriste za treniranje modela, a koji uključuju različite domene, u pravilu su ipak premaleni da bi bili pogodni za treniranje specijaliziranih modela te prevođenje vrlo specifičnih domena. Drugim riječima, premaleni su da bi pokrili terminologiju specifičnu za određenu domenu, karakteristične fraze i izraze ili posebne rečenične konstrukcije. Jednako tako, sustav treniran na specifičnim korpusima vrlo loše prevodi testne podatkovne skupove izvan trenirane domene (Carpuat et al., 2012; Pecina et al., 2012).

Načelno u statističkom modelu vrijedi, što je više podataka uključeno to će i strojni prijevodi biti bolji (Turchi et al., 2012b), posebno ukoliko je na raspolaganju velika količina podataka iz specifične domene za koju se izgrađuje sustav za strojno prevođenje (Offersgaard i Hansen, 2012). Nadalje, kako bi se poboljšala kvaliteta strojnog prijevoda, ugađanje se također uvijek vrši prema specifičnoj domeni, tj. pomoću domenski specifičnog skupa za ugađanje modela i time predstavlja prvi korak u adaptaciji domene u sustavu za statističko strojno prevođenje.

Idealno, podatkovni skupovi za treniranje statističkih modela odgovaraju budućoj primjeni sustava za strojno prevođenje. Međutim, kako vrlo često nema dovoljno velikih podatkovnih skupova za treniranje statističkih modela za određenu domenu, u postupku treniranja koriste se dostupni korpusi koji su vrlo često znatno različiti u odnosu na domenu koju u konačnici treba prevoditi sustav za statističko strojno prevođenje (Koehn i Schroeder, 2007). Izazov je stoga, kako model sustava za strojno prevođenje koji je istreniran na jednoj domeni prilagoditi, tj. adaptirati za prevođenje druge domene?

Neusklađenost između domene za koju postoji dovoljna količina podatkovnih skupova i specifične domene za koju ne postoji dovoljna količina podatkovnih skupova, rezultira potrebnom za adaptacijom domene (eng. *domain adaptation*) u sustavu za strojno prevođenju. Pod pojmom adaptacija (prilagodba) domene podrazumijevaju se aktivnosti koje se poduzimaju radi povećanja kvalitete strojnog prijevoda u određenoj domeni. Ideja je izgraditi i/ili prilagoditi statističke modele koji poboljšavaju performanse strojnog prevođenja u specifičnoj domeni, s potencijalno negativnim utjecajem na performanse prevođenja tekstova iz drugih domena (Sennrich, 2012).

U pravilu se već istreniran, tj. izgrađen sustav za strojno prevođenje adaptira određenoj domeni. Naime, redovito se sustav za statističko strojno prevođenje izgrađuje općenitim

podatkovnim skupovima, a zatim se sustav prilagođava prevođenju tekstova iz vrlo specifičnih domena. Računalna adaptacija domene posredovana je računalom, tj. računalno se pripremaju ili prilagođavaju podatkovni skupovi za određenu domenu ili se procesi prilagodbe sustava za strojno prevođenje izvode pomoću računala.

Ograničavanjem sustava za strojno prevođenje na određenu domenu postižu se kvalitetniji strojni prijevodi, a razlozi su mnogobrojni: smanjuju se vokabular, leksička varijabilnost i semantička dvoznačnost u modelu strojnog prevođenja, nove rečenice koje treba prevesti slične su onima koje su korištene za treniranje modela te ih je stoga jednostavnije statistički opisati i strojno prevesti. Ograničavanjem na određenu domenu smanjuje se perpleksnost modela (Sennrich, 2012b), prijevodi su konzistentniji, a statistički modeli kompaktniji.

Adaptacijom domene utječe se i na količinu riječi izvan vokabulara. Primjerice, korpus koji pokriva domenu sporta vjerojatno ne sadrži riječi poput ***Cortijev organ*** ili ***Fibonaccijev niz***. Takve riječi se vjerojatnije mogu pronaći u korpusima koji pokrivaju medicinsku, odnosnu matematičku terminologiju. Pa čak i da se nađu u tom obliku u korpusima koji se odnose na sport, navedeni termini su vjerojatno opisani drugim kontekstom, tj. različitim glagolima, pridjevima, objektima i subjektima itd.

Postoji mnoštvo metoda za adaptaciju postojećih sustava za statističko strojno prevođenje kada je na raspolaganju specifičan podatkovni skup iz određene domene (Duma i Vertan, 2013; Giménez Linares, 2008): može se primjerice upotrijebiti leksičko-semantička mreža WordNet (Jurafsky i Martin, 2013) za ekstrakciju glosa, tj. kratkih rječničkih definicija koje se pojavljuju uz svaki skup sinonima (eng. *synset*), i to radi anotiranja specifičnih korpusa za određenu domenu ili radi izgradnje prijevodnih modela. Predložena je i integracija domenski specifičnog rječnika u prijevodni model, koji onda pokriva unutardomenske pojmove (Langlais, 2002).

Mogu se upotrijebiti vanjski izvori jezičnog znanja, kao što su jednojezični rječnici radi izgradnje prijevodnih modela (Zhang i Zong, 2013) i dodatnih jezičnih modela koji se mogu implementirati u log-linearni model sustava za statističko strojno prevođenje (Giménez Linares, 2008; Wu et al., 2008). Naime, log-linearni model dozvoljava dodavanje i kombiniranje brojnih domenskih i izvandomenskih podmodela, tj. značajki (Farzindar i Khreich, 2012; Niehues i Waibel, 2012; Razmara et al., 2012; Koehn, 2010; Lavergne et al., 2011; Schwenk i Koehn, 2008; Bulyko et al., 2007; Koehn i Schroeder, 2007). Više jezičnih modela može se kombinirati pomoću zasebnih funkcija značajki, čije se težine mogu odrediti pomoću treniranja s minimalnom stopom pogreške (MERT) primjenom podatkovnog skupa za ugađanje (Niehues et al., 2010).

Različiti podmodeli mogu se linearno interpolirati (Mansour i Ney, 2012; Banerjee et al., 2011; Mohit et al., 2009), kao i tablice prijevoda fraza, tj. fraznih struktura (Durrani et al., 2013). Linearna interpolacija dvaju prijevodnih modela treniranih na različitim korpusima može se formalizirati kao u (2.63), pri čemu se težina $\boldsymbol{\lambda}$ kreće u rasponu od $\mathbf{0}$ do $\mathbf{1}$ i može se ugađati (Vertan i Duma, 2013).

$$p(f|e) = \lambda p_1(f|e) + (1-\lambda)p_2(f|e) \tag{2.63}$$

Predložena je i uporaba paralelnog korpusa iz specifične domene za adaptaciju prijevodnog modela, s fokusom na greške koje, pri prevođenju teksta iz specifične domene, generira sustav za strojno prevođenje koji je treniran na podatkovnom skupu izvan domene (Formiga et al., 2012).

Mogu se primijeniti i ontologije (Vertan i Duma, 2013), jednojezični korpusi za adaptaciju jezičnog modela te paralelni korpusi za adaptaciju modela sravnjivanja riječi pomoću interpolacije leksičke vjerojatnosti, fertiliteta i distorzije (Wu et al., 2005).

Nadalje, može se izvršiti konkatenacija domenski specifičnog korpusa i općenitog korpusa prije izgradnje jezičnog i prijevodnog modela (Koehn i Schroeder, 2007). Predložena je i kombinacija jezičnog modela treniranog na specifičnom korpusu te prijevodnog modela treniranog na konkateniranom podatkovnom skupu (Koehn i Schroeder, 2007). Na taj način preferiraju se rečenične konstrukcije iz određene domene, dok sustav istovremeno raspolaže velikom tablicom prijevoda fraza, tj. fraznih struktura. Dodavanjem što većeg domenski specifičnog korpusa mogu se poboljšati rezultati strojnog prevođenja. Međutim, prikupljanje velike količine domenski specifičnih korpusa je zahtjevno i skupo (Giménez Linares, 2008).

Specijalizirani sustavi za strojno prevođenje izgrađeni su pomoću statističkih modela treniranih na vrlo specifičnim korpusima, kao npr. onima koji sadrže filmske titlove (podslova) (Sousa et al., 2011; Volk i Harder, 2007), terminologiju iz domene računalnog softvera (Khalilov i Choudhury, 2012) ili automobilske industrije (Läubli et al., 2013), specifični internetski žargon i izraze s društvenih mreža (Jiang et al., 2012), ekonomsko-pravnu (Pouliquen et al., 2013; Pouliquen et al., 2012) ili patentnu dokumentaciju (Junczys-Dowmunt i Pouliquen, 2014; Ceauşu et al., 2011; Hardt i Elming, 2010).

Prijevodni model može se trenirati korpusom u kojemu su rečenice pomno odabrane (eng. *corpus pre-selection/text discarding*), tj. korpusom koji je unaprijed pripremljen te koji sadrži vrlo slične rečenice kao i testni podatkovni skup (Hildebrand et al., 2005), što se može odrediti pomoću

izračuna perpleksnosti jezičnog modela (Yasuda et al., 2008). Jednako tako, i jezični model može se izgraditi pomoću sličnih rečenica (Eck et al., 2004). Rečenice svojstvene jednoj domeni mogu se ekstrahirati iz usporedivih korpusa (eng. *comparable corpora*) (Stein, 2013; Sharoff, 2012; Mohammadi i GhasemAghaee, 2010; Smith et al., 2010; Munteanu i Marcu, 2005) primjenom metoda iz područja pretraživanja i dohvaćanja informacija kroz različite jezike (eng. *cross-language information retrieval, CLIR*) (Ruopp i Xia, 2008; Snover et al., 2008).

Usporedivi korpus je skup tekstova u izvornom i ciljnom jeziku koji raspravljaju o istoj ili sličnoj tematici, međutim, takvi tekstovi nisu paralelni niti nužno rečenično sravnjeni, tj. ne radi se o međusobnim prijevodima (npr. isti članci na Wikipediji na različitim jezicima). Vrlo često se usporedivi korpusi koriste iz domene vijesti i novinskih članaka, budući da se iste vijesti često prevode na više jezika (Vertan i Duma, 2013; Matsoukas et al., 2009). S obzirom da je sustav za statističko strojno prevođenje temeljno na frazama vrlo robustan na pogrešno sravnjene rečenice (Goutte et al., 2012), usporedivi korpusi mogu se kombinirati i s paralelnim korpusima te se time može umanjiti količina riječi izvan vokabulara, što je vrlo pogodno za jezike, za koje je na raspolaganju manja količina resursa (eng. *low-resourced languages*) (Irvine i Callison-Burch, 2013).

Može se izgraditi domensko specifični jezični model koji pokriva isključivo rečenične konstrukcije svojstvene ciljnoj domeni (Sethy et al., 2006). Predlaže se i kombinacija statističkih podmodela treniranih na podatkovnim skupovima iz različitih domena uz adekvatno podešavanje težina svakog podmodela (Foster i Kuhn, 2007). Sustav za statističko strojno prevođenje treniran na malenom domenski specifičnom paralelnom korpusu može se unaprijediti dodavanjem izvandomenskih korpusa (Vogel i Tribble, 2002). Analiziran je i utjecaj dodavanja izvandomenskog korpusa u različitim fazama izgradnje sustava za strojno prevođenje na performanse sustava (Haddow i Koehn, 2012).

Moguće je interpolirati više jezičnih modela (općeg i specifičnog) te zatim dati veću težinu preferiranom jezičnom modelu (Koehn i Schroeder, 2007). Ispitane su i mogućnosti kombiniranja statističkih modela iz općenite domene i domene patenata (Ceauşu et al., 2011). Predloženo je i dodavanje kvalitetnih strojnih prijevoda iz specifične domene u podatkovne skupove za treniranje statističkih modela (Bertoldi i Federico, 2009; Schwenk i Senellart, 2009; Schwenk, 2008).

Za smanjivanje broja riječi izvan vokabulara u prijevodnim modelima i izgradnju specifičnih jednojezičnih i višejezičnih resursa iz domene medicine mogu se pretraživati izvori na internetu (Lu et al., 2014). Potrebni podatkovni skupovi za adaptiranje sustava za statističko strojno prevođenje mogu se prikupiti pretraživanjem interneta koje je orijentirano na specifičnu domenu

(eng. *domain-focused web crawling*) (Pecina et al., 2011). Mogu se primijeniti i tehnike za normalizaciju podatkovnih skupova te poluautomatska klasifikacija riječi radi smanjivanja broja riječi izvan vokabulara (Banerjee et al., 2012).

Integracija dvojezične terminologije morfološki označene XML-om u sustav za statističko strojno prevođenje te uporaba pseudo-paralelnog korpusa za treniranje statističkog modela također mogu poboljšati kvalitetu strojnog prijevoda (Weller et al., 2014).

Upotrebom identifikatora korpusa u faktoriranom prijevodnom modelu moguće je preferirati frazne parove iz određene domene (Niehues i Waibel, 2010). Predloženo je i kombiniranje prijevodnog modela i modela preslagivanja/premještanja iz specifične domene s modelima izvan domene (Nakov, 2008). Nadalje, za potrebe adaptacije domene moguće je uključiti i HM model sravnjivanja u model statističkog strojnog prevođenja (Civera i Juan, 2007).

Eksperimentirano je i s adaptacijom sustava za više domena pomoću klasifikatora domena/tema (Hasler et al., 2014) koji detektira rečenicu na izvornom jeziku te ju adekvatno klasificira kako bi sustav za strojno prevođenje mogao odabrati prikladne težine značajki specifične za određenu domenu te prikladni jezični model (Wang et al., 2012).

Sustav se može adaptirati i modificiranom Moore-Lewis metodom filtriranja (Axelrod et al., 2011): ideja je trenirati jezične modele na općenitom i specijaliziranom korpusu te zatim ukloniti sve rečenične parove koji postižu malene vjerojatnosti u jezičnom modelu treniranom na specijaliziranom korpusu.

Metoda kombiniranja modela s popunjavanjem (eng. *fill-up combination*) je pogodna ukoliko je relevantnost pojedinih modela *a priori* poznata (Bisazza et al., 2011). Ovom metodom mogu se koristiti vjerojatnosti iz prijevodnog modela treniranog na specifičnom korpusu, dok za sve ostale frazne parove koji se ne pojavljuju u specifičnom korpusu koristi se tablica prijevoda fraza, tj. fraznih struktura iz prijevodnog modela treniranog na korpusu izvan domene, kao neka vrsta *back-off* modela (Niehues, 2014).

Navedenom metodom mogu se sačuvati unosi i vjerojatnosti iz prvog modela i dodavati unosi iz drugog modela samo ukoliko se radi o novim, tj. različitim unosima. Za razlikovanje fraznih parova iz općenitog i specifičnog korpusa predlaže se uporaba dodatne značajke u obliku indikatora porijekla fraznih parova (Green et al., 2014; Bisazza et al., 2011).

Da bi se različiti prijevodni modeli mogli kombinirati, potrebno ih je prvo zasebno izgraditi pomoću dva različita podatkovna skupa (Haque et al., 2009), najčešće pomoću općenitih te specijaliziranih korpusa. Nakon što se izgrade prijevodni modeli mogu se kombinirati na više

načina (Nakov i Ng, 2012; Haque et al., 2009; Foster i Kuhn, 2007; Koehn i Schroeder, 2007), kao npr.:

- linearnom interpolacijom,
- log-linearnom interpolacijom,
- tablice prijevoda fraza, tj. fraznih struktura mogu se spojiti, a porijekla fraznih parova može se pratiti primjerice pomoću indikatora u obliku dodatne značajke,
- istovremenom primjenom oba prijevodna modela pomoću metode alternativnog puta dekodiranja (eng. *alternate/alternative/multiple decoding path*) (Koehn, 2015; Gong et al., 2012; Sennrich, 2012b; Birch et al., 2007; Koehn et al., 2007; Koehn i Schroeder, 2007).

U ovom doktorskom radu računalna adaptacija domene izvršena je pomoću:

- ugađanja podatkovnim skupom iz specifične domene,
- kombinacije domenskih i izvandomenskih značajki u modelu sustava za statističko strojno prevođenje,
- kombinacije različitih prijevodnih modela (eng. *translation model mixture*), pri čemu se jedan prijevodni model preferira metodom alternativnog puta dekodiranja,
- te *back-off* modela.

Radi se o intrinzičnom tipu adaptacije, s obzirom da se ne koriste vanjski izvori znanja, već se performanse izgrađenog sustava za statističko strojno prevođenje nastoje poboljšati isključivo s resursima koji su već na raspolaganju.

Istovremenom uporabom oba prijevodna modela pomoću metode alternativnog puta dekodiranja moguće je za svaku rečenicu u izvornom jeziku iz oba prijevodna modela odrediti skup mogućnih prijevoda (eng. *translation options*) koje treba uzeti u obzir za vrijeme dekodiranja (Niehues i Waibel, 2010). Naime, svaki put dekodiranja odnosi se na jednu tablicu prijevoda fraza, tj. fraznih struktura (Haque et al., 2009).

Nadalje, ovim pristupom moguće je odrediti način na koji dekoder odabire frazne parove, s obzirom na vjerojatnost svakog fraznog para (eng. *phrase pair score*). Ideja kombiniranja različitih prijevodnih modela metodom alternativnog puta dekodiranja jest dekoder prisiliti na odabir fraznih parova (puta izgradnje fraze) s najvećom vjerojatnošću ukoliko se frazni parovi pojave u više prijevodnih modela. Drugim riječima, ukoliko se primjenjuje metoda alternativnog puta dekodiranja i mogući prijevod (eng. *translation option*) se pojavi u identičnom obliku u obje tablice fraznih struktura, tada se odabire onaj prijevod s većom vjerojatnošću (eng. *higher scoring*) (Koehn, 2015).

U ovom doktorskom istraživanju zasebno su izgrađena po dva prijevodna modela pomoću dva različita podatkovna skupa koji se zatim istovremeno koriste pomoću metode alternativnog puta dekodiranja (Sennrich, 2011; Birch et al., 2007). Takav pristup ima brojne prednosti (Nakov i Ng, 2012; Sennrich, 2012b):

- određivanjem puta dekodiranja moguće je preferirati frazne parove iz jednog prijevodnog modela po izboru (najčešće model treniran na specifičnoj domeni),
- moguće je razlikovati porijeklo fraznih parova, tj. odrediti kojem prijevodnom modelu pripadaju,
- kombiniranje različitih prijevodnih modela uz primjenu adekvatnih težina nužno rezultira boljim ili barem jednako dobrim performansama sustava za strojno prevođenje,
- na raspolaganju je veći broj leksičkih vjerojatnosti te vjerojatnosti prevođenja fraza te se time smanjuje broj riječi izvan vokabulara,
- primjenom dodatnog većeg prijevodnog modela (najčešće model treniran na općenitom korpusu) pokrivaju se frazni parovi koji nisu nužno pokriveni manjim prijevodnim modelom,
- odvojenim, tj. zasebnim korištenjem većeg prijevodnog modela ne umanjuju se vjerojatnosti prevođenja fraza iz manjeg prijevodnog modela,
- kombiniranjem većeg i manjeg prijevodnog modela proširuje se skup mogućih prijevoda.

Međutim, metoda alternativnog puta dekodiranja ima i određene nedostatke (Sennrich, 2012b):

- alternativno dekodiranje povećava kompleksnost generiranja strojnog prijevoda,
- težine modela mogu biti podešene tako da se mogući prijevodi u određenom modelu ignoriraju,
- metoda nije pogodna za velik broj različitih prijevodnih modela,
- odabir fraznih parova s najvećom vjerojatnošću (eng. *highest-scoring phrase pair*) pogoduje odabiru fraznih parova koji možda značajno odstupaju od prosječnih vrijednosti (eng. *statistical outlier*), s obzirom na određenu količinu šuma i nedostatak potrebne količine podataka (eng. *data sparseness*) te se time umanjuje i robusnost sustava za strojno prevođenje.

Treba ponoviti da n-gramski jezični modeli mogu koristiti *back-off* metodu kao mogućnost povlačenja s jezičnog modela višeg reda na jezični model nižeg reda i time pokriti specifične rečenične konstrukcije (Koehn, 2010). Međutim, *back-off* metoda može se primijeniti i u slučaju prijevodnih modela, i to kada prvi prijevodni model ne pokriva određenu frazu (Habash et al., 2011). Naime, problem oskudne ili nedovoljne količine podataka (eng. *data sparsity*) za treniranje prijevodnog modela glavni je razlog povlačenja prema drugom prijevodnom modelu.

Metoda alternativnog puta dekodiranja zahtijeva dva (ili više) prijevodna modela, a *back-off* metoda implicira niz prijevodnih modela, pri čemu se model s manjim prioritetom koristi isključivo kada prijevodni model većeg prioriteta nije dovoljan, tj. kada ne pokriva specifične frazne parove. Drugim riječima, prijevodni model koji je treniran na općenitom korpusu koristi se samo ukoliko se u (preferiranom) prijevodnom modelu treniranom na specifičnom korpusu ne mogu pronaći mogući prijevodi, ili obrnuto (Yıldırım i Tantuğ, 2015).

Primjenom alternativnog puta dekodiranja dozvoljava se izgradnja skupa mogućih prijevoda pomoću oba prijevodna modela. U takvom hibridnom sustavu za strojno prevođenje, koji ne sadrži isključivo podmodele svojstvene samo jednoj domeni već kombinira prijevodne modele iz različitih domena, skup mogućih prijevoda pretražuje se prvo u prijevodnom modelu treniranom na domenski specifičnom podatkovnom skupu, a tek zatim u prijevodnom modelu treniranom na nekoj drugoj (općenitoj) domeni ukoliko mogući prijevodi nisu pronađeni u prvom, tj.

preferiranom modelu (Yıldırım i Tantuğ, 2015). U tom slučaju, prijevodni model treniran na općenitoj domeni predstavlja *back-off* prijevodni model za riječi i fraze koje nisu viđene u prvoj, tj. preferiranoj tablici prijevoda fraza, tj. fraznih struktura. Time se mogu poboljšati performanse sustava za statističko strojno prevođenje (Habash et al., 2011). U hibridnom sustavu mogu se kombinirati i drugi domenski te izvandomenski modeli, poput modela distorzije itd.

U ovom doktorskom istraživanju za izračun vjerojatnosti prevođenja fraze (eng. *scoring*) koriste se obje tablice prijevoda fraza, tj. fraznih struktura (općenita i specifična). To znači da se svaki mogući prijevod (eng. *translation option*) prikuplja iz obje tablice prijevoda fraza, tj. fraznih struktura te se izračun vjerojatnosti prevođenja fraze vrši s obzirom na svaku tablicu prijevoda fraza, tj. fraznih struktura. To isto tako znači da se pri primjeni metode alternativnog puta dekodiranja za izračun vjerojatnosti prevođenja fraze, mogući prijevod mora nalaziti u obje tablice prijevoda fraza, tj. fraznih struktura (Koehn, 2015; Koehn et al., 2007).

U ovom istraživanju eksperimentirano je s više različitih prijevodnih modela, a *back-off* model je korišten ukoliko primjena metode alternativnog puta dekodiranja nije bila uspješna, tj. ukoliko preferirana tablica prijevoda fraza, tj. fraznih struktura nije sadržavala odgovarajući prijevod. *Back-off* model je primijenjen samo za nepoznate n-grame duljine **1** (unigrame), tj. riječi duljine **1** koje u preferiranom prijevodnom modelu predstavljaju riječi izvan vokabulara.

4. EVALUACIJA KVALITETE STROJNOG PRIJEVODA

Evaluacija, tj. vrednovanje kvalitete strojnog prijevoda vrlo je važan element u ciklusu izgradnje sustava za strojno prevođenje (España-Bonet i Gonzàlez, 2014; Gonzàlez, 2014), međutim, kvalitetu strojnog prijevoda nije jednostavno procijeniti. U nastavku je dan primjer (adaptiran prema Koehn, 2010) izvorne rečenice na engleskom jeziku te četiri pripadajuće strojno prevedene rečenice na ciljnom jeziku (hrvatskom), generirane od četiri različita sustava za strojno prevođenje:

Croatia's authorities are keeping the stadiums safe.

sustav za strojno prevođenje 1: Hrvatska je zadužena za sigurnost na stadionu.

sustav za strojno prevođenje 2: Hrvatske vlasti su odgovorne za stadionsku sigurnost.

sustav za strojno prevođenje 3: Hrvatske službe održavaju red na stadionu.

sustav za strojno prevođenje 4: Službe u Hrvatskoj vode računa o sigurnosti na stadionu.

Postavlja se pitanje, da li strojni prijevodi na hrvatskom jeziku odgovaraju izvornoj rečenici na engleskom jeziku? Ako odgovaraju, do koje mjere odgovaraju izvornoj rečenici? Koji je sustav za strojno prevođenje generirao najbolji prijevod, a koji najlošiji? Koji je prijevod najjednostavniji, a koji najkompleksniji? Da li su prijevodi namijenjeni određenoj skupini korisnika (npr. izvornim govornicima)? Da li je kvaliteta strojnog prijevoda zadovoljavajuća s obzirom na namjenu strojnog prijevoda (visokokvalitetni profesionalni prijevod ili grubo razumijevanje sadržaja)?

Iz gornjeg primjera vidljivo je da evaluacija strojnog prijevoda za čovjeka može predstavljati vrlo zahtjevan zadatak i da je ponekad zaista vrlo teško objektivno procijeniti da li je strojni prijevod dobar ili loš, prikladan ili neprikladan. Ponekad je skoro nemoguće jednoznačno izdvojiti najbolji mogući prijevod iz skupa generiranih strojnih prijevoda.

Ljudska evaluacija je vrlo subjektivna. Naime, ljudski mozak može vrlo brzo nadomjestiti eventualne nedostatke u strojnom prijevodu, zanemariti pravopisne pogreške ili logički zaključiti koja bitna informacija nedostaje u prijevodu.

Jednako tako, na ljudsku prosudbu kvalitete utječe je li evaluator prije procjene strojnog prijevoda pročitao referentni, tj. ljudski prijevod ili tekst na izvornom jeziku. Pored toga, da li evaluator razumije materiju, odnosno tekst za koji ocjenjuje kvalitetu prijevoda? Kakvi su mu afektivni stavovi općenito prema strojnom prevođenju (npr. averzija) itd.? Kakvo mu je psiho-fizičko stanje i raspoloženje za vrijeme ocjenjivanja kvalitete strojnog prijevoda? U svakom slučaju, može se postaviti puno pitanja. Odnosno, mnoštvo elemenata izravno utječe na eksperimentalno okruženje u kojem ljudski evaluatori donose prosudbu o kvaliteti strojnih prijevoda.

Situacija se može i obrnuti, može se zamisliti strojno prevedena rečenica na ciljnom jeziku te nekoliko različitih, no vrlo sličnih pripadajućih ljudski prevedenih rečenica, također na ciljnom jeziku. Problem ocjenjivanja kvalitete ljudskih prijevoda bi bio jednako složen. Naime, ljudski prevoditelji primjenjuju različite strategije pri prevođenju te koriste različit vokabular za prenošenje informacije pa čak i kad prevode vrlo kratke rečenice. Takva raznolikost u ljudskom prevođenju otežava evaluaciju kvalitete prijevoda.

Postupak evaluacije strojnih prijevoda složen je proces, a ispitivanje kvalitete treba biti pomno osmišljeno kako bi se izbjegla pristranost te dobile korisne povratne informacije (Dillinger i Marciano, 2012; O'Brien, 2010).

Primjerice, važni su način odabira i veličina podatkovnog skupa za testiranje (ispitivanje) kvalitete strojnog prijevoda, domena prijevoda, broj i profil evaluatora, kriteriji i skale za ocjenjivanje kvalitete, načini prezentiranja i interpretiranja rezultata itd.

Nadalje, za ljudsku evaluaciju strojnog prijevoda potrebni su često evaluatori s posebnim vještinama te eksperti u određenoj domeni koji uz to razumiju kako izvorni tako i ciljni jezik. Međutim, takvi evaluatori u pravilu su skupi, a i rijetko su na raspolaganju (Koehn, 2010). Stoga se redovito angažiraju evaluatori koji razumiju samo ciljni jezik, te koji uz dane referentne ljudske prijevode mogu ocijeniti gramatičku ispravnost i prenesenost značenja u strojnom prijevodu ili prema kvaliteti rangirati dva sustava za strojno prevođenje (eng. *system ranking*) (Koehn, 2010).

Evaluacija strojnog prijevoda od iznimne je važnosti za razvoj sustava za strojno prevođenje, s obzirom da rezultati evaluacije otkrivaju slabosti i nedostatke takvog sustava, omogućavaju identifikaciju pogrešaka koje se pojavljuju za vrijeme prevođenja te omogućuju njihovu složenu analizu.

No, ljudska evaluacija je preskupa i prespora pa se zbog toga upotrebljavaju metrike koje automatski mogu ocijeniti strojni prijevod bez intervencije čovjeka.

Ipak, čovjekova procjena realnije opisuje prikladnost i kvalitetu strojnog prijevoda. Stoga ni ne čudi što se danas u pravilu primjenjuju metrike koje pozitivno koreliraju s ljuskom evaluacijom.

Također treba ponoviti da rezultati automatske evaluacije omogućuju ugađanje težina pojedinih značajki u modelu strojnog prevođenja te podešavanje različitih parametara prema vlastitim potrebama. Isto tako, daju smjernice kako implementirati značajke u modelu strojnog prevođenja te kako testirati, ispitati performanse sustava za strojno prevođenje.

4.1. Ljudska evaluacija kvalitete strojnog prijevoda

Za ljudsku evaluaciju strojnog prijevoda potrebni su: strojni prijevod na ciljnom jeziku, tekst na izvornom jeziku te eventualno referenti ljudski prijevod. Ljudska evaluacija strojnog prijevoda u pravilu je uvijek orijentirana prema procjeni dvaju ključnih kriterija: adekvatnost/točnost (eng. *adequacy*) i fluentnost/tečnost (eng. *fluency*) (España-Bonet i Gonzàlez, 2014; Gonzàlez, 2014; Koehn, 2010).

Adekvatnost je kriterij prema kojemu se procjenjuje količina prenesenog značenja iz izvornog u ciljni jezik. Drugim riječima, adekvatnost ispituje da li je značenje u ciljnome jeziku sačuvano u potpunosti, ili da li je dio informacije izgubljen, dodan ili izmijenjen (Koehn, 2010). Fluentnost ocjenjuje gramatičko oblikovanje rečenice, izbor riječi te pridržavanje jezičnih standarda. Tim kriterijem procjenjuje se da li je strojni prijevod „u duhu jezika".

Strojni prijevod je kvalitetan ukoliko je tražena (tj. potrebna) razina adekvatnosti i fluentnosti zadovoljena, ukoliko prijevod odgovara ciljnoj skupini i namjeni te ukoliko zadovoljava sve ostale bitne specifikacije i potrebe korisnika (Lommel, 2013).

Kriteriji adekvatnosti i fluentnosti ocjenjuju se na Likertovoj skali od 1-4, 1-5 ili 1-10 (Schaefer et al., 2014; O'Brien, 2010; Roturier i Bensadoun, 2011) ili po principu „prošao-nije prošao" (eng. *pass-fail*) (Dillinger i Marciano, 2012). Primjer takve skale dan je u nastavku (Tablica 13).

Tablica 13. Prikaz skale za procjenu fluentnosti/tečnosti i adekvatnosti/točnosti strojnog prijevoda.

fluentnost/tečnost		**adekvatnost/točnost**	
prijevod je…		unačenje je preneseno…	
4	besprijekoran	**4**	u potpunosti
3	dobar	**3**	u većoj mjeri
2	dijelom razumljiv	**2**	u manjoj mjeri
1	nerazumljiv	**1**	nije preneseno

Na manjim skalama ocjenjivanje je konzistentnije, s obzirom da čovjek teško razlučuje granične razlike među ocjenama, kao npr. između ocjena **7** i **8**, ili **8** i **9**. Sve popularnija skala je ona s rasponom ocjena od 1-3, s time da je **3** najbolja, a **1** najlošija ocjena (eng. *good*, *bad*, *ugly*).

Prijevod koji je ocijenjen s ***good***, tj. s **3**, je dobar i razumljiv prijevod. Prijevod ocijenjen s ***bad***, tj. s ocjenom **2**, vrijedi ručno ispravljati, tj. doraditi. Strojni prijevod ocijenjen s ***ugly***, tj. s **1**, treba u potpunosti odbaciti, jer ga se ne može ispraviti uz minimalan napor.

Takve skale često su popraćene i pojašnjenjem ocjena, kao što je prikazano na jednom primjeru u nastavku (Tablica 14) (Dillinger i Marciano, 2012).

Tablica 14. Primjer skale s pojašnjenjem ocjena.

ocjena prijevoda		**pojašnjenje**
5	odličan	prijevod je točan i jasan gramatika, vokabular i stil su prikladni
4	vrlo dobar	manje pogreške u prijevodu značenje je jasno preneseno
3	dovoljno dobar	pogreške u prijevodu značenje u prijevodu odgovara značenju u izvornom jeziku prijevod razumljiv
2	nedovoljno dobar	pogreške u gramatici, vokabularu i stilu otežavaju razumijevanje prijevoda značenje jedva razumljivo
1	loš	velik broj pogrešaka u gramatici i vokabularu onemogućuju razumijevanje prijevoda značenje u prijevodu izgubljeno
0	greška u sustavu	za prijevode koji se ne mogu ocijeniti prema gornjim ocjenama nerazumljivi znakovi cijele rečenice neprevedene

Postupak ljudskog, tj. ručnog dorađivanja strojnog prijevoda naziva se naknadno uređivanje ili post-editiranje (eng. *post-editing*) (Hutchins, 2005b). U statističkom strojnom prijevodu često se mogu pronaći razni tipovi pogrešaka, poput dodavanja riječi, ispuštanja riječi, netočnog velikog ili malog slova, gubitka interpunkcije, a pored toga, neke fraze mogu biti vrlo tečne, dok druge nisu nimalo fluentne (O'Brien, 2010). No, unatoč određenom broju pogrešaka u strojnom prijevodu, post-editiranje može strojni prijevod učiniti vrlo korisnim (Federico et al., 2012). U procesu post-editiranja uzimaju se u obzir vrijeme koje je potrebno da se strojni prijevod ispravi, broj potrebnih izmjena riječi te broj potrebnih operacija i manipulacija tekstom, broj pritisaka na

tipkovnici (eng. *keystrokes*) itd. (Koehn i Germann, 2014; Morado Vázquez et al., 2013; Koehn, 2012; De Almeida i O'Brien, 2010; Plitt i Masselot, 2010). Može se načelno tvrditi da se post-editiranje isplati ako je trošak naknadne dorade strojnog prijevoda manji od troška ljudskog prevođenja ispočetka (Koehn, 2010). Za post-editiranje strojnog prijevoda, klasifikaciju i anotiranje tipova pogrešaka u strojnom prijevodu koriste se brojni alati, kao što su PET (eng. *Post-Editing Tool*) (Aziz et al., 2012; Aziz i Specia, 2012), DQF (eng. *Dynamic Quality Evaluation Framework*) (Garcia, 2014; Görög, 2014), Translog II (Carl, 2012), BLAST (eng. *BiLingual Annotator/Annotation/Analysis Support Tool*) (Stymne, 2012), translate5 (Lommel et al., 2014) itd.

Ljudska evaluacija je spora, a pored toga i vrlo subjektivna. Naime, određeni tipovi pogrešaka više utječu na ljudsku percepciju kvalitete; npr. strojni prijevod koji se sastoji dijelom od neprevedenih riječi, bit će u pravilu ocijenjen najlošijom ocjenom. Stoga se predlaže veći broj evaluatora, nakon čega se računa razina (ne)slaganja između ljudskih evaluatora (eng. *inter-annotator agreement*), primjerice pomoću Cohenovog kappa koeficijenta (3.1) (Callison-Burch et al., 2012; Koehn, 2011).

$$K = \frac{p(A) - p(E)}{1 - p(E)} \tag{3.1}$$

$\boldsymbol{p(A)}$ je relativno slaganje između ljudskih evaluatora (omjer), dok $\boldsymbol{p(E)}$ predstavlja vjerojatnost da se radi o slučajnom slaganju među evaluatorima. Npr. ako je skala od $\mathbf{1}$ do $\mathbf{4} \rightarrow \boldsymbol{p(E)} = \frac{\mathbf{1}}{\mathbf{4}}$. Za mjerenje razine interne konzistentnosti (eng. *internal consistency*) među evaluatorima predlaže se i uporaba Cronbach alphe (Seljan et al., 2015). Ako je (3.2) tada se Cronbach alpha može definirati kao u (3.3), pri čemu $\boldsymbol{\sigma_X^2}$ predstavlja varijancu opaženih ukupnih testnih vrijednosti (eng. *total test scores*), a $\boldsymbol{\sigma_{Y_i}^2}$ predstavlja varijancu komponente $\boldsymbol{i}$ na trenutnom uzorku evaluatora. Vrijednost $\boldsymbol{\alpha \geq 0.9}$ upućuje na visoku konzistentnost među evaluatorima, $\boldsymbol{0.8 \leq \alpha < 0.9}$ dobru konzistentnost, $\boldsymbol{0.7 \leq \alpha < 0.8}$ prihvatljivu, $\boldsymbol{0.6 \leq \alpha < 0.7}$ upitnu, $\boldsymbol{0.5 \leq \alpha < 0.6}$ slabu i $\boldsymbol{\alpha < 0.5}$ neprihvatljivu konzistentnost.

$$X = Y_1 + Y_2 + \cdots + Y_K \tag{3.2}$$

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K} \sigma_{Y_i}^2}{\sigma_X^2}\right) \tag{3.3}$$

Konzistentnost ocjenjivanja samo jednog evaluatora također se može ispitati (eng. *intra-annotator agreement*) (Callison-Burch et al., 2012). Za izračun korelacije automatskih metrika ($\boldsymbol{x}$) s ljudskom evaluacijom kvalitete ($\boldsymbol{y}$) koriste se Pearsonova korelacija $\boldsymbol{r_{xy}}$ (3.4) (na razini sustava; Gonzàlez, 2014) ili Spearmanova korelacija (Graham i Baldwin, 2014; Koehn, 2010) te Kendallov tau (na razini segmenta; Gonzàlez, 2014).

$$r_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \tag{3.4}$$

Srednja vrijednost $\boldsymbol{\bar{x}}$ i varijanca uzorka $\boldsymbol{s_x^2}$ definirani su u nastavku (3.5, 3.6) (Koehn, 2010).

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.5}$$

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{3.6}$$

Jednako tako mogu se izračunati i srednja vrijednost $\boldsymbol{\bar{y}}$ te varijanca uzorka $\boldsymbol{s_y^2}$. Vrijednost Pearsonovog koeficijenta korelacije kreće se od $\boldsymbol{+1}$ (savršena pozitivna korelacija) do $\boldsymbol{-1}$ (savršena negativna korelacija). Pri $\boldsymbol{r_{xy} = 0}$ dvije su varijable međusobno nezavisne (Koehn, 2010). Ljudska evaluacija je konzistentnija pri rangiranju cjelovitih sustava prema kvaliteti strojnog prijevoda nego pri dodjeljivanju apsolutnih ocjena pojedinim strojnim prijevodima (Koehn, 2010). Ljudska evaluacija često se provodi i na način da evaluatori odgovaraju na pitanja o razumijevanju sadržaja (eng. *content understanding assessment*): tko?, što?, kada?, gdje? itd.

4.2. Automatska evaluacija kvalitete strojnog prijevoda

Zadatak automatske evaluacije kvalitete strojnog prijevoda je izračunati sličnost strojnog prijevoda s jednim ili više referentnih, tj. ljudskih prijevoda (eng. *reference translation*). Stoga se i metrika može zamisliti kao funkcija koja definira udaljenost između dva elementa koja se ispituju. Referentni prijevod je prijevod visoke kvalitete i najčešće se ljudski prijevod jedne rečenice uzima kao referentni prijevod.

Sličnost se odnosi na razinu preklapanja, tj. podudaranja strojnog prijevoda s referentnim prijevodom, što u praksi nije pogodno za jezike s bogatom morfologijom i relativno slobodnim poretkom riječi. Današnje metrike pretežno se temelje na leksičkoj sličnosti skupova podataka (eng. *lexical similarity*) koje treba evaluirati (España-Bonet i Gonzàlez, 2014). Automatske metrike za izračun kvalitete strojnog prijevoda ubrzavaju postupak izgradnje sustava za strojno prevođenje (Gonzàlez, 2014). One olakšavaju složenu analizu pogrešaka u procesu strojnog prevođenja i analizu mogućih uzroka, omogućuju rangiranje prijevodnih kandidata i optimiziranje vrijednosti parametara radi poboljšavanja performansi sustava. Uz to, omogućuju i usporedbu s drugim sustavima strojnog prevođenja. Dobra automatska metrika ima nekoliko ključnih karakteristika (Koehn, 2011; Koehn, 2010):

- malen trošak (eng. *low cost*) – metrika ne troši puno vremena niti računalnih resursa,
- brzina (eng. *speed*) – metrika je brza,
- točnost (eng. *correct*) – metrika bolje rangira kvalitetnije sustave za strojno prevođenje,
- mogućnost ugađanja (eng. *tuneable*) – parametri u modelu strojnog prevođenja trebaju se moći optimizirati prema zadanoj metrici,
- suvislost (eng. *meaningful*) – rezultati evaluacije trebaju omogućiti smislenu interpretaciju rezultata i prosudbu kvalitete strojnog prijevoda,
- konzistentnost/dosljednost (eng. *consistent*) – ponavljanje postupka evaluacije određenom metrikom treba dati jednake rezultate,

- stabilnost/pouzdanost/općenitost (eng. *stable/reliable/general*) – rezultati evaluacije jednog dijela podatkovnog skupa za testiranje konzistentna je s rezultatima evaluacije drugog dijela podatkovnog skupa za testiranje.

Metrikama se mjeri učinkovitost sustava, no one omogućuju i detektiranje slabosti sustava. Slabosti se mogu ukloniti podešavanjem parametara u sustavu za statističko strojno prevođenje (Denkowski i Lavie, 2010; Agarwal i Lavie, 2008). Identifikacija slabosti time izravno doprinosi povećanju kvalitete automatskog prijevoda, a posebno je važna za morfološki bogate jezike ili jezike s ograničenom količinom kvalitetnih paralelnih korpusa.

Glavne prednosti automatskih metrika za evaluaciju kvalitete strojnog prijevoda u odnosu na ljudsku evaluaciju su objektivnost, brzina i ponovna iskoristivost.

Aktualne metrike danas temelje se na (España-Bonet i Gonzàlez, 2014):

- udaljenosti uređivanja (eng. *edit distance*), npr. WER (eng. *word-error rate*), PER (eng. *position-independent word error rate*; Popović i Ney, 2011), TER (eng. *translation error/edit rate*) itd.
- preciznosti, npr. BLEU (eng. *BiLingual Evaluation Understudy*), NIST (eng. *National Institute of Standards and Technology*) itd.
- odziv, npr. ROUGE (eng. *Recall-Oriented Understudy for Gisting Evaluation*; Lin, 2004) itd.
- omjeru preciznosti i odziva, npr. GTM (eng. *General Text Matcher*), METEOR (eng. *Metric for Evaluation of Translation with Explicit ORdering*) itd.

WER metrika temelji se na Levenshteinovoj udaljenosti (eng. *Levenshtein distance*) (Jurafsky i Martin, 2013) i poretku riječi, a u kontekstu strojnog prevođenja predstavlja najmanji broj potrebnih izmjena da bi se strojni prijevod pretvorio u referenti prijevod (jednadžba 3.7) (Matusov et al., 2005; Tomás et al., 2003; Nießen et al., 2000). Prema WER-u tri su moguće izmjene:

- zamjena riječi, tj. jedna se riječ zamjenjuje drugom (eng. *substitution*)
- umetanje riječi (eng. *insertion*)

- ispuštanje riječi (eng. *drop/deletion*)

$$WER = \frac{broj\_zamjena + broj\_umetanja + broj\_ispuštanja}{duljina\_referentnog\_prijevoda} \quad (3.7)$$

WER metrika ne prepoznaje, tj. ne dozvoljava premještanje riječi: riječ koja je prevedena ispravno no koja se nalazi na krivoj poziciji u rečenici bit će penalizirana kao ispuštena riječ u odnosu na strojni prijevodu, odnosno kao umetnuta riječ u odnosu na referentni prijevod (Specia, 2010).

TER metrika odražava potreban broj izmjena da se strojni prijevod preoblikuje u referentni prijevod, uzevši u obzir i broj riječi u referentnom prijevodu (tj. prosječan broj riječi ukoliko se koristi više referentnih prijevoda) (Cer et al., 2010; Snover et al., 2005), što je prikazano u (3.8).

$$TER = \frac{broj\_izmjena}{prosječan\_broj\_referentnih\_riječi} \quad (3.8)$$

TER metrikom se želi aproksimirati napor, tj. količina post-uređivanja potrebna za ljudsko ispravljanje strojnog prijevoda (Specia, 2010). TER može koristiti više referentnih prijevoda; u tom slučaju se za svaki referenti prijevod računa minimalni broj izmjena potrebnih da bi se strojni prijevod preoblikovao u referenti prijevod. Sam TER je tada omjer minimalnog broja izmjena te prosječnog broja riječi u referentnom prijevodu. TER se također temelji na Levenshteinovoj udaljenosti, no za razliku od WER-a, dozvoljena je još jedna operacija – premještanje/skok bloka teksta (eng. *block movement/jump*) (Koehn, 2010) te se time obuhvaćaju operacije nad frazama (premještanje) (Snover et al., 2006). TER dozvoljava sljedeće izmjene (Liu et al., 2011):

- zamjena riječi, tj. jedna se riječ zamjenjuje drugom (eng. *substitution*)
- umetanje riječi (eng. *insertion*)
- ispuštanje riječi (eng. *drop/deletionn*)
- pomak bloka teksta (eng. *shift*)

Pomak bloka teksta odnosi se na pomak niza riječi unutar strojnog prijevoda. Pomak proizvoljnog niza riječi (bez obzira na udaljenost) računa se kao jedna izmjena (eng. *one edit; same edit cost*). Na taj način se ublažavaju rezultati WER metrike koja drastično penalizira strojni prijevod ukoliko se poredak riječi iz strojnog prijevoda točno ne podudara s poretkom riječi u referentnom prijevodu. Naime, WER ne pridaje nikakav značaj strojnom prijevodu ukoliko se sastoji od točnih riječi na krivim pozicijama. Optimalan niz izmjena, uključujući i pomake blokova teksta (fraza) vrlo je teško pronaći. Zapravo radi se o NP-kompletnom problemu (eng. *NP-complete problem*) (Koehn, 2010), stoga se i koristi *greedy* algoritam pretraživanja (eng. *greedy search algorithm*) za odabiranje skupa pomaka bloka teksta (Specia, 2010; Snover et al., 2006):

1. pri svakom koraku izračunati broj umetanja, ispuštanja i zamjena riječi (Levenshteinova udaljenost) primjenom dinamičkog programiranja,
2. odabrati pomak bloka teksta koji najviše, tj. najefikasnije umanjuje Levenshteinovu udaljenost,
3. ponoviti sve dok više ne preostaje niti jedan pomak koji umanjuje Levenshteinovu udaljenost.

TER s jednim referentnim prijevodom može jednako dobro korelirati s ljudskom evaluacijom kvalitete strojnog prijevoda kao i BLEU metrika koja koristi četiri referentna prijevoda (Snover et al., 2006). Jedna varijanta TER metrike je poluautomatska metrika HTER (eng. *human-targeted TER*). Ideja je da ljudski evaluator kreira referentni prijevod koristeći kao podlogu strojni prijevod. Radi se o minimalnom broju potrebnih izmjena da se strojni prijevod pretvori u fluentan prijevod s ispravnim značenjem (Stymne, 2008). Zatim se takav prijevod može koristiti kao referenti prijevod za izračun drugih metrika. Takva varijacija TER metrike vrlo je skupa, s obzirom da zahtijeva dugotrajno ljudsko post-uređivanje strojnog prijevoda (3-7 minuta po rečenici), no zato visoko korelira s ljudskom evaluacijom kvalitete strojnog prijevoda (bolje od TER-a) (Snover et al., 2006; Specia, 2010).

Za razliku od metrika GTM, BLEU, NIST i METEOR, za WER i TER metrike vrijedi „manje je bolje“. Vrijednost **0** za metrike WER ili TER upućuje na to da nije potrebno izvršiti izmjene strojnog prijevoda (savršeno podudaranje s referentnim prijevodom). U teoriji, gornja vrijednost TER (i WER) metrike može biti i veća od **1**, s obzirom da duljine strojnog i referentnog

prijevoda mogu biti beskonačno različite (npr. u slučaju netočnog prijevoda). No, u pravilu se rezultati WER i TER metrika kreću u rasponu $[\mathbf{0}, \mathbf{1}]$.

GTM metrika zbraja riječi koje se podudaraju u strojnom i referentnom prijevodu te na taj način računa njihovu međusobnu sličnost. Preciznije, GTM računa ispravan broj unigrama, tj. ispravan broj preklapanja unigrama koji se odnose na riječi koje se ne ponavljaju u strojnom i referentnom prijevodu. GTM favorizira dulja preklapanja n-grama u ispravnom poretku (redoslijedu), dodjeljuje im veću težinu te se temelji na preciznosti (broj točnih tokena podijeljen s brojem tokena u strojnom prijevodu), odzivu (broj točnih tokena podijeljen s brojem tokena u referentnom prijevodu) i izračunu F-mjere (eng. *F-measure*) (Koehn, 2010; Melamed et al., 2003; Turian et al., 2003; Manning i Schütze, 1999). F-mjera omogućuje kombiniranje preciznosti i odziva, a temelji se na njihovoj harmonijskoj sredini. Vrijednosti GTM metrike kreću se u rasponu $[\mathbf{0}, \mathbf{1}]$, pri čemu $\mathbf{1}$ predstavlja potpuno podudaranje strojnog i referentnog prijevoda (tekstovi su identični), a $\mathbf{0}$ znači da nema preklapanja riječi.

U nastavku je prikazan primjer rečenice s pripadajućim izračunima, pri čemu su dijelovi rečenice koji se preklapaju u strojnom i referentnom prijevodu podcrtani (3.9, 3.10, 3.11):

strojni prijevod: Kuhamo večeru i uređujemo kolač jer nam parents u goste dolaziti .

referentni prijevod: Kuhamo večeru i pripremamo puno kolača jer nam roditelji dolaze u goste .

$$preciznost = \frac{broj\_preklapajućih\_tokena}{duljina\_strojnog\_prijevoda\_u\_tokenima} = \frac{8}{12} = 0.67 \quad (3.9)$$

$$odziv = \frac{broj\_preklapajućih\_tokena}{duljina\_referentnog\_prijevoda\_u\_tokenima} = \frac{8}{13} = 0.62 \quad (3.10)$$

$$F\text{-}mjera = \frac{preciznost * odziv}{\left(\frac{preciznost + odziv}{2}\right)} = \frac{0.42}{0.65} = 0.65 \quad (3.11)$$

BLEU je danas standardna metrika za evaluaciju kvalitete strojnog prijevoda, a oslanja se na preklapanje, tj. podudarnost n-grama između strojnog prijevoda i jednog ili više referentnih prijevoda te je orijentirana prema preciznosti (Coughlin, 2003; Doddington, 2002; Papineni et al.,

2002). Metrika BLEU temelji se na modificiranoj n-gramskoj preciznosti $\boldsymbol{P_n}$ koja uzima u obzir strojni prijevod i referentni prijevod te se računa za n-grame duljine od 1 do 4 ($\boldsymbol{N}$), najčešće 4 (Papineni et al., 2002). Primjerice, BLEU-3 metrika je BLEU metrika s redom n-grama 3.

Modificirana n-gramska preciznost ($\boldsymbol{MNP}$) računa se prema jednadžbama (3.12 i 3.13), a služi za izbjegavanje zbrajanja točnih, tj. preklapajućih n-grama ($\boldsymbol{zbroj_{ograničen}}$, eng. *clipping*) koji se pojave više puta u strojnom prijevodu nego što se pojave u referentnom prijevodu (Jurafsky i Martin, 2013; Koehn, 2010).

$$zbroj_{ograničen} = \min(zbroj_{strojni\_prijevod}, maksimalni\_zbroj_{referetni\_prijevod}) \tag{3.12}$$

$$MNP = \frac{\sum_{c \in \{strojni\ prijevodi\}} \sum_{ngram \in c} zbroj_{ograničen}(ngram)}{\sum_{c \in \{strojni\ prijevodi\}} \sum_{ngram \in c} zbroj(ngram)} \tag{3.13}$$

Primjer izračuna unigramske preciznosti, $\boldsymbol{preciznost_1}$, dan je u nastavku (3.14), a analogno tome računaju se preciznosti i za bigrame, trigrame i 4-grame:

strojni prijevod: čaj čaj čaj čaj čaj

referentni prijevod: Pijemo čaj jer čaj volimo

$$zbroj_{ograničen}(čaj) = 2$$

$$klasična\ preciznost\ = \frac{5}{5} = 1 \tag{3.14}$$

$$modificirana\ unigramska\ preciznost, preciznost_1 = \frac{2}{5} = 0.4$$

Nakon izračuna svih n-gramskih preciznosti, računa se geometrijska sredina n-gramske preciznosti (jednadžba 3.15) (Tiedemann et al., 2014).

$$P_n = \sqrt[n]{preciznost_1 * preciznost_2 * \dots * preciznost_n}$$
$$= (preciznost_1 * preciznost_2 * \dots * preciznost_n)^{\frac{1}{n}}$$
$$= \left(\prod_{i=1}^{n} preciznost_i\right)^{\frac{1}{n}} = \exp\left(\frac{1}{n}\sum_{i=1}^{n} log_e(preciznost_n)\right) \quad (3.15)$$

BLEU se temelji na geometrijskoj sredini svih modificiranih n-gramskih preciznosti $\boldsymbol{P_n}$, pri čemu $\boldsymbol{w_n}$ predstavlja red n-grama (3.16), a pored toga uključuje i kaznu za kratkoću prijevoda, što je razvidno iz jednadžbe (3.17) (Simard i Fujita, 2012).

$$w_n = \frac{1}{N} \quad (3.16)$$

$$BLEU = kazna\_za\_kratkoću * \exp\left(w_n \sum_{n=1}^{N} log_e(preciznost_n)\right) \quad (3.17)$$

Međutim, kako se radi o geometrijskoj sredini, BLEU je jednako osjetljiv na proporcionalne razlike u pogodcima za sve vrijednosti $\boldsymbol{N}$. Tj. malen broj pogodaka za velik $\boldsymbol{N}$ znatno utječe na konačan rezultat BLEU-a (Specia, 2010).

BLEU kažnjava strojne prijevode koji su prekratki u odnosu na referenti prijevod. Kazna za kratkoću (eng. *brevity penalty*) definirana je prema (3.18), pri čemu $\boldsymbol{c}$ predstavlja duljinu strojnog prijevoda (eng. *candidate*), a $\boldsymbol{r}$ duljinu referentnog prijevoda (eng. *reference*).

Ukoliko se za svaki strojni prijevod primjenjuje više referentnih prijevoda, $\boldsymbol{r}$ se računa kao suma duljina najbližeg referentnog prijevoda u odnosu na svaki strojni prijevod (Specia, 2010).

BLEU favorizira strojne prijevode čiji je odabir riječi sličan odabiru riječi u referentnom prijevodu te čiji je poredak riječi u strojnom prijevodu sličan poretku riječi u referentnom prijevodu (Jurafsky i Martin, 2013; Specia, 2010).

$$kazna\_za\_kratkoću = \begin{cases} 1 & ako\ c > r \\ e^{1-\frac{r}{c}} & ako\ c \leq r \end{cases} \tag{3.18}$$

Ukoliko se objedine prethodne jednadžbe, dobije se (3.19).

$$BLEU = \min\left(1, \exp\left(1 - \frac{\mathrm{r}}{\mathrm{c}}\right)\right) \exp\left(\frac{1}{N}\sum_{n=1}^{N} log_e(preciznost_n)\right) \tag{3.19}$$

Na primjeru sljedeće rečenice pokazat će se izračun BLEU-a (Tablica 15), pri čemu su dijelovi rečenice koji se preklapaju u strojnom i referentnom prijevodu podcrtani:

strojni prijevod: Kuhamo večeru i pripremamo kolač jer nam parents u goste dolaziti .

referentni prijevod: Kuhamo večeru i pripremamo puno kolača jer nam roditelji dolaze u goste .

Tablica 15. Primjer izračun BLEU-a.

	$\boldsymbol{p_n}$	$\boldsymbol{log_e p_n}$
1-*gramska preciznost*	$\frac{9}{12} = 0.75$	-0.288
2-*gramska preciznost*	$\frac{5}{11} = 0.455$	-0.787
3-*gramska preciznost*	$\frac{2}{10} = 0.2$	-1.609
4-*gramska preciznost*	$\frac{1}{9} = 0.111$	-2.198
$\boldsymbol{P_n}$	-4.882	
$\boldsymbol{w_n}$	0.25	
kazna_za_kratkoću	0.92	
BLEU	**0.27**	

BLEU je **0** ako je jedna od n-gramskih preciznosti **0**, što znači da se niti jedan n-gram određene duljine u strojnom prijevodu ne preklapa s n-gramima u referentnom prijevodu. S obzirom da je BLEU vrijednost **0** česta pojava pri izračunu za 4-grame, BLEU se prilikom

evaluacije kvalitete strojnog prijevoda u pravilu računa na razini cijelog korpusa, a ne pojedinačne rečenice, osim u slučaju određenih tehnika adaptacije domene (npr. modifikacija postupka ugađanja sustava za vrijeme treniranja s minimalnom stopom pogreške).

Ipak, BLEU na razini rečenica (segmenata) može se koristiti pri detaljnijoj evaluaciji kvalitete pojedinih strojnih prijevoda ili pri određivanju intervala pouzdanosti (Koehn, 2004c).

Svakako treba naglasiti, da aritmetička sredina BLEU vrijednosti na razini rečenica ne odgovara ukupnoj BLEU vrijednosti na razini cijelog testnog skupa. Naime, pri izračunu BLEU-a na razini cijelog testnog skupa istovremeno se uzimaju sve rečenice i računaju preklapanja n-grama preko svih rečenica (Madnani, 2011). Iz toga proizlazi da BLEU na razini korpusa ne odgovara aritmetičkoj sredini BLEU vrijednosti na razini rečenica. No, postoje metode izglađivanja BLEU vrijednosti na razini rečenica koje nastoje aproksimirati ukupnu BLEU vrijednost na razini korpusa (Chen i Cherry, 2014).

BLEU vrijednosti kreću se u rasponu od $\mathbf{0}$ (nema preklapanja riječi u strojnom prijevodu s riječima u referentnom prijevodu) do $\mathbf{1}$ (savršeno podudaranje s referentnim prijevodom, tj. strojni i referenti prijevod su identični), pri čemu vrijednosti veće od $\mathbf{0.3}$ u pravilu odražavaju razumljive strojne prijevode, a BLEU rezultat veći od $\mathbf{0.5}$ upućuje na dobre i tečne strojne prijevode (Lavie et al., 2010).

BLEU (na razini korpusa, ne rečenice, tj. segmenta) pozitivno korelira s ljudskom evaluacijom kvalitete strojnog prijevoda (Ye et al., 2007) i može razlikovati sustave za strojno prevođenje koji se temelje na istom pristupu strojnom prevođenju.

Empirijski je utvrđeno da veća vrijednost BLEU metrike upućuje na kvalitetniji strojni prijevod, a time i na bolji sustav za strojno prevođenje (Koehn, 2006). Jednako tako, strojni prijevod s niskom vrijednošću BLEU-a bit će u pravilu lošije ocijenjen od strane ljudskog evaluatora.

Međutim, za pouzdan izračun BLEU-a potreban je veći broj referentnih prijevoda za istu rečenicu, pri čemu se onda promatra podudarnost n-grama u svim referentnim prijevodima. Pored toga, predlaže se primjena većeg podatkovnog skupa za ispitivanje kvalitete strojnog prijevoda (eng. *test set*) sa što heterogenijim, tj. reprezentativnijim referentnim prijevodima.

Nedostatak BLEU metrike jest što sve riječi imaju jednaku težinu, tj. ispuštanje sadržajne riječi (koja je važna za semantiku) jednako se penalizira kao i ispuštanje funkcijske riječi (Stymne, 2008; Callison-Burch et al., 2006). Nadalje, BLEU je više namijenjen evaluaciji istog sustava za strojno prevođenje nakon uvođenja određenih modifikacija ili malenih izmjena, poput adaptiranja

određenih parametara, nego usporedbi različitih arhitektura sustava za strojno prevođenje (Stymne, 2008; Callison-Burch et al., 2006).

Metrika **NIST** je inačica BLEU metrike. BLEU metrika pomoću n-gramske preciznosti svakoj riječi pridaje jednaku težinu (eng. *uniform weights*). NIST pak umjesto frekvencije pojavljivanja n-grama računa težinu informativnosti za svaku riječ, tj. veću težinu pridaje rjeđim n-gramima koji se smatraju informativnijima (Adeyanju, 2010). NIST koristi aritmetičku sredinu frekvencija n-grama umjesto geometrijske sredine (Doddington, 2002) te informativniji n-grami dobivaju veću težinu (Adeyanju, 2010).

NIST-ova kazna za kratkoću se također razlikuje u odnosu na BLEU-ovu kaznu za kratkoću, tj. radi se o tzv. modificiranoj kazni za kratkoću (eng. *modified brevity penalty*): malene razlike u duljini prijevoda ne utječu na ukupan rezultat NIST metrike (Specia, 2010), s obzirom da pri ljudskoj evaluaciji kvalitete strojnog prijevoda takve neznatne razlike ne utječu bitno na ljudsku prosudbu kvalitete (Adeyanju, 2010). Korjenovanje riječi može poboljšati rezultate NIST (i BLEU) metrike (Lavie et al., 2004).

Modificirana kazna za kratkoću (***MKK***) dana je u (3.20) (Adeyanju, 2010).

$$MKK = \begin{cases} 1 & ako\ L_{strojni} > \bar{L}_{referentni} \\ exp\left\{\beta log^2\left[min\left(\frac{L_{strojni}}{\bar{L}_{referentni}}, 1\right)\right]\right\} & ako\ L_{strojni} \leq \bar{L}_{referentni} \end{cases} \quad (3.20)$$

Težina informativnosti n-grama $\boldsymbol{Info(w_1 \dots w_n) = I}$ računa se kao u (3.21), i to s obzirom na zbroj n-grama preko cijelog skupa referentnih prijevoda (Specia, 2010) .

$$I = \log_2\left(\frac{zbroj\_pojavljivanja\ (w_1, \dots, w_{n-1})\ u\ referetnim\_prijevodima}{zbroj\_pojavljivanja\ (w_1, \dots, w_n)\ u\ referetnim\_prijevodima}\right) \quad (3.21)$$

Na temelju težina informativnosti n-grama NIST se računa prema (3.22), pri čemu $\boldsymbol{w}$ predstavlja riječ u strojnom prijevodu, $\boldsymbol{N}$ maksimalnu duljinu n-grama, npr. $\mathbf{5}$ (tj. $\boldsymbol{n = 1 \dots N}$), $\boldsymbol{L_{strojni}}$ broj riječi (duljinu) u strojnom prijevodu, a $\boldsymbol{\bar{L}_{referentni}}$ prosječan broj riječi u

referentnim prijevodima. Faktor kazne za kratkoću $\boldsymbol{\beta}$ odabran je tako da kazna za kratkoću $\boldsymbol{MKK}$ bude $\mathbf{0.5}$ kada je $\frac{\boldsymbol{L_{strojni}}}{\boldsymbol{\bar{L}_{referentni}}} = \frac{\mathbf{2}}{\mathbf{3}}$ (Adeyanju, 2010; Doddington, 2002).

$$NIST = MKK * \sum_{n=1}^{N} \left\{ \frac{\sum_{\forall\, n-gram\, w_1 \ldots w_n \in\, strojni\_prijevod} Info(w_1 \ldots w_n)}{broj_{n-grama\, w_1 \ldots w_n\, u\, strojnom\, prijevodu}} \right\} \qquad (3.22)$$

NIST može biti $\mathbf{0}$ ili veći od $\mathbf{0}$ (nema fiksnog maksimuma). Veći NIST (ili BLEU) rezultat upućuje na bolju korelaciju strojnog prijevoda s ljudskom prosudbom kvalitete strojnog prijevoda (Adeyanju, 2010).

Općenito prevladava mišljenje da su automatske metrike bolje što je veća korelacija s ljudskom evaluacijom (Seljan et al., 2012; Brkić et al. 2011; Seljan et al., 2011; Koehn, 2010). Međutim, osnovni problem NIST-a (kao i BLEU metrike) jest što je korelacija s ljudskom evaluacijom relativno slaba na razini rečenice, tj. segmenta (Federmann, 2011).

Metrika **METEOR** evaluira prijevod pomoću izračuna eksplicitnog podudaranja riječi (eng. *explicit word-to-word matches*) između strojnog i referentnog prijevoda (Agarwal i Lavie, 2008). Ukoliko je na raspolaganju više od jednog referentnog prijevoda, strojni prijevod se evaluira prema svakom referentnom prijevodu zasebno te se zatim odabire najbolje bodovani (eng. *best scored*) par (Agarwal i Lavie, 2008).

METEOR balansira odnos odziva i preciznosti unigramskih preklapanja (Specia, 2010). Ipak, odzivu je dana veća težina (Lavie i Agarwal, 2007). Ova metrika može uključiti morfološke informacije, tj. lingvističko znanje (eng. *linguistic knowledge*), što je tada čini ovisnom o jeziku (Banerjee i Lavie, 2005), ali i pogodnom metrikom za morfološki bogate jezike (Denkowski i Lavie, 2011). METEOR podržava i parafraziranje, korjenovanje riječi te pretraživanje sinonima pomoću leksičko-semantičke mreže WordNet (Denkowski i Lavie, 2014). METEOR (kao i GTM) favorizira dulja preklapanja riječi s ispravnim poretkom.

Osnovne komponente metrike METEOR su parametrizirana harmonijska sredina ($\boldsymbol{F_{harmonijska\_sredina}}$) i fragmentacijska kazna $\boldsymbol{PEN}$, koja umanjuje rezultate izračuna harmonijske sredine (Agarwal i Lavie, 2008). METEOR vrši sravnjivanje riječi iz strojnog i referentnog prijevoda prema tipu sravnjivanja **1-1**. Tj. svaka riječ rečenice strojnog prijevoda inkrementalno se mapira u riječ referentnog prijevoda (Agarwal i Lavie, 2008).

Na temelju broja mapiranih unigrama ($\boldsymbol{m}$) koji se podudaraju između dva segmenta, ukupnog broj unigrama u strojnom prijevodu ($\boldsymbol{t}$) i ukupnog broja unigrama u referentnom prijevodu ($\boldsymbol{r}$), moguće je izračunati unigramsku preciznost $\boldsymbol{P} = \frac{\boldsymbol{m}}{\boldsymbol{t}}$ i unigramski odziv $\boldsymbol{R} = \frac{\boldsymbol{m}}{\boldsymbol{r}}$ (Agarwal i Lavie, 2008). Zatim se računa parametrizirana harmonijska sredina $\boldsymbol{F_{harmonijska\_sredina}}$ (eng. *parameterised harmonic mean*) od $\boldsymbol{P}$ i $\boldsymbol{R}$, kao što je prikazano u (3.23). Preciznost, odziv i harmonijska sredina temelje se isključivo na podudaranju unigrama. Parametar $\boldsymbol{\alpha}$ određuje relativnu težinu preciznosti i odziva.

$$F_{harmonijska\_sredina} = \frac{P * R}{\alpha * P + (1 - \alpha) * R} \tag{3.23}$$

Kako bi se u obzir uzela razina ispravnog (tj. identičnog) redoslijeda riječi za unigrame koji se podudaraju u strojnom i referentnom prijevodu, METEOR računa kaznu za danu sravnjenost riječi (Agarwal i Lavie, 2008; Specia, 2010): niz unigrama koji se podudaraju između strojnog i referentnog prijevoda dijeli se u najmanji mogući broj (najduljih) djelića rečenice (eng. *chunks*) tako da podudarajući unigrami, tj. riječi u djeliću rečenice (*chunku*), budu susjedni (eng. *adjacent*) u oba prijevoda (i u strojnom i u referentnom) i s ispravnim poretkom riječi.

Broj *chunkova* ($\boldsymbol{ch}$) i ukupan broj podudaranja riječi u svim *chunkovima* ($\boldsymbol{m}$) se zatim koriste za izračun fragmentacijskog odsječka (eng. *fragmentation fraction*), tj. $\boldsymbol{frag} = \frac{\boldsymbol{ch}}{\boldsymbol{m}}$. Sama kazna računa se kao u (3.24).

$$PEN = \gamma * (frag)^{\beta} \tag{3.24}$$

Vrijednost $\boldsymbol{\gamma}$ određuje težinu maksimalne kazne ($\boldsymbol{0 \leq \gamma \leq 1}$). Vrijednost $\boldsymbol{\beta}$ određuje odnos fragmentacije i same kazne. Izračun METEOR-a dan je u nastavku (3.25).

$$METEOR = (1 - PEN) * F_{harmonijska\_sredina} \tag{3.25}$$

METEOR se temelji na podudaranju djelića teksta (*chunkova*) te pored težinskih varijabli $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ može uključiti i parametar $\boldsymbol{\delta}$ koji služi za razlikovanje sadržajnih od funkcijskih riječi. Izvorno su vrijednosti parametara bile $\boldsymbol{\alpha = 0.9}$, $\boldsymbol{\beta = 3.0}$, $\boldsymbol{\gamma = 0.5}$ (eksperimentalno utvrđene), međutim, danas se mogu ugađati (Denkowski i Lavie, 2014) kako bi se postigla maksimalna korelacija s ljudskom prosudbom kvalitete strojnog prijevoda (Lavie i Agarwal, 2007) ili optimalan omjer adekvatnosti i fluentnosti (Stymne, 2008), sve ovisno o jezičnom paru, podatkovnim skupovima, razini evaluacije (rečenica, dokument, sustav) (Specia, 2010).

METEOR evaluira strojni prijevod sravnjujući ga u odnosu na referenti prijevod te računa njihovu međusobnu sličnost na razini rečenice (Denkowski i Lavie, 2014). Za svaki par strojnog i referentnog prijevoda, prostor pretraživanja se izgrađuje identificiranjem svih mogućih podudaranja između strojnog i referentnog prijevoda primjenom različitih algoritama za analizu dijelova koji se podudaraju (eng. *matcher*). Drugim riječima, različiti parametri METEOR-a dozvoljavaju različite tipove podudaranja unigrama, stoga se koriste i zasebni moduli za (Denkowski i Lavie, 2014; Specia, 2010):

- potpuno podudaranje riječi,
    - mapiranje riječi samo ukoliko su njihovi oblici riječi (eng. *surface forms*) identični
- generalizaciju prema korijenu riječi,
    - korjenovanje riječi pomoću jezično-odgovarajućeg *stemmera* te zatim mapiranje riječi samo ukoliko su korijeni riječi identični
- podudaranje sinonima,
    - mapiranje riječi ukoliko su pridružene jednom sinskupu, tj. skupu sinonima u bazi WordNet
- parafraziranje
    - mapiranje fraza samo ukoliko su navedene kao parafraze u tablici parafraza koje dozvoljavaju mapiranje tipa $\boldsymbol{n}$-$\boldsymbol{n}$ (više prema više)

Primjerice, modul potpunog, tj. egzaktnog podudaranja riječi mapira dvije riječi (jednu iz strojnog te drugu iz referentnog prijevoda) samo ukoliko su identične. Modul generalizacije prema korijenu riječi mapira dvije riječi samo ukoliko su identične nakon izvođenja korjenovanja

riječi (eng. *stemming*). Modul koji koristi podudaranje sinonima može mapirati dvije riječi ukoliko se radi o sinonimima, tj. ukoliko riječi pripadaju istom sinskupu u WordNet leksičko-semantičkoj mreži. Modul koji se oslanja na parafraziranje mapira dvije fraze samo ukoliko su navedene kao parafraze u pripadajućoj tablici parafraza.

Prema tome, metrika METEOR danas ima nekoliko varijacija, a pretežno se razlikuju prema algoritmu identificiranja zajedničkih dijelova (eng. *matching algorithm*) rečenica koji se preklapaju, uzimajući u obzir samo potpuno identične dijelove (eng. *exact match*), djelomične pogotke korijena riječi (eng. *stem match*) ili djelomične pogotke sinonima (eng. *synonym match*) (Giménez i Gonzàlez, 2011; Giménez i Màrquez, 2010).

METEOR je pogodan za mjerenje kvalitete na razini segmenta, odnosno rečenice (Agarwal i Lavie, 2008). Rezultati METEOR metrike kreću se u rasponu od $\mathbf{0}$ do $\mathbf{1}$, a uobičajeno su veći od BLEU-a i odražavaju razumljive prijevode kada su veći od $\mathbf{0.5}$, a dobre i fluentne prijevode kada su veći od $\mathbf{0.7}$ (Lavie, 2010).

Treba ponoviti, da METEOR uključuje tzv. izračun fragmentacije (eng. *fragmentation score*) koji uzima u obzir redoslijed, tj. poredak riječi, proširuje preklapanja tokena uzevši u obzir korjenovanje riječi te pretraživanje sinonima, a pored toga dozvoljava ugađanje težina svojih komponentni kako bi se poboljšala korelacija s ljudskom evaluacijom kvalitete (Specia, 2010).

Pri usporedbi dvaju sustava za strojno prevođenje koriste se razne metode. Međutim, ukoliko je jedan sustav postigao bolje rezultate metrike, to ne znači nužno da se zaista radi o boljem sustavu. Možda su oba sustava jednako dobra, a razlika u rezultatima metrika je slučajna. Takva tvrdnja naziva se nul-hipoteza, za razliku od alternativne hipoteze koja tvrdi da je jedan sustav zaista i bolji od drugog (Koehn, 2010).

Testiranje hipoteze (eng. *hypothesis testing*) je postupak odlučivanja koja je hipoteza istinita, tj. treba provjeriti da li je razlika u rezultatima statistički značajna. Ukoliko postoji vjerojatnost manja od $\mathbf{1\%}$ da je razlika u rezultatima metrika slučajna, kaže se da su sustavi različiti s $\mathbf{99\%}$-$\boldsymbol{tnom}$ statističkom značajnosti, što se označava i s p-vrijednošću $\boldsymbol{p < 0.01}$.

Ako se za statističku značajnost postavi jedan prag (npr. 95%) može se ispitati raspon rezultata metrika. Taj raspon s udaljenošću $\boldsymbol{d}$ od vrijednosti rezultata $\boldsymbol{\bar{x}}$ naziva se interval pouzdanosti (eng. *confidence interval*), a opisan je kao u (3.26). Interval pouzdanosti predstavlja raspon vrijednosti koje su konzistentne s podatcima i za koje se vjeruje da obuhvaćaju „pravu“, tj. stvarnu (eng. *true*) vrijednost s visokom vjerojatnošću, primjerice 95% (Zhang et al., 2004). Širi intervali upućuju na

nižu preciznost, dok uži intervali na veću preciznost. Raspon vrijednosti računa se na uzorku podataka (Zhang et al., 2004).

$$[\bar{x} - d, \bar{x} + d] \tag{3.26}$$

Ukoliko se uspoređuju dva sustava, potrebno je izračunati njihove intervali pouzdanosti, a ako se dva intervala pouzdanosti razlikuju, zaista postoji i razlika među sustavima na zadanoj razini statističke značajnosti (Koehn, 2010).

Još jedna metoda za određivanje statističke značajnosti razlike između sustava za strojno prevođenje jest metoda ponovnog uzorkovanja (eng. *bootstrap resampling*) (Søgaard et al., 2014; Koehn, 2010; Koehn, 2004c; Efron i Tibshirani, 1986). Radi se o samodopunjujućoj metodi koja generira intervale pouzdanosti koji se zatim mogu međusobno uspoređivati.

To je neparametarska statistička metoda za procjenu pouzdanosti i zahtijeva rezultate metrika na razini rečenica (ne cijelog korpusa). Ideja metode ponovnog uzorkovanja je iz jednog inicijalnog testnog podatkovnog skupa strojnih prijevoda generirati velik broj različitih testnih korpusa koji se sastoje od nasumičnih rečenica iz istog skupa rečenica. Rečenice, tj. segmenti se pritom mogu pojaviti nula, jednom ili više puta, a smiju se i ponavljati u različitim korpusima, (eng. *sampling with replacement*). Zatim se može primijeniti metrika na svakom od korpusa te se nakon toga mogu odrediti intervali pouzdanosti (Zhang i Vogel, 2004).

Međutim, jednostavnije je prvo odrediti vrijednosti metrika dobivenih evaluacijom sustava, koje se zatim primjenjuju u metodi ponovnog uzorkovanja radi generiranje velikog broj umjetnih skupova te definiranja njihovih intervala pouzdanosti. Ukoliko je jedan sustav za strojno prevođenje bolji od drugog u najmanje 95% testnih korpusa (uzoraka), tada je on i statistički značajno bolji na razini $\boldsymbol{p \leq 0.05}$ (Koehn, 2010).

Treba isto naglasiti da su rezultati BLEU-a primijenjeni na ljudskim prijevodima u pravilu dosta loši zbog veće varijabilnost riječi, subjektivne motivacije ljudskih prevoditelja za odabirom riječi itd. Pored toga, leksička sličnost dviju riječi nije nužan niti dovoljan uvjet za procjenu količine prenesenoga značenja. Stoga i ne čudi što automatske metrike postižu bolje rezultate ukoliko podatkovni skupovi za treniranje i testiranje pripadaju istoj ili bliskoj domeni koja obuhvaća specifičnu terminologiju i kontekst.

Jedna od najvažnijih kritika upućena automatskim metrikama jest što u pravilu ne uzimaju u obzir relevantnost riječi, iako su primjerice imena, glagoli ili imenice semantički vrjednije od članova ili interpunkcijskih znakova u rečenici. Pored toga, mnoštvo metrika poput BLEU-a ne razlikuje različite tipove pogrešaka (ACCEPT, 2012). Npr. sljedeći n-grami se potpuno jednako tretiraju: ***dao sam joj***, ***dao sam njoj***, ***dao sam woman*** (Way i Hassan, 2009).

Nadalje, interpretacija rezultata automatskih metrika također nije uvijek jednostavna. Naime, treba uzeti u obzir i veličinu testnog korpusa, kvalitetu korpusa, određene specifičnosti korpusa, broj referentnih ljudskih prijevoda po strojnom prijevodu, jezični par, domenu, radnje pri pripremi testnog korpusa (npr. segmentacija, sravnjivanje, tokenizacija itd.) (eng. *preprocessing*).

S obzirom da automatske metrike koje se baziraju na leksičkoj sličnosti dvaju niza znakova ili podatkovnih skupova ne mogu jamčiti točnu evaluaciju, predložene su i druge strategije za ocjenjivanje kvalitete strojnog prijevoda koje se ne temelje nužno na leksičkoj sličnosti (España-Bonet i Gonzàlez, 2014; Gonzàlez, 2014).

To su strategije koje se fokusiraju na sličnost s obzirom na sintaksu (eng. *syntactic similarity*) ili semantiku (eng. *semantic similarity*). Sintaksna sličnost može se ispitati pomoću lematizacije, razdjeljivanja (eng. *chunking*), sintaksne analize i sintaksnih stabala, POS tagiranja itd. Semantička sličnost može se ispitati provjerom semantičkih uloga riječi u rečenici (eng. *semantic roles*), analizom diskursa (eng. *discourse analysis*) ili analizom imenovanih entiteta (eng. *named entities*), s obzirom da se imena i specifični nazivi organizacija rijetko nalaze u korpusima koji se koriste kao ulazni skupovi za treniranje sustava za strojno prevođenje.

Radi pouzdanije informacije o kvaliteti strojnog prijevoda predlaže se kombinacija više različitih automatskih metrika, s obzirom da različite metrike obuhvaćaju i različite aspekte sličnosti (Gonzàlez, 2014).

Intencija znanstvene zajednice okupljene oko strojnog prevođenja jest definirati jedinstvenu metriku koja je u stanju prepoznati da li se značenje u strojnom prijevodu podudara sa značenjem u referentnom prijevodu. Stoga je razvijen „Multidimensional Quality Metrics“ (MQM), složen model za ocjenjivanje kvalitete strojnog prijevoda, koji se u potpunosti može prilagoditi potrebama korisnika (Burchardt i Lommel, 2014; Lommel et al., 2013). MQM razvijen je u sklopu projekta „QTLaunchPad“, a implementiran je primjerice u alatu translate5 (Lommel et al., 2014b).

Nadalje, istražuju se i mogućnosti automatske procjene kvalitete bez upotrebe referentnih prijevoda (eng. *quality estimation*) (Bojar et al., 2014; Rapp, 2009), već samo primjenom rečenica na izvornome jeziku i strojnih prijevoda te podataka o samom sustavu za strojno prevođenju (Specia

et al., 2009). Takav postupak bi zasigurno ubrzao ugađanje težina pojedinih značajki te usporedbu i rangiranje raznih sustava za strojno prevođenje. Radi se o vrlo kompleksnom poduhvatu, pri čemu se u obzir uzimaju parametri kao što su težine značajki, omjer težina značajki, prosječna duljina strojnog prijevoda, perpleksnost, razina dvoznačnosti identificirana pomoću rječnika ili leksičko-semantičkih mreža, brojnost interpunkcijskih znakova, sravnjenost riječi itd. (Gonzàlez, 2014).

Treba još jednom naglasiti kako interpretacija rezultata automatskih metrika nije uvijek jednostavna. Npr. BLEU vrijednost od 35 ne odgovara zaista na pitanje je li strojni prijevod dobar ili loš. Možda je prijevod samo djelomično točan ili netočan, djelomično fluentan ili nefluentan? Visoka vrijednost metrike stoga može dati lažan osjećaj sigurnosti u kvalitetu strojnog prijevoda. Isto tako treba naglasiti da automatska metrika ne ukazuje izravno na uzroke niskih ili visokih rezultata metrike, već je potrebna detaljna analiza strojnih prijevoda i pogrešaka u strojnim prijevodima te analiza karakteristika sustava za strojno prevođenje.

5. ISTRAŽIVANJE

U narednim poglavljima opisane su metode, postupci i rješenja koja su primijenjena u doktorskom istraživanju. Prvo su definirani ciljevi i hipoteze istraživanja te znanstveni doprinos. Zatim su opisani metodologija i tijek istraživanja, a nakon toga podatkovni skupovi upotrijebljeni u različite svrhe, tj. za treniranje, ugađanje i testiranje (ispitivanje) sustava za statističko strojno prevođenje. Zatim su definirani i eksperimentalno okruženje te korišteni resursi. Rezultati eksperimentalnog dijela istraživanja opisani su u posebnom poglavlju, a nakon toga prikazani su rezultati evaluacije kvalitete strojnih prijevoda primjenom različitih pristupa te pripadajuća analiza rezultata.

Statističko strojno prevođenje propulzivan je segment informacijskih tehnologija sa sve širom uporabom u industriji prevođenja, poput lokalizacije koja predstavlja jedno od najbrže rastućih grana u području jezičnih tehnologija. Istraživanja na području strojnog prevođenja u Hrvatskoj nisu provedena u dostatnoj mjeri, unatoč realnoj potrebi za sve bržim prevođenjem sa i na hrvatski jezik. Utjecaj računalne adaptacije (prilagodbe) domene, karakteristika ulaznih podatkovnih skupova te modifikacija značajki modela sustava za statističko strojno prevođenje temeljeno na frazama na kvalitetu strojnog prijevoda od posebnog su interesa znanstvene zajednice i šire jezične industrije.

5.1. Cilj i hipoteze istraživanja

Cilj ovog doktorskog istraživanja jest ustanoviti utjecaj adaptacije domene te obilježja ulaznoga podatkovnog skupa i modifikacija značajki modela sustava za statističko strojno prevođenje temeljeno na frazama na kvalitetu strojnog prijevoda za hrvatski jezik. Radom su analizirane postojeće teorije i metodologije izgradnje sustava za statističko strojno prevođenje temeljeno na frazama te istraženi novi pristupi povećanju kvalitete automatskog strojnog prijevoda u općoj domeni i domeni računalnog softvera za englesko-hrvatski i hrvatsko-engleski jezični par.

Ideja ovoga rada bila je ispitati da li se kombinacijom izvandomenskih značajki može, te u kojoj mjeri, utjecati na povećanje kvalitete strojnog prijevoda za hrvatski jezik. Nakon izgradnje 8 različitih sustava za statističko strojno prevođenje hrvatsko-engleskog i englesko-hrvatskog jezičnog para izvršena je komparativna analiza kvalitete prijevoda, s obzirom na, u doktorskom istraživanju izgrađene sustave, te postojeće web servise za statističko strojno prevođenje.

Glavne hipoteze istraživanja navedene su u nastavku (H1, H2 i H3):

H1: adaptacija (prilagodba) domene i karakteristike ulaznoga podatkovnog skupa utječu na kvalitetu strojnog prijevoda, tj. na performanse sustava za statističko strojno prevođenje temeljeno na frazama za hrvatsko-engleski i englesko-hrvatski jezični par

H2: primjenom hibridnog sustava za statističko strojno prevođenje temeljeno na frazama za hrvatsko-engleski i englesko-hrvatski jezični par moguće je postići poboljšanje kvalitete strojnog prijevoda u odnosu na sustave koji koriste značajke svojstvene isključivo jednoj domeni

H3: izgrađeni sustavi za statističko strojno prevođenje temeljeno na frazama u komparaciji s postojećim web servisima za statističko strojno prevođenje mogu polučiti bolje rezultate za određenu domenu

U ovom doktorskom istraživanju pod pojmom hibridni sustav za statističko strojni prevođenje podrazumijeva se primjena alternativnog puta dekodiranja koja dozvoljava izgradnju skupa mogućih prijevoda pomoću više prijevodnih modela, a ne kombinacija s pristupom temeljenim na pravilima ili implementacija jezičnoga znanja.

Takav hibridni sustav za strojno prevođenje ne sadrži isključivo podmodele svojstvene samo jednoj domeni (tj. ne koristi samo modele trenirane na jednoj domeni) već kombinira prijevodne modele (ali i druge statističke modele) iz drugih, tj. različitih domena.

Skup mogućih prijevoda pretražuje se prvo u prijevodnom modelu treniranom na specifičnom podatkovnom skupu, a tek zatim u prijevodnom modelu treniranom na nekoj drugoj (općenitoj) domeni ukoliko mogući prijevodi nisu pronađeni u prvom, tj. preferiranom modelu.

U tom slučaju, prijevodni model treniran na općenitoj domeni predstavlja *back-off* prijevodni model za riječi i fraze koje nisu viđene u prvoj, tj. preferiranoj tablici prijevoda fraza, tj. fraznih struktura. Ovom metodom računalne adaptacije domene nastoje se poboljšati performanse sustava za statističko strojno prevođenje za hrvatsko-engleski i englesko-hrvatski jezični par.

5.2. Znanstveni doprinos

Doprinos ovog doktorskog rada jest što se navedeni utjecaji ispituju na primjeru strojnog prevođenja za hrvatsko-engleski jezični par. Iz istraživačkih rezultata i analize spomenutih utjecaja proizlazi teorijska podloga i upute „dobre prakse" za izgradnju statističkih modela za hrvatsko-engleski jezični par. Pored toga, doktorski rad doprinosi općem razumijevanju osobina statističkih modela i hibridnog pristupa izgradnji sustava za statističko strojno prevođenje. Također ukazuje na važnost razlikovanja opće i specifične domene, tj. domenskih i izvandomenskih podatkovnih skupova. Uloga težina značajki u težinskom modelu sustava za statističko strojno prevođenje te modifikacija značajki modela također su istraženi na primjeru hrvatskog jezika. Naime, doktorskim radom utvrđeno je da li se kombinacijom izvandomenskih značajki može utjecati na povećanje kvalitete strojnog prijevoda za hrvatski jezik, te u kojoj mjeri.

Nadalje, izgrađeni su sustavi za statističko strojno prevođenje temeljeno na frazama za hrvatsko-engleski jezični par u oba smjera te je, nakon provedbe evaluacije, komparativnom analizom ustanovljena kvaliteta strojnih prijevoda u usporedbi s postojećim web servisima za statističko strojno prevođenje.

Rezultati ovog istraživanja predstavljaju doprinos informacijskim znanostima, budući da se primjenom modela statističkog strojnog prevođenja temeljenog na fraza mogu izgraditi informacijski sustavi koji podupiru dohvaćanje i obradu podataka te upravljanje informacijama.

Nova saznanja koja proizlaze iz ovog doktorskog rada mogu pomoći budućim istraživačima pri izgradnji sustava za statističko strojno prevođenje, nadogradnji te optimizaciji modela za statističko strojno prevođenje temeljeno na frazama.

Takvi sustavi omogućuju automatsko prevođenje odabrane domene, što otvara nove mogućnosti komunikacije i pretraživanja informacija, te suradnje u digitalnom višejezičnom europskom i svjetskom znanstvenom okruženju, a rezultati znanstvenog istraživanja na području strojnog prevođenja mogu se izravno i praktično primijeniti u gospodarstvu, tj. industriji.

5.3. Metodologija i tijek istraživanja

Podatkovni skupovi koji su korišteni za treniranje sustava za statističko strojno prevođenje temeljeno na frazama sastoje se od hrvatsko-engleskih i englesko-hrvatskih paralelnih korpusa. Različite su veličine i pripadaju općoj, odnosno specijaliziranoj domeni (računalni softver). Korpusi su dohvaćeni putem TAUS repozitorija[5]. Korpus opće domene čine filmski podnaslovi (eng. *subtitles*) koji pokrivaju šaroliku terminologiju: od svakodnevnog razgovora, meteorologije, sporta, prava, oružja, hrane i pića, osobnih imena itd. Korpus računalnog softvera čine razne korisničke upute za korištenje softvera, a pokriva terminologiju informacijske tehnologije.

Ovdje treba naglasiti, da podatkovni skup općenite domene nije sadržavao samo rečenice na hrvatskom jeziku, već i u manjoj mjeri rečenice srodnih jezika (bosanski, slovenski i srpski), što je unosilo određenu količinu šuma, s obzirom da se jezici razlikuju prema uporabi pisma (ćirilica, latinica) i dijakritičkih znakova, prema vokabularu, morfologiji itd. Nadalje, općeniti podatkovni skup sadržavao je i neke vrlo loše prijevode, prijevode bez dijakritičkih znakova ili su pak nedostajali dijelovi prijevoda. Time se svakako utjecalo na performanse sustava za statističko strojno prevođenje, međutim, od istraživačkog interesa bilo je ispitati je li moguće takvim sustavom generirati kvalitativno prihvatljive prijevode.

Podatkovni skup korišten za treniranje domenski specifičnih statističkih modela pripadao je području računalnog softvera te se radilo o visokokvalitetnom paralelnom korpusu, sastavljenom od tehničke dokumentacije te korisničkih uputa koje su sadržavale vrlo specifičnu terminologiju, poput elemenata grafičkog sučelja, tipkovnih prečaca, akronima iz informacijske tehnologije, tehnoloških pojmova itd.

Motivacija za uključivanje korpusa različitih domena bila je ispitivanje utjecaja prilagodbe značajki u sustavu za statističko strojno prevođenje koje se koriste prilikom dekodiranja specifičnoj domeni (računalni softver).

Korpus koji pokriva općenitu domenu bio je desetak puta veći od specijaliziranog korpusa iz domene računalnog softvera. Istraživanjima su ispitani utjecaj veličine korpusa, domene, smjera jezičnog para i korištenih značajki modela na kvalitetu strojnog prijevoda.

Podatkovni skup za ugađanje nastao je ekstrakcijom prvih 1000 rečenica, tj. segmenata iz inicijalnog domenski specifičnog podatkovnog skupa. Podatkovni skup za testiranje generiran je

[5] https://www.tausdata.org/index.php/data

ekstrakcijom zadnjih 1000 rečenica iz inicijalnog skupa specifične domene. Sve ekstrahirane rečenice (ili segmenti) su zatim uklonjene iz inicijalnog domenski specifičnog podatkovnog skupa, koji je potom iskorišten za treniranje jezičnih i prijevodnih modela za specifičnu domenu, tj. domenu računalnog softvera.

Svi korišteni paralelni korpusi trebali su biti sravnjeni na razini rečenice, tj. segmenta. Rečenična sravnjenost korpusa osigurana je ručnom provjerom pomoću alata CorAl[6] te primjenom Gale-Church metode sravnjivanja korpusa. Korpusi su kao običan tekst (eng. *plain text*) pohranjeni u UTF-8 kodnom zapisu čime su sačuvani svi znakovi.

Pored navedenog, moguće je korpuse i anotirati pomoću XML-a, koristiti rešetku riječi (eng. *word lattice*) ili *fuzzy* mrežu (eng. *confusion network*), što je posebno korisno ukoliko je ulazni podatkovni skup rezultat procesa prepoznavanja govora (Koehn, 2015). Nakon provjere rečenične sravnjenosti uslijedili su drugi postupci pretprocesiranja ulaznih podatkovnih skupova, poput, tokenizacije, postupka pretvaranja početnog znaka riječi u malo ili veliko slovo te čišćenja korpusa.

U sklopu ovog doktorskog istraživanja, pomoću pretprocesiranih podatkovnih skupova razvijeno je više sustava za statističko strojno prevođenje temeljeno na frazama za dvije domene različitih veličina te karakteristika i za različite jezične smjerove. Istraživanje je provedeno na ukupno 8 sustava, odnosno 4 hrvatsko-engleska i 4 englesko-hrvatska sustava:

1. sustav treniran na korpusu **opće domene**

   (**sustav 1**) *smjer: hrvatsko-engleski*

2. sustav treniran na specijaliziranom korpusu, tj. **domeni računalnog softvera** (**sustav 2**) *smjer: hrvatsko-engleski*

3. hibridni sustav koji koristi tablice prijevoda fraza, tj. fraznih struktura iz **obje domene**, a preostale značajke modela (leksičko premještanje, jezični model itd.) iz **domene računalnog softvera**

   (**sustav 3**) *smjer: hrvatsko-engleski*

4. hibridni sustav koji koristi tablice prijevoda fraza, tj. fraznih struktura iz **obje domene**, a preostale značajke modela (leksičko premještanje, jezični model itd.) iz **opće domene**

[6] http://takelab.fer.hr/coral/

(**sustav 4**) *smjer: hrvatsko-engleski*

5. sustav treniran na korpusu **opće domene**

   (**sustav 5**) *smjer: englesko-hrvatski*

6. sustav treniran na specijaliziranom korpusu **domene računalnog softvera**

   (**sustav 6**) *smjer: englesko-hrvatski*

7. hibridni sustav koji koristi tablice prijevoda fraza, tj. fraznih struktura iz **obje domene**, a preostale značajke modela (leksičko premještanje, jezični model itd.) iz **domene računalnog softvera**

   (**sustav 7**) *smjer: englesko-hrvatski*

8. hibridni sustav koji koristi tablice prijevoda fraza, tj. fraznih struktura iz **obje domene**, a preostale značajke modela (leksičko premještanje, jezični model itd.) iz **opće domene**

   (**sustav 8**) *smjer: englesko-hrvatski*

Postupak izgradnje sustava za statističko strojno prevođenje jednak je za sve sustave koji koriste isključivo domenske značajke, a razlikuju se samo po jezičnom smjeru te namjeni, tj. domeni. Za hibridne sustave nije bilo potrebno trenirati jezične i prijevodne modele. Naime, hibridni sustav koristi resurse koji su na raspolaganju, npr. jezični model treniran na korpusu opće ili specifične domene, model leksičkog preslagivanja/premještanja fraza, tj. distorzije fraza itd. Hibridni sustavi za statističko strojno prevođenje koriste tablice prijevoda fraza, tj. fraznih struktura iz obje domene, tj. iz općenite domene i specifične domene (računalni softver). Na taj način je, putem metode alternativnog puta dekodiranja i primjenom *back-off* modela, moguće iskoristiti prednosti oba prijevodna modela i time utjecati na poboljšanje kvalitete strojnog prijevoda.

U svim hibridnim sustavima favorizirana je domenski specifična tablica prijevoda fraza, tj. fraznih struktura, pri čemu je samo ciljni jezik određivao odabir statističkih modela. Time je u sustavima za statističko strojno prevođenje dana veća težina prijevodnom modelu treniranom na domeni računalnog softvera.

Model povlačenja, tj. *back-off* model je korišten samo za nepoznate unigrame koje u preferiranom (domenski specifičnom) prijevodnom modelu predstavljaju riječi izvan vokabulara. Sve preostale težine značajki preuzimaju se iz sustava treniranih isključivo na jednoj domeni (ili

općenitoj ili specifičnoj): jezični model, model leksičkog preslagivanja/premještanja (za oba smjera), model leksičkih težina za prevođenje riječi (za oba smjera), model penaliziranja riječi, model penaliziranja fraza itd. Stoga hibridni sustavi u ovom doktorskom istraživanju predstavljaju pristup računalnoj adaptaciji domene. Primjenom hibridnog sustava nastojalo se ispitati poboljšanje kvalitete strojnog prijevoda unutar jednog sustava za strojno prevođenje adaptacijom, tj. prilagodnom vrijednosti težina značajki.

Za učenje (treniranje), pohranjivanje, analizu i pristup n-gramskim jezičnim modelima korišten je IRST Language Modeling Toolkit (IRSTLM) i pripadajući programski alati, koji omogućuju izglađivanje modela modificiranom inačicom Kneser-Ney metode izglađivanja, dodavanje graničnih oznaka rečenicama (eng. *sentence boundary symbols*), binarizaciju jezičnog modela kao jedne vrste kompresije itd. (Federico et al., 2008). Radi se o besplatnom i *open-source* alatu razvijenom na Fondazione Bruno Kessler (FBK) u Trentu (Italija). Jezični model osigurava tečnost strojnog prijevoda, stoga se i izgrađuje za ciljni jezik.

Alat GIZA++ upotrijebljen je za sravnjivanje riječi na razini fraza. Za pretprocesiranje skupova, ekstrakciju fraza (eng. *phrase extraction*) te izračun vjerojatnosti za svaki frazni par (eng. *scoring*), generiranje tablica prijevoda fraza i leksičkog premještanja fraza (eng. *lexicalised reordering tables*), ugađanje *(eng. tuning*) sustava pomoću učenja s minimalnom stopom pogreške (MERT) te dekodiranje, tj. statističko strojno prevođenje temeljeno na frazama koristit će se Moses 1.0 Toolkit[7], pisan u programskom jeziku C++.

Ispitat će se značaj ugađanja sustava kojim se podešavaju težinske vrijednosti raznih značajki modela sustava za statističko strojno prevođenje (leksičko premještanje, jezični model, prijevodni model itd.).

Dekoder pomoću treniranih statistički modela s ugođenim težinama prevodi izvornu rečenicu u ciljni jezik (Koehn, 2015), pri čemu vrlo važnu ulogu ima jezični model kojim se nastoji osigurati fluentan prijevod u ciljnom jeziku.

Za svaki od 8 izgrađenih sustava provedena je automatska evaluacija strojnog prijevoda primjenom metrika: WER, TER, BLEU, NIST, METEOR i GTM te pripadajućih programskih skripti za izračun i alata Asiya[8] (Giménez i Gonzàlez, 2011; Giménez i Màrquez, 2010). Navedene metrike primijenjene su u postupku evaluacije istog podatkovnog skupa korištenog za testiranje (ispitivanje) kvalitete strojnog prijevoda svih izgrađenih sustava.

[7] https://github.com/moses-smt/mosesdecoder
[8] http://asiya.lsi.upc.edu/

S obzirom da se BLEU metrika koristi za usporedbu sličnih sustava ili različitih inačica istog sustava, dobiveni rezultati za vlastite sustave usporedit će se s rezultatima dobivenih primjenom web servisa za automatsko strojno prevođenje za iste jezične parove: Google Translate (https://translate.google.com/) i Yandex Translate (https://translate.yandex.com/). Isto će se učiniti i za preostale evaluacijske metrike.

Na taj način su prema kvaliteti strojnog prijevoda rangirani vlastiti trenirani sustavi koji se međusobno razlikuju prema vrsti domene treniranog sustava, jezičnom smjeru itd., i to usporedno s postojećim web servisima za automatsko strojno prevođenje. Ljudsku evaluaciju proveli su evaluatori koji su izvorni govornici hrvatskog jezika te tečni govornici engleskog jezika, prema standardnim kriterijima evaluacije strojnih prijevoda, tj. koristeći Likertovu skalu s obzirom na kriterije adekvatnosti/točnosti i fluentnosti/tečnosti. Stupanj međusobne složnosti ljudskih evaluatora određen je mjerom Cronbach alpha.

Za mjerenje korelacije između rezultata automatskih metrika i ljudske evaluacije korišten je Pearsonov koeficijent korelacije. Radom je istraženo i koja vrsta pogreške u strojnom prijevodu najviše utječe na ljudsku percepciju kvalitete strojnog prijevoda.

Za statističku obradu i analizu rezultata istraživanja upotrijebljeni su programski jezik R[9] i pripadajuće okruženje RStudio[10] te Microsoft Excel.

Za provođenje predloženih istraživanja primijenjene su različite znanstvene metode: kvalitativna analiza znanstvenog dosega i metoda na području statističkog strojnog prevođenja, eksperiment, promatranje, mjerenje i kvantitativna analiza rezultata, komparativna analiza rezultata, induktivno i deduktivno zaključivanje te dokazivanje i opovrgavanje hipoteza.

Na provođenje doktorskog istraživanja utrošeno je oko godinu dana: za prikupljanje, konverziju, pretprocesiranje, sravnjivanje i analizu podatkovnih skupova utrošeno je oko mjesec dana. Na inicijalno podešavanje sustava i eksperimentiranje sa statističkim strojnim prevođenjem utrošeno je oko 4 mjeseci, a za razvijanje svih 8 sustava za statističko strojno prevođenje temeljeno na frazama bila su potrebna dodatna 2 mjeseca. Za eksperimentiranje s faktorima koji utječu na kvalitetu strojnog prijevoda utrošeno je oko 3 mjeseca. Za kvantitativnu i kvalitativnu analizu rezultata i evaluaciju sustava bilo je potrebno oko 2 mjeseca. Za izradu dodatnih pomoćnih programskih alata za pretprocesiranje, ekstrakciju te analizu korištenih podatkovnih skupova za statističko strojno prevođenje utrošeno je oko mjesec dana.

[9] http://www.r-project.org/
[10] http://www.rstudio.com/

5.4. Podatkovni skupovi za treniranje, ugađanje i testiranje

Za izgradnju, tj. treniranje n-gramskih jezičnih modela potreban je podatkovni skup za treniranje jezičnog modela, a kojeg čini jednojezični korpus ciljnog jezika (eng. *monolingual training set*), ovisno o tome na koji jezik sustav prevodi. Za izgradnju, tj. treniranje prijevodnih modela potreban je rečenično sravnjeni paralelni korpus (eng. *bilingual training set*) koji se sastoji od tekstova na izvornom i ciljnom jeziku, tj. međusobnih prijevoda.

Za analizu podatkovnih skupova za treniranje, ugađanje i testiranje (ispitivanje) korišteni su programski jezik Python (Bird et al., 2009) i NLTK 3.0 (Natural Language Toolkit)[11] platforma te vlastite Perl skripte.

U nastavku (Tablice 16 i 17) prikazane su karakteristike podatkovnih skupova (prije pretprocesiranja) iskorištenih za treniranje jezičnih modela iz općenite domene (eng. *general domain*). Pod pojmom „segment" podrazumijevaju se fragmenti rečenice koji se prevode zasebno, tj. koji se u korpusu nalaze u zasebnom retku, kao npr. niz riječi ***potpuno drugačije***.

Tablica 16. Karakteristike podatkovnih skupova za treniranje jezičnih i prijevodnih modela iz **općenite** domene.

Karakteristike podatkovnog skupa	hrvatski	engleski
broj znakova	8916077	10726358
broj riječi	1633166	2105795
broj različitih riječi (vokabular)	185281	94647
broj rečenica/segmenata	289080	289080
maksimalni broj riječi po rečenici/segmentu	64	220
minimalni broj riječi po rečenici/segmentu	1	1
prosječan broj znakova po rečenici/segmentu	30.84	37.11
prosječan broj riječi po rečenici/segmentu	5.65	7.28
standardna devijacija (broj riječi po rečenici/segmentu)	3.64	4.73

[11] http://www.nltk.org/

Tablica 16 prikazuje karakteristike paralelnog korpusa namijenjenog treniranju prijevodnih modela iz općenite domene. Radi se o korpusu prije provođenja pretprocesiranja. Sastojao se od ukupno cca. 600000 rečenica/segmenata, sastavljenih od 3738961 riječi, tj. cca. 20 milijuna znakova. Uobičajeno je engleski dio paralelnog korpusa bio veći, tj. sastojao se od više riječi. Taj fenomen u engleskom jeziku vjerojatno se može pripisati većoj upotrebi funkcijskih riječi pri sastavljanju rečenica.

Treba svakako uočiti da je razlika u veličini hrvatskog i engleskog vokabulara oko 100%, što ukazuje na bogatstvo jezika i morfoloških varijacija u hrvatskom jeziku. Najdulja rečenica u hrvatskom dijelu paralelnog korpusa sastojala se od najviše 64 riječi, dok u engleskom dijelu od čak 220 riječi. Najkraća rečenica/segment u oba jezika bila je duljine jedan. Da je za formiranje rečenice u hrvatskom dijelu paralelnog korpusa potreban manji broj riječi pokazuje i podatak da je prosječan broj riječi po rečenici/fragmentu iznosio 5.65 (za hrvatski), odnosno 7.28 (za engleski).

To potvrđuje i podatak da je prosječan broj znakova po rečenici/segmentu na hrvatskom jeziku bio 30.84, a u engleskom dijelu korpusa 37.11. Standardna devijacija broja riječi po rečenici, tj. segmentu u hrvatskom dijelu korpusa iznosila je 3.64, dok je u engleskom dijelu bila nešto veća, tj. 4.73. Jednojezični dio tog istog paralelnog korpusa korišten je za treniranje n-gramskog jezičnog modela ciljnog jezika, ovisno o namjeni sustava za statističko strojno prevođenje.

Tablica 17 prikazuje distribuciju riječi u skupovima za treniranje jezičnih i prijevodnih modela iz općenite domene, prema frekvencijama riječi.

Tablica 17. Distribucija riječi u podatkovnim skupovima za treniranje jezičnih i prijevodnih modela iz **općenite domene**, prema frekvencijama riječi.

Frekvencije riječi unutar podatkovnog skupa	**Broj riječi**	**Udio u vokabularu**	**Udio u ukupnom broju riječi**
hrvatski			
1, broj riječi koje se pojavljuju točnom jednom (hapax legomena)	105782	57.09%	6.48%
2, broj riječi koje se pojavljuju točno dva puta (dis legomena)	29353	15.84%	3.59%
3, broj riječi koje se pojavljuju točno tri puta (tris legomena)	13345	7.20%	2.45%
4	7792	4.21%	1.91%
5	5035	2.72%	1.54%

6	3520	1.90%	1.29%
7	2591	1.40%	1.11%
8	2049	1.11%	1.00%
9	1626	0.88%	0.90%
10	1255	0.68%	0.77%
> 10	12933	6.98%	78.96%
engleski			
1, broj riječi koje se pojavljuju točnom jednom (hapax legomena)	45109	47.66%	2.14%
2, broj riječi koje se pojavljuju točno dva puta (dis legomena)	14948	15.79%	1.42%
3, broj riječi koje se pojavljuju točno tri puta (tris legomena)	7582	8.01%	1.08%
4	4775	5.05%	0.91%
5	3244	3.43%	0.77%
6	2350	2.48%	0.67%
7	1735	1.83%	0.58%
8	1452	1.53%	0.55%
9	1131	1.19%	0.48%
10	957	1.01%	0.45%
> 10	11364	12.01%	90.94%

Kada se detaljnije analizira distribucija riječi prema frekvencijama u vokabularu može se uočiti da riječi koje se pojavljuju točnom jednom u hrvatskom dijelu paralelnog korpusa čine 57% svih riječi u vokabularu, međutim, to čini samo cca. 6.5% svih riječi u hrvatskom dijelu paralelnog korpusa (Tablica 17). U engleskom dijelu korpusa, hapax legomena čine cca. 48% vokabulara te cca. samo 2% ukupnog broja engleskih riječi u korpusu. Dis legomena u hrvatskom dijelu korpusa čine oko 16% vokabulara, a tris legomena oko 7%. Trend opadanja udjela u vokabularu i ukupnom broju riječi nastavlja se porastom frekvencija riječi unutar podatkovnog skupa.

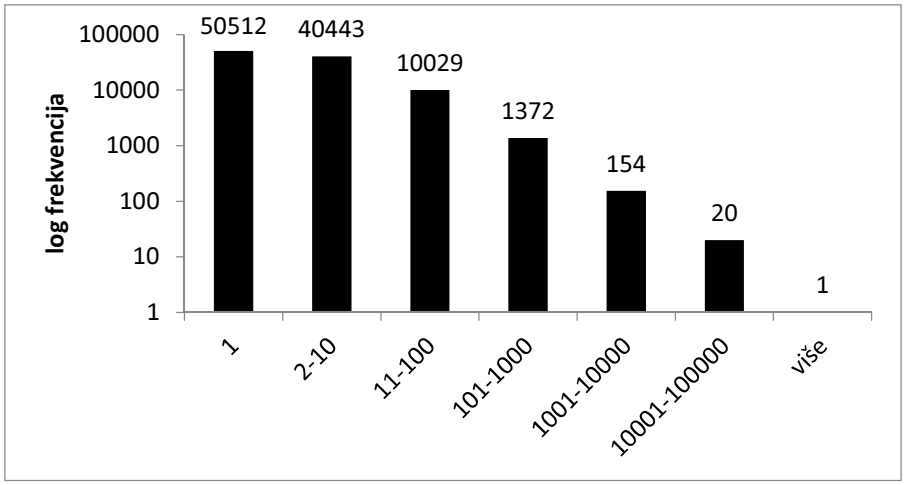
Međutim, u hrvatskom dijelu korpusa, riječi s frekvencijom većom od 10 čine čak 79% ukupnog broja riječi u korpusu, odnosno oko 7% vokabulara. Isti trendovi mogu se uočiti i u engleskom dijelu korpusa, no, u slučaju engleskog jezika riječi s frekvencijom većom od 10 čine čak 91% ukupnog broja riječi u korpusu, tj. 12% vokabulara. To upućuje na vrlo čestu pojavu funkcijskih riječi, poput veznika, članova, zamjenica itd., ali istovremeno i na mogućnost sastavljanja velikog broja (semantički) različitih rečenica pomoću jednakog skupa riječi.

Tablica 18 prikazuje dvadeset najfrekventnijih tokena u podatkovnim skupovima korištenim pri treniranju jezičnih i prijevodnih modela iz općenite domene. Iz tablice očekivano proizlazi da se radi o interpunkcijskim znakovima te funkcijskim riječima.

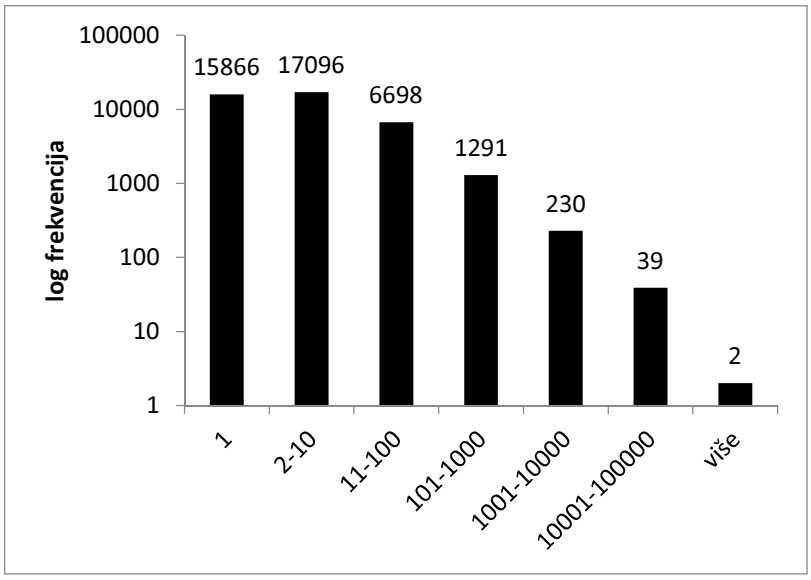
Tablica 18. Dvadeset najfrekventnijih tokena u podatkovnim skupovima korištenim pri treniranju jezičnih i prijevodnih modela iz **općenite domene**.

hrvatski		**engleski**	
token	**frekvencija**	**token**	**frekvencija**
.	223073	.	234311
,	94335	,	128598
je	55686	you	89875
?	54256	I	83808
da	41437	the	62219
-	35724	-	58229
se	32671	?	57423
i	28112	to	51294
!	27152	a	46711
ne	27102	's	40098
u	25218	it	38275
...	23503	't	36346
sam	22492	and	29469
to	19935	that	28205
na	15105	of	28091
za	13491	...	28045
što	12890	in	23045
mi	12670	!	20655
ti	12388	me	20253
li	11497	he	17259

Slike u nastavku prikazuju distribucije tokena u hrvatskom, odnosno engleskom dijelu podatkovnog skupa korištenog pri treniranju jezičnih i prijevodnih modela iz općenite domene. Najčešći tokeni su oni koji se pojavljuju samo jednom (Slika 18), odnosno 2-10 puta (Slika 19).



Slika 18. Histogram distribucije tokena u hrvatskom dijelu podatkovnog skupa korištenog pri treniranju jezičnih i prijevodnih modela iz **općenite domene**.



Slika 19. Histogram distribucije tokena u engleskom dijelu podatkovnog skupa korištenog pri treniranju jezičnih i prijevodnih modela iz **općenite domene**.

Tablica 19 prikazuje karakteristike podatkovnog skupa korištenog za treniranje jezičnih i prijevodnih modela iz specifične domene, tj. domene računalnog softvera. Karakteristike se odnose na korpus prije pretprocesiranja. S obzirom na broj rečenica/segmenata, ovaj skup je cca. 16 puta manji od paralelnog korpusa iz općenite domene. Sastoji se od ukupno 36236 rečenica na hrvatskom i engleskom jeziku, sastavljenih od ukupno 378469 riječi, tj. 2499530 znakova.

Razlika u veličini vokabulara je 7510 riječi. Najdulje rečenice sastojale su se maksimalno od 73 riječi (hrvatski), odnosno 86 riječi u engleskom dijelu korpusa. Prosječan broj riječi po rečenici, tj. segmentu iznosio je 10.04 (za hrvatski), odnosno 10.85 za engleski jezik, a prosječan broj znakova po rečenici, tj. segmentu iznosio je 71.92 u hrvatskom dijelu korpusa, a 66.04 u engleskom dijelu. Standardna devijacija broja riječi po rečenici/segmentu u ovom je slučaju znatno veća: 7.25 za hrvatski dio korpusa, odnosno 7.99 za engleski dio paralelnog korpusa.

Tablica 19. Karakteristike podatkovnih skupova za treniranje jezičnih i prijevodnih modela iz **specifične domene, tj. domene računalnog softvera**.

Karakteristike podatkovnog skupa	**hrvatski**	**engleski**
broj znakova	1303106	1196424
broj riječi	181910	196559
broj različitih riječi (vokabular)	22498	14988
broj rečenica/segmenata	18118	18118
maksimalni broj riječi po rečenici/segmentu	73	86
minimalni broj riječi po rečenici/segmentu	1	1
prosječan broj znakova po rečenici/segmentu	71.92	66.04
prosječan broj riječi po rečenici/segmentu	10.04	10.85
standardna devijacija (broj riječi po rečenici/segmentu)	7.25	7.99

Tablica 20 prikazuje analizu distribucije riječi u podatkovnim skupovima za treniranje jezičnih i prijevodnih modela iz specifične domene, tj. domene računalnog softvera, i to prema frekvencijama riječi. I u ovom slučaju vidljiv je već spomenuti trend opadanja udjela u vokabularu i ukupnom broju riječi za oba jezika. Riječi koje se pojavljuju samo jednom, dva ili tri puta čine oko 75% vokabulara, no, istovremeno samo 13% od ukupnog broja riječi u hrvatskom dijelu paralelnog korpusa. Riječi koje se pojavljuju od četiri do 10 puta čine oko 15-16% vokabulara u hrvatskom i engleskom dijelu korpusa, pri čemu čine oko 11% (za hrvatski), odnosno 7% (za engleski) od ukupnog broja riječi u paralelnom korpusu. Riječi koje se pojavljuju više od 10 puta čine 10.32% vokabulara u hrvatskom dijelu korpusa, odnosno 13.69% vokabulara u engleskom dijelu, te 78.81% od ukupnog broja riječi u hrvatskom dijelu korpusa, odnosno 85.01% u engleskom dijelu. U engleskom dijelu korpusa, prve tri kategorije čine 70.3% vokabulara, a istovremeno samo 7.58% od ukupnog broja riječi.

Tablica 20. Distribucija riječi u podatkovnim skupovima za treniranje jezičnih i prijevodnih modela iz **specifične domene, tj. domene računalnog softvera**, prema frekvencijama riječi.

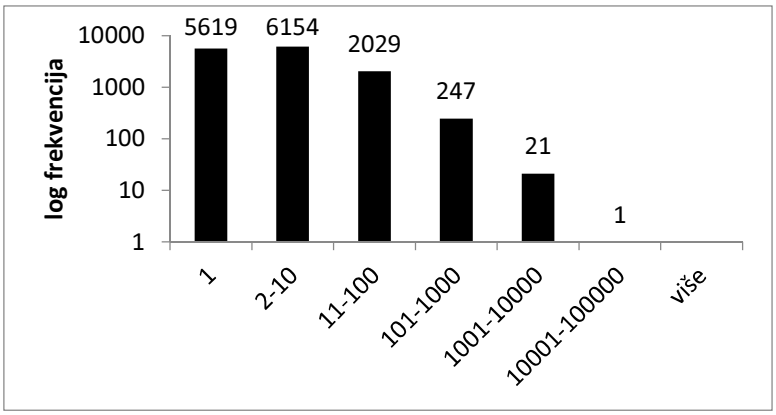
Frekvencije riječi unutar podatkovnog skupa	Broj riječi	Udio u vokabularu	Udio u ukupnom broju riječi
hrvatski			
1, broj riječi koje se pojavljuju točnom jednom (hapax legomena)	11618	51.64%	6.39%
2, broj riječi koje se pojavljuju točno dva puta (dis legomena)	3466	15.41%	3.81%
3, broj riječi koje se pojavljuju točno tri puta (tris legomena)	1696	7.54%	2.80%
4	981	4.36%	2.16%
5	742	3.30%	2.04%
6	492	2.19%	1.62%
7	390	1.73%	1.50%
8	302	1.34%	1.33%
9	255	1.13%	1.26%
10	234	1.04%	1.29%
> 10	2322	10.32%	75.81%
engleski			
1, broj riječi koje se pojavljuju točnom jednom (hapax legomena)	7274	48.53%	3.70%
2, broj riječi koje se pojavljuju točno dva puta (dis legomena)	2155	14.38%	2.19%
3, broj riječi koje se pojavljuju točno tri puta (tris legomena)	1107	7.39%	1.69%
4	701	4.68%	1.43%
5	464	3.10%	1.18%
6	359	2.40%	1.10%
7	280	1.87%	1.00%
8	229	1.53%	0.93%
9	178	1.19%	0.82%
10	189	1.26%	0.96%
> 10	2052	13.69%	85.01%

Tablica 21 prikazuje dvadeset najfrekventnijih tokena u podatkovnim skupovima korištenim za treniranje jezičnih i prijevodnih modela iz specifične domene, domene računalnog softvera. Također očekivano prevladavaju funkcijske riječi i interpunkcijski znakovi.

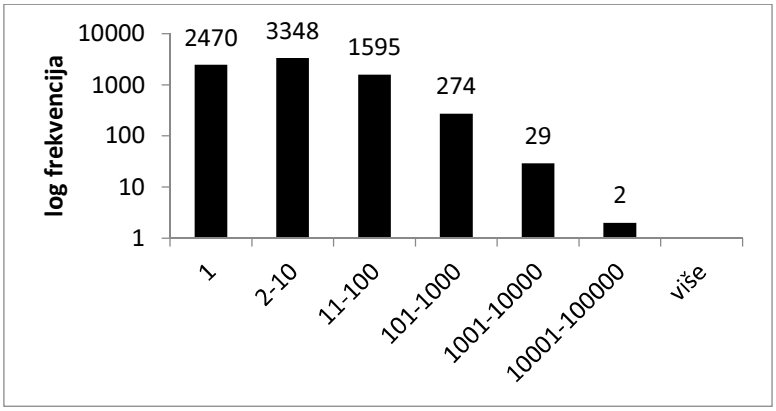
Tablica 21. Dvadeset najfrekventnijih tokena u podatkovnim skupovima korištenim pri treniranju jezičnih i prijevodnih modela iz **specifične domene, tj. domene računalnog softvera**.

hrvatski		**engleski**	
token	**frekvencija**	**token**	**frekvencija**
.	11202	the	14895
,	6192	.	11174
u	5404	,	9599
i	4571	to	6026
za	4050	a	4924
na	3404	and	4553
se	2651	in	4058
:	2447	you	3682
ili	2286	of	2862
odaberite	2071	PDF	2771
PDF	1996	:	2434
da	1716	or	2400
je	1547	for	2329
/	1511	can	1781
kliknite	1462	that	1682
(	1336	is	1619
)	1335	click	1527
ako	1272	/	1455
s	1247	select	1434
možete	1244	;	1282

Slike 20 i 21 prikazuju distribucije tokena u hrvatskom, odnosno engleskom dijelu podatkovnog skupa korištenog pri treniranju jezičnih i prijevodnih modela iz specifične domene, tj. domene računalnog softvera. Najčešći tokeni su oni koji se u korpusu pojavljuju 2-10 puta.



Slika 20. Grafički prikaz distribucije tokena u hrvatskom dijelu podatkovnog skupa korištenog pri treniranju jezičnih i prijevodnih modela iz **specifične domene, tj. domene računalnog softvera**.



Slika 21. Grafički prikaz distribucije tokena u engleskom dijelu podatkovnog skupa korištenog pri treniranju jezičnih i prijevodnih modela iz **specifične domene, tj. domene računalnog softvera**.

Ugađanje pomoću podatkovnog skupa iz specifične domene također predstavlja oblik adaptacije domene. Tablica u nastavku prikazuje karakteristike (prije pretprocesiranja) dvojezičnih podatkovnih skupova korištenih za ugađanje (eng. *tuning/development set*) statističkih modela za sve izgrađene sustave za statističko strojno prevođenje (sustave 1-8), tj. prevođenje općenite i specifične domene računalnog softvera.

Podatkovni skupovi sastojali su se od ukupno 8926 riječi sastavljenih od 59072 znakova, raspoređenih u po 1000 rečenica, tj. segmenata za svaki jezik (Tablica 22).

Broj različitih riječi bio je očekivano veći u hrvatskom dijelu korpusa. U hrvatskom dijelu korpusa najdulja rečenica sastojala se od 21 riječi, a engleskom dijelu od 6 riječi više. Najkraći segment također je bio duljine jedan u oba dijela paralelnog korpusa. Prosječan broj riječi po rečenici/segmentu bio 4.26 u hrvatskom dijelu korpusa, te 4.66 u engleskom dijelu. Standardna devijacija u hrvatskom dijelu iznosila je cca. 3.3, a u engleskom dijelu oko 4.

Tablica 22. Karakteristike podatkovnih skupova za ugađanje statističkih modela za **općenitu i specifičnu domenu, tj. domenu računalnog softvera**.

Karakteristike podatkovnog skupa	**hrvatski**	**engleski**
broj znakova	30538	28534
broj riječi	4265	4661
broj različitih riječi (vokabular)	1818	1445
broj rečenica/segmenata	1000	1000
maksimalni broj riječi po rečenici/segmentu	21	27
minimalni broj riječi po rečenici/segmentu	1	1
prosječan broj znakova po rečenici/segmentu	30.53	28.53
prosječan broj riječi po rečenici/segmentu	4.26	4.66
standardna devijacija (broj riječi po rečenici/segmentu)	3.36	4.08

U sljedećoj tablici (Tablica 23) dane su distribucije riječi prema frekvencijama u podatkovnom skupu specifične domene (računalni softver) upotrijebljenim za ugađanje statističkih modela u svim izgrađenim sustavima za statističko strojno prevođenje (sustavi 1-8). Skoro 90% vokabulara čine prve tri kategorije frekvencija riječi, a time pokrivaju više od 50% od ukupnog broja riječi u hrvatskom dijelu korpusa. Oko 80% vokabulara u engleskom dijelu otpada na prve 3 kategorije, čime pokrivaju oko 35% vokabulara. Zanimljivo je kako riječi koje se pojavljuju više od 10 puta čine samo cca. 26% od ukupnog broja riječi u hrvatskom dijelu paralelnog korpusa, što odgovara 2.53% vokabulara.

To upućuje na manju upotrebu funkcijskih riječi te većeg značaja sadržajnih riječi. To je i cilj podatkovnog skupa za ugađanje statističkih modela kojim se nastoje pokriti određene rečenične konstrukcije specifične određenoj domeni, u ovom slučaju domeni računalnog softvera. U engleskom dijelu korpusa, riječi koje se pojavljuju više od 10 puta ima znatno više, tj. preko 40% od ukupnog broja riječi, pri čemu čine oko 5% vokabulara.

Tablica 23. Distribucija riječi u podatkovnim skupovima za ugađanje statističkih modela iz **specifične domene, tj. domene računalnog softvera**, prema frekvencijama riječi.

Frekvencije riječi unutar podatkovnog skupa	Broj riječi	Udio u vokabularu	Udio u ukupnom broju riječi
hrvatski			
1, broj riječi koje se pojavljuju točnom jednom (hapax legomena)	1153	63.42%	27.03%
2, broj riječi koje se pojavljuju točno dva puta (dis legomena)	308	16.94%	14.44%
3, broj riječi koje se pojavljuju točno tri puta (tris legomena)	130	7.15%	9.14%
4	71	3.91%	6.66%
5	41	2.26%	4.81%
6	22	1.21%	3.09%
7	20	1.10%	3.28%
8	11	0.61%	2.06%
9	9	0.50%	1.90%
10	7	0.39%	1.64%
> 10	46	2.53%	25.93%
engleski			
1, broj riječi koje se pojavljuju točnom jednom (hapax legomena)	805	55.71%	17.27%
2, broj riječi koje se pojavljuju točno dva puta (dis legomena)	253	17.51%	10.86%
3, broj riječi koje se pojavljuju točno tri puta (tris legomena)	117	8.10%	7.53%
4	65	4.50%	5.58%
5	57	3.94%	6.11%
6	29	2.01%	3.73%
7	17	1.18%	2.55%
8	14	0.97%	2.40%
9	14	0.97%	2.70%
10	3	0.21%	0.64%
> 10	71	4.91%	40.61%

Za provjeru performansi svih izgrađenih sustava za statističko strojno prevođenje (sustavi 1-8) te evaluaciju kvalitete strojnih prijevoda generiranih pomoću web servisa (Google Translate i Yandex Translate) korišten je skup za testiranje (ispitivanje) koji je različit u odnosu na skupove za treniranje i ugađanje (nema preklapanja rečenica, tj. segmenata). Podatkovni skupovi (prije pretprocesiranja) namijenjeni testiranju, tj. ispitivanju (eng. *test set*) svih sustava za statističko strojno prevođenje, tj. i za općenitu i specifičnu domenu (računalni softver) opisani su u nastavku (Tablica 24). Za testiranje sustava za statističko strojno prevođenje korišten je skup od po 1000 rečenica/segmenata za svaki jezik. Sadrži oko 17000 riječi ukupno, sastavljenih od oko 115000 znakova. Vokabular na hrvatskom jeziku veći je za oko 600 riječi u odnosu na engleski dio korpusa. Prosječan broj riječi po rečenici, tj. segmentu otprilike je jednak u oba jezika. Najdulja rečenica/segment broji 36 riječi na hrvatskom jeziku, odnosno 3 riječi više na engleskom jeziku. Standardna devijacija iznosi oko 7 riječi po rečenici, tj. segmentu.

Tablica 24. Karakteristike podatkovnih skupova za testiranje sustava za statističko strojno prevođenje **za općenitu** i **specifičnu domenu, tj. domenu računalnog softvera**.

Karakteristike podatkovnog skupa	**hrvatski**	**engleski**
broj znakova	58678	54337
broj riječi	8146	8962
broj različitih riječi (vokabular)	2686	2028
broj rečenica/segmenata	1000	1000
maksimalni broj riječi po rečenici/segmentu	36	39
minimalni broj riječi po rečenici/segmentu	1	1
prosječan broj znakova po rečenici/segmentu	58.68	54.34
prosječan broj riječi po rečenici/segmentu	8.15	8.96
standardna devijacija (broj riječi po rečenici/segmentu)	6.69	7.58

Distribucija riječi prema frekvencijama riječi u podatkovnim skupovima za testiranje sustava za statističko strojno prevođenje za općenitu i specifičnu domenu računalnog softvera opisana je u nastavku (Tablica 25). Hapax legomena čine također velik udio u vokabularu (oko 61%), što odgovara 20% od ukupnog broja riječi u hrvatskom dijeli korpusa. U engleskom dijelu podatkovnog skupa ta je brojka nešto manja, tj. cca. 56% vokabulara te oko 13% od ukupnog broja riječi u korpusu. Riječi koje se pojavljuju više od 10 puta čine oko 41% od ukupnog broja riječi u hrvatskom dijelu korpusa, odnosno cca. 56% u engleskom dijelu korpusa. Riječi koje se

pojavljuju između 4 i 10 puta čine cca. 11% vokabulara u hrvatskom dijelu korpusa, odnosno 14.5% u engleskom dijelu korpusa, što odgovara cca. 21% od ukupnog broja riječi u hrvatskom dijelu korpusa te oko 20% u engleskom dijelu korpusa.

Tablica 25. Distribucija riječi u podatkovnim skupovima za testiranje sustava za statističko strojno prevođenje za **općenitu** i **specifičnu domenu**, **tj. domenu računalnog softvera**, prema frekvencijama riječi.

Frekvencije riječi unutar podatkovnog skupa	**Broj riječi**	**Udio u vokabularu**	**Udio u ukupnom broju riječi**
hrvatski			
1, broj riječi koje se pojavljuju točnom jednom (hapax legomena)	1636	60.91%	20.08%
2, broj riječi koje se pojavljuju točno dva puta (dis legomena)	469	17.46%	11.51%
3, broj riječi koje se pojavljuju točno tri puta (tris legomena)	174	6.48%	6.41%
4	97	3.61%	4.76%
5	59	2.20%	3.62%
6	56	2.08%	4.12%
7	29	1.08%	2.49%
8	25	0.93%	2.46%
9	19	0.71%	2.10%
10	10	0.37%	1.23%
> 10	112	4.17%	41.21%
engleski			
1, broj riječi koje se pojavljuju točnom jednom (hapax legomena)	1128	55.62%	12.59%
2, broj riječi koje se pojavljuju točno dva puta (dis legomena)	328	16.17%	7.32%
3, broj riječi koje se pojavljuju točno tri puta (tris legomena)	150	7.40%	5.02%
4	77	3.80%	3.44%
5	66	3.25%	3.68%
6	53	2.61%	3.55%
7	33	1.63%	2.58%
8	28	1.38%	2.50%

9	24	1.18%	2.41%
10	14	0.69%	1.56%
> 10	127	6.26%	55.36%

Sve prethodno navedene podatkovne skupove, tj. korpuse za treniranje, ugađanje i testiranje trebalo je pretprocesirati. Pretprocesiranje u ovom istraživanju sastojalo se od sljedećih procesa (Koehn, 2015):

- tokenizacija (eng. *tokenisation*), kojom se umeću razmaci između riječi i interpunkcije,
- postupak pretvaranja početnog znaka riječi u malo ili veliko slovo (eng. *truecasing*), tj. najvjerojatniji oblik riječi,
- čišćenje korpusa (eng. *cleaning*), kojim se uklanjaju predugačke, prekratke, neuparene, nekompatibilne (9:1 omjer za GIZA++ sravnjivanje) i prazne rečenice/segmenti te suvišni razmaci.

Za tokenizaciju podatkovnih skupova korišteni su Perl skripta `tokenizer.perl`[12] koja je sastavni dio Moses 1.0 Toolkit-a te vlastita tokenizacijska pravila za hrvatski jezik koja su sačuvana u datoteci `nonbreaking_prefix.hr`. Tokenizacijska pravila su važna zbog definiranja izuzeća pri tokenizaciji, tj. radi se o popisu kratica koje ne označavaju kraj rečenice ukoliko završavaju točkom. Nakon tokenizacije, uslijedio je postupak pretvaranja početnog znaka riječi u prikladno malo ili veliko slovo. U tu svrhu korišteni su tokenizirani podatkovni skupovi, a bilo je potrebno trenirati i alat za pretvaranje početnog znaka riječi (eng. *truecaser*) `truecase.perl`[13] pomoću skripte `train-truecaser.perl`[14]. Zatim su skupovi očišćeni alatom `clean-corpus-n.perl`[15], tj. uklonjene su rečenice dulje od **80** riječi. Time su se podatkovni skupovi smanjili za oko 0.01% (korpus za općenitu domenu), odnosno 0.04% (korpus korišten za specifičnu domenu, tj. domenu računalnog softvera).

[12] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl
[13] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/recaser/truecase.perl
[14] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/recaser/train-truecaser.perl
[15] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/training/clean-corpus-n.perl

5.5. Izgradnja sustava za statističko strojno prevođenje temeljeno na frazama

Računalo koje je korišteno pri izgradnji sustava za statističko strojno prevođenje i u eksperimentalne svrhe imalo je sljedeće karakteristike:

- operativni sustav: Ubuntu 12.04.03 LTS
- Linux kernel: 3.8.0-32 generic
- radna memorija: 6 GB
- procesor: dvojezgreni, Inter Core2 CPU 6300@1.86 GHz

Izgradnja i evaluacija sustava za statističko strojno prevođenje temeljeno na frazama (sustavi 1-8) prolazili su kroz više faza, koje se načelno mogu svrstati u jednu od dvije glavnih kategorija (Koehn, 2015):

- procesi u kanalu treniranja statističkih modela (eng. *training pipeline*), koji obuhvaćaju treniranje jezičnih, prijevodnih i ostalih modela te ugađanje težina značajki,
- procesi dekodiranja, tj. generiranja strojnih prijevoda, koji obuhvaćaju prevođenje podatkovnog skupa za testiranje.

Pretprocesirani ulazni podatkovni skupovi korišteni su kao jednojezični, odnosno paralelni korpusi za izgradnju jezičnih, odnosno prijevodnih statističkih modela, na koje se oslanja model sustava za statističko strojno prevođenje temeljeno na frazama. Primijenjene su i heuristike za uklanjanje rečeničnih parova, tj. parova segmenata koji su vjerojatno netočno sravnjeni, a pored toga, uklonjene su i predugačke rečenice.

Vrlo važan element modela sustava za statističko strojno prevođenje jest jezični model, tj. statistički model koji se, s obzirom da osigurava fluentnost u ciljnom jeziku, izgrađuje pomoću jednojezičnog korpusa ciljnog jezika. Za izgradnju jezičnih modela upotrijebljen je IRSTLM

5.80.03[16] (Bertoldi, 2008; Federico et al., 2008b). Svi jezični modeli izgrađeni su s redom **3**. Prije same izgradnje jezičnih modela dodani su granični simboli koji označavaju početak i kraj rečenice: $< s >$ i $</s >$. U svim jezičnim modelima uklonjeni su n-grami koji se pojavljuju samo jednom u korpusu korištenom za treniranje (Koehn, 2015). Time se znatno smanjuje jezični model pri čemu je utjecaj na performanse sustava za statističko strojno prevođenje minimalan. Za izglađivanje korištena je modificirana inačica Kneser-Ney metode izglađivanja. Dobiveni jezični modeli u ARPA formatu su zatim binarizirani pomoću Moses 1.0 Toolkit-a.

Tablica 26 prikazuje broj jedinstvenih n-grama u jezičnim modelima. Iako je jezični model za općenitu domenu treniran na korpusu koji je 16 puta veći, broj unigrama i bigrama u jezičnom modelu treniranom za specifičnu domenu samo je cca. 7 puta manji u hrvatskom jeziku, odnosno oko 5-6 puta u engleskom jeziku. Broj trigrama (za oba jezika) u jezičnom modelu za općenitu domenu veći je oko 10 puta u odnosu na broj trigrama u jezičnom modelu specifične domene.

Tablica 26. Broj jedinstvenih n-grama u jezičnim modelima treniranim na općenitom korpusu, odnosnu specifičnom korpusu, tj. na domeni računalnog softvera.

jedinstveni n-grami	**općenita domena**		**specifična domena**	
	hrvatski	**engleski**	**hrvatski**	**engleski**
unigrami	102572	41258	14110	7731
bigrami	612645	391221	88396	66333
trigrami	222911	288112	22499	29092

Tablica 27 prikazuje vrijednosti perpleksnosti jezičnih modela. Perpleksnosti su očekivano vrlo velike za jezične modele iz općenite domene kada se računaju za podatkovne skupove za ugađanje, odnosno za ispitivanje, tj. testiranje. Naime oba skupa (skupovi za ugađanje i za ispitivanje), pripadaju vrlo specifičnoj domeni, tj. domeni računalnog softvera, dok skup za treniranje jezičnih modela pripada domeni šarolikih filmskih titlova (podslova).

Perpleksnost ukazuje na vrlo velik nesrazmjer domena, tj. veliku različitost podatkovnog skupa za treniranje jezičnog modela i podatkovnih skupova za ugađanje i ispitivanje, tj. testiranje. Perpleksnost u jezičnim modelima treniranim na skupovima za treniranje domenski specifičnih jezičnih modela je 16 puta manja za hrvatski i čak 25 puta manja za engleski jezik ako se računa za podatkovni skup za ugađanje. Čak 220 puta manja je za hrvatski jezik kada se računa za

[16] http://sourceforge.net/projects/irstlm/

podatkovni skup za testiranje, a 94 puta manja za engleski jezik. To ukazuje da su oba podatkovna skupa, iako pripadaju istoj specifičnoj domeni (računalni softver), također međusobno vrlo različita.

Tablica 28 prikazuje udio riječi izvan vokabulara u jezičnim modelima. Najveći udio riječi izvan vokabulara nalazi se u hrvatskom jezičnom modelu treniranom za općenitu domenu, dok je najmanji udio nepoznatih riječi u engleskom domenski specifičnom jezičnom modelu.

Tablica 27. Izračun perpleksnosti jezičnih modela s obzirom na skup za ugađanje, odnosno skup za ispitivanje, tj. testiranje.

skupovi	**općenita domena**		**specifična domena**	
	hrvatski	**engleski**	**hrvatski**	**engleski**
skup za ugađanje	84140.92	28721.39	5099.87	1136.56
skup za ispitivanja, tj. testiranje	150569.02	28245.26	683.05	300.19

Tablica 28. Udio riječi izvan vokabulara u jezičnim modelima.

skupovi	**općenita domena**		**specifična domena**	
	hrvatski	**engleski**	**hrvatski**	**engleski**
skup za ugađanje	18.64%	11.33%	10.59%	3.91%
skup za ispitivanja, tj. testiranje	25.21%	13.89%	4.11%	1.83%

Paralelne rečenice, tj. segmenti zatim su sravnjeni na razini riječi pomoću alata GIZA++ 1.0.7[17] (Och i Ney, 2003), koji sadrži niz statističkih modela implementiranih 1980-ih godina u kompaniji IBM (Koehn, 2015). Dobivene sravnjenosti riječi korištene su za ekstrakciju fraznih parova koji su zatim sačuvani u prijevodnom modelu, tj. tablici prijevoda fraza/fraznih struktura. Nadalje, sravnjenost riječi iskorištena je za ekstrakciju statističkih podataka, kao npr. podataka o orijentaciji sravnjivanja koja se može iščitati iz matrice sravnjenosti riječi ili podataka o vjerojatnostima sravnjivanja fraza.

Za učenje leksičke vjerojatnosti, ekstrakciju fraza i izračun vjerojatnosti prevođenja fraze, generiranje modela leksičkog preslagivanja/premještanja fraza i konfiguracijskih datoteka upotrijebljena je skripta `train-model.perl`[18]. Riječi u frazi mogle su se permutirati do

[17] http://www.statmt.org/moses/giza/GIZA++.html
[18] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/training/train-model.perl

maksimalne udaljenosti od **6** riječi (maksimalna distorzija u svim modelima). Kao simetrizacijska heuristika korišten je *grow-diag-final-and*. Model leksičkog preslagivanja/premještanja fraza predviđa orijentaciju tipa $\boldsymbol{m}$, $\boldsymbol{s}$, $\boldsymbol{d}$ i treniran je za oba jezična smjera. Veličine tablica prijevoda fraza, tj. fraznih struktura iznosile su između 56 MB (za domenski specifične prijevodne modele) i 428 MB (za prijevodne modele općenite domene).

Log linearni model sustava za statističko strojno prevođenje temeljeno na frazama standardno se sastojao od značajki kao što su: jezični model, prijevodni model, inverzni prijevodni model (obrnuti jezični smjer), model leksičkih težina za prevođenje riječi (za oba smjera), model penaliziranja riječi, model penaliziranja fraza, model leksičkog preslagivanja/premještanja fraza, tj. distorzije fraza (za oba smjera). Sve težine navedenih značajki trebalo je ugađati kako bi sustav bio što efikasniji.

Nakon što su statistički modeli izgrađeni pokrenut je proces ugađanja težina u log-linearnom modelu sustava za statističko strojno prevođenje temeljeno na frazama, s ciljem da sustav u konačnici proizvede što bolje strojne prijevode. Za ugađanje težina značajki modela korišten je algoritam MERT (Koehn, 2015) te metrika BLEU uz pomoć skripte `mert-moses.pl`[19] koja generira nove konfiguracijske datoteke, a primjenjuje pretprocesirani podatkovni skup za ugađanje, u obliku paralelnog korpusa iz domene računalnog softvera. Podatkovni skup za ugađanje različit je u odnosu na skupove za treniranje jezičnog i prijevodnog modela, tj. radi se o disjunktnim skupovima. Nakon procesa ugađanja binarizirane su tablice prijevoda fraza, tj. fraznih struktura i model leksičkog preslagivanja/premještanja fraza radi bržeg učitavanja u radnu memoriju računala.

Za mjerenje performansi svih sustava za strojno prevođenje korišten je isti pretprocesirani podatkovni skup za testiranje. Radi se također podatkovnom skupu različitom u odnosu na skupove za treniranje i ugađanje, tj. radi se i u ovom slučaju o disjunktnim skupovima. Disjunktnost je osigurana vlastitom sed skriptom.

Za dekodiranje podatkovnih skupova za testiranje (ispitivanje) upotrijebljen je dekoder koji primjenjuje zrakasto pretraživanje s odbacivanjem, a koji u Moses 1.0 Toolkit-u dolazi kao zasebna aplikacija pisana u jeziku C++. Za kompiliranje Moses 1.0 Toolkit-a bili su potrebni boost, tj. skup biblioteka za C++, bjam (sustav Boost.Build) i gcc (eng. *GNU Compiler Collectio*n), te zlib i bzip2 alati za kompresiju podataka.

[19] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/training/mert-moses.pl

Zadaća dekodera u izgrađenim sustavima bila je za danu izvornu rečenicu iz skupa za testiranje (ispitivanje) pronaći najvjerojatniju (prema statističkim modelima) rečenicu u ciljnom jeziku (eng. *highest scoring sentence*). Radi dodatnog ubrzavanja postupka dekodiranja izvršeno je filtriranje tablice prijevoda fraza, tj. fraznih struktura pomoću skripte `filter-model-given-input.pl`[20] kako bi se koristili samo oni unosi iz tablice prijevoda fraza koji su zaista i potrebni za dekodiranje podatkovnog skupa za testiranje. Dekoderu su u procesu generiranja strojnog prijevoda proslijeđene ranije binarizirane inačice prijevodnog modela i modela leksičkog preslagivanja/premještanja fraza.

[20] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/training/filter-model-given-input.pl

5.6. Performanse sustava za statističko strojno prevođenje temeljeno na frazama

Ranije je već navedeno da je u ovom doktorskom istraživanju izgrađeno ukupno 8 sustava:

smjer: hrvatsko-engleski

- sustav 1: treniran na korpusu opće domene
- sustav 2: treniran na specijaliziranom korpusu, tj. domeni računalnog softvera
- sustav 3: hibridni sustav koji koristi tablice prijevoda fraza, tj. fraznih struktura iz obje domene, a preostale značajke modela (leksičko premještanje, jezični model itd.) iz domene računalnog softvera
- sustav 4: hibridni sustav koji koristi tablice prijevoda fraza, tj. fraznih struktura iz obje domene, a preostale značajke modela (leksičko premještanje, jezični model itd.) iz opće domene

smjer: englesko-hrvatski

- sustav 5: treniran na korpusu opće domene
- sustav 6: treniran na specijaliziranom korpusu domene računalnog softvera
- sustav 7: hibridni sustav koji koristi tablice prijevoda fraza, tj. fraznih struktura iz obje domene, a preostale značajke modela (leksičko premještanje, jezični model itd.) iz domene računalnog softvera
- sustav 8: hibridni sustav koji koristi tablice prijevoda fraza, tj. fraznih struktura iz obje domene, a preostale značajke modela (leksičko premještanje, jezični model itd.) iz opće domene

Svi sustavi ugođeni su pomoću istog podatkovnog skupa, tj. skupa za ugađanje koji se sastojao od 1000 rečenica, tj. segmenata iz specifične domene, tj. domene računalnog softvera. Jednako tako, pri evaluaciji automatskim metrika svi izgrađeni sustavi (i web servisi) ispitani su pomoću

istog podatkovnog skupa za testiranje, tj. ispitivanje, a sastojao se također od 1000 rečenica, tj. segmenata iz domene računalnog softvera.

Performanse svih izgrađenih sustava za statističko strojno prevođenje temeljeno na frazama izmjerene su pomoću automatskih evaluacijskih metrika: BLEU, NIST, METEOR, GTM, WER i TER. U svim provedenim istraživanjima korišten je METEOR-ov modul koji pri identificiranju zajedničkih preklapajućih dijelova uzima u obzir samo potpuno identične dijelove.

Rezultati mjerenja performansi prvog sustava, tj. sustava treniranog na korpusu opće domene za hrvatsko-engleski smjer dani su u Tablici 29. Sustav 1 očekivano je prošao poprilično loše. Ispitivanje performansi pomoću toliko različitog podatkovnog skupa ima snažne posljedice na sustave koji nisu trenirani na istoj ili barem srodnoj domeni (u ovom slučaju domena računalnog softvera). Relativno slabe performanse potvrđene su svim metrikama. BLEU vrijednost 0.07 vrlo je niska što upućuje na relativno slabo (rijetko) preklapanja n-grama iz referentnog i strojnog prijevoda. Vrijednost NIST-a je također relativno niska, što ukazuje na manji broj rjeđih n-grama koji se smatraju informativnijima. METEOR koji je inače vrlo pogodan za evaluaciju kvalitete na razini rečenice, također postiže vrlo nisku vrijednost (cca. 0.12). S obzirom da METEOR veći od 0.5 upućuje na razumljive prijevode, strojni prijevodi generirani sustavom 1 mogu se smatrati relativno nerazumljivima. GTM metrika također postiže niske rezultate, što upućuje na malen broj preklapanja n-grama s ispravnim poretkom riječi. WER i TER vrlo su visoki, što ukazuje na velik broj potrebnih izmjena da bi se strojni prijevod preoblikovao u referentni prijevod.

Tablica 29. Performanse sustava za statističko strojno prevođenje temeljeno na frazama: **opća domena, hrvatsko-engleski smjer (sustav 1)**.

Metrika	**opća domena**
BLEU	0.0732
NIST	2.9351
METEOR	0.1169
GTM	0.3401
WER	0.7953
TER	0.7726

Tablica 30 prikazuje rezultate mjerenja performansi sustava za statističko strojno prevođenje broj 2, tj. sustava treniranog na specijaliziranom korpusu, tj. domeni računalnog softvera, za

hrvatsko-engleski smjer. Ovdje je vidljiv velik porast kvalitete strojnog prijevoda. Sve metrike upućuju na razumljive i relativno kvalitetne strojne prijevode. Ponajviše se ističe metrika BLEU koja ukazuje na velik broj podudaranja n-grama u strojnom i referentnom prijevodu. Sustav 2 potvrđuje da se izgradnjom domenski specifičnih sustava drastično povećava kvaliteta dekodiranja domenskog testnog skupa za hrvatsko-engleski smjer. Drugim riječima, prilagodba domene i karakteristike ulaznoga podatkovnog skupa utječu na kvalitetu strojnog prijevoda, što dijelom potvrđuje prvu hipotezu (H1).

Tablica 30. Performanse sustava za statističko strojno prevođenje temeljeno na frazama: **specifična domena, tj. domena računalnog softvera, hrvatsko-engleski smjer (sustav 2)**.

metrika	**domena računalnog softvera**
BLEU	0.3734
NIST	7.4487
METEOR	0.3316
GTM	0.6750
WER	0.4966
TER	0.3681

Tablica 31 prikazuje rezultate mjerenja performansi hibridnog sustava za statističko strojno prevođenje temeljeno na frazama za hrvatsko-engleski smjer, koji koristi tablice prijevoda fraza, tj. fraznih struktura iz obje domene, a preostale značajke modela (leksičko premještanje, jezični model itd.) iz domene računalnog softvera. Hibridni sustav 3 pokazao se kvalitetnijim u odnosu na sustav 1 koji koristi isključivo prijevodni model iz općenite domene. To potvrđuju i rezultati svih automatskih evaluacijskih metrika.

Razlog tome je što statistički modeli trenirani na izvandomenskom podatkovnom skupu (općenita domena) osiguravaju relativno visok odziv, dok istovremeno domenski specifičan podatkovni skup pruža preciznost u modelu sustava za statističko strojno prevođenje. Time je dijelom potvrđena hipoteza H2 da se hibridnim pristupom može poboljšati kvaliteta strojnog prijevoda za hrvatsko-engleski smjer u odnosu na sustave koji koriste značajke svojstvene isključivo jednoj domeni.

Tablica 31. Performanse **hibridnog sustava** za statističko strojno prevođenje temeljeno na frazama: tablice prijevoda fraza, tj. fraznih struktura iz **obje domene**, a preostale značajke modela **iz specifične domene, tj. domene računalnog softvera, hrvatsko-engleski smjer (sustav 3)**.

metrika	hibridni sustav
BLEU	0.1229
NIST	3.5689
METEOR	0.1367
GTM	0.3923
WER	0.7272
TER	0.7103

Tablica 32 prikazuje rezultate mjerenja performansi hibridnog sustava za statističko strojno prevođenje temeljeno na frazama za hrvatsko-engleski smjer, koji koristi tablice prijevoda fraza, tj. fraznih struktura iz obje domene, a preostale značajke modela (leksičko premještanje, jezični model itd.) iz opće domene. Sustav 4 nešto je slabiji u odnosu na sustav 3. To je dijelom i očekivano, s obzirom da sve bitne značajke (osim prijevodnog modela iz domene računalnog softvera) proizlaze iz općenite domene što se nije pokazalo jednako efikasnim rješenjem za prevođenje skupa za testiranje iz domene računalnog softvera.

Istako tako, razlog tome jest što jezični model općenite domene nije pokrivao specifične rečenične konstrukcije iz testirane domene, a to je vidljivo i iz izračuna perpleksnosti jezičnih modela (Tablica 27). Ipak, hibridni sustav 4 bolji je u odnosu na sustav 1, što potvrđuje hipotezu H2, da se primjenom hibridnog sustava za statističko strojno prevođenje temeljeno na frazama za hrvatsko-engleski može postići poboljšanje kvalitete strojnog prijevoda u odnosu na sustave koji koriste značajke svojstvene isključivo jednoj domeni.

Tablica 32. Performanse **hibridnog sustava** za statističko strojno prevođenje temeljeno na frazama: tablice prijevoda fraza, tj. fraznih struktura iz **obje domene**, a preostale značajke modela iz **opće domene, hrvatsko-engleski smjer (sustav 4)**.

metrika	hibridni sustav
BLEU	0.1064
NIST	3.3956
METEOR	0.1315

GTM	0.3817
WER	0.7244
TER	0.7092

U nastavku su prikazani rezultati mjerenja performansi sustava broj 5. Radi se o sustavu treniranom na korpusu opće domene, za engleski-hrvatski smjer (Tablica 33). Sustav treniran na općoj domeni za englesko-hrvatski smjer očekivano je prošao najlošije od svih 8 sustava. BLEU vrijednost od cca. 0.05 izrazito je niska. Budući da se metrika NIST temelji na BLEU-u ne iznenađuje ni relativno niska NIST vrijednost od cca. 2.39. Također vrlo niska vrijednost METEOR-a upućuje na relativno neupotrebljive strojne prijevode iz domene računalnog softvera. To ne iznenađuje, s obzirom da je u ovom doktorskom radu korištena jezično neovisna inačica METEOR-a koja mapira riječi samo ukoliko su njihovi oblici riječi identični. Vjerojatno bi se veća vrijednost metrike METEOR mogla ostvariti uporabom WordNet leksičko semantičke mreže ili primjenom modula koji pri identificiranju mogućih podudaranja između strojnog i referentnog prijevoda uzima u obzir i korijen riječi.

Tablica 33. Performanse sustava za statističko strojno prevođenje temeljeno na frazama: **opća domena, englesko-hrvatski smjer (sustav 5)**.

metrika	**opća domena**
BLEU	0.0552
NIST	2.3918
METEOR	0.0995
GTM	0.2921
WER	0.8308
TER	0.8195

Tablica 34 prikazuje rezultate mjerenja performansi sustava za statističko strojno prevođenje broj 6, tj. sustava treniranog na specijaliziranom korpusu, tj. domeni računalnog softvera, za engleski-hrvatski smjer. Sustav 6 generirao je najbolje strojne prijevode za englesko-hrvatski smjer. Relativno visoka vrijednost BLEU (oko 0.3) odražava razumljive strojne prijevode. NIST vrijednost od oko 6.3 to potvrđuje. GTM vrijednost (0.6008) ne razlikuje se mnogo od GTM vrijednosti iz sustava 2 (0.6750). Iako ni u ovom slučaju nije postignuta METEOR vrijednost veća od 0.7, a koja upućuje na dobre i fluentne prijevode, vrijednost od 0.27 poprilično je visoka

ako se uzme u obzir kompleksnost domene na kojoj je sustav treniran (domena računalnog softvera). No, sustav 2 ipak je demonstrirao bolje performanse ako se analiziraju vrijednosti svih metrika. To potvrđuje činjenicu da je jednostavnije strojno prevoditi s morfološki bogatog jezika na morfološki manje bogate jezike. Drugim riječima, strojno prevođenje je efikasnije kad se prevodi s hrvatskog na engleski. Ipak, i kad je sustav 6 u pitanju hipoteza H1 se može potvrditi.

Tablica 34. Performanse sustava za statističko strojno prevođenje temeljeno na frazama: **specifična domena, tj. domena računalnog softvera, engleski-hrvatski smjer (sustav 6)**.

metrika	**domena računalnog softvera**
BLEU	0.2839
NIST	6.3419
METEOR	0.2732
GTM	0.6008
WER	0.5310
TER	0.4780

Tablica 35 prikazuje rezultate mjerenja performansi hibridnog sustava za statističko strojno prevođenje (sustav 7) koji koristi tablice prijevoda fraza, tj. fraznih struktura iz obje domene, a preostale značajke modela (leksičko premještanje, jezični model itd.) iz domene računalnog softvera, za englesko-hrvatski smjer. I u slučaju englesko-hrvatskog smjera hibridni sustav za strojno prevođenje dokazao je da je uporabom modela iz različitih domena, tj. i općenite i specifične, moguće ostvariti poboljšanje kvalitete strojnog prijevoda. Rezultati svih metrika bolji su u odnosu na rezultate sustava 5. Time je još jednom potvrđena hipoteza H2 da se ciljanom uporabom većeg broja domenski različitih značajki mogu poboljšati performanse sustava za strojno prevođenje.

Tablica 35. Performanse **hibridnog sustava** za statističko strojno prevođenje temeljeno na frazama: tablice prijevoda fraza, tj. fraznih struktura iz **obje domene**, a preostale značajke modela **iz specifične domene, tj. domene računalnog softvera, englesko-hrvatski smjer (sustav 7).**

metrika	**hibridni sustav**
BLEU	0.1008
NIST	3.0877

METEOR	0.1311
GTM	0.3555
WER	0.7806
TER	0.7720

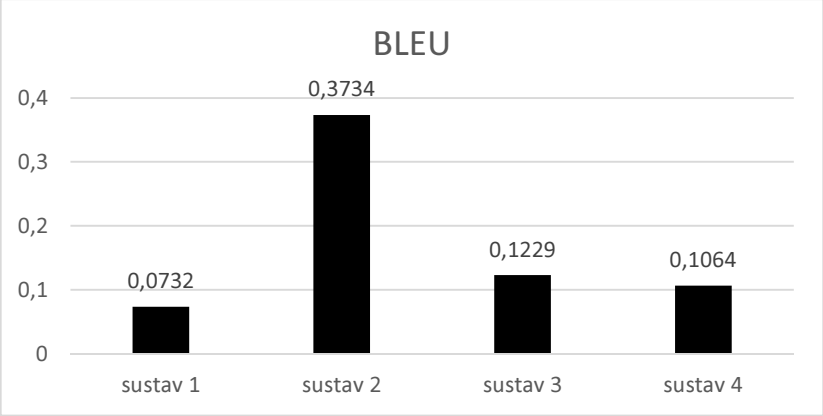
Tablica 36 prikazuje rezultate mjerenja performansi hibridnog sustava za statističko strojno prevođenje (sustav 8) koji koristi tablice prijevoda fraza, tj. fraznih struktura iz obje domene, a preostale značajke modela (leksičko premještanje, jezični model itd.) iz opće domene, za englesko-hrvatski smjer. Kao i u slučaju sustava 3 i 4, sustav koji koristi preostale značajke modela iz opće domene (sustav 8) demonstrirao je nešto slabije performanse od sustava 7 (preostale značajke modela iz domene računalnog softvera). Također, u svim segmentima evaluacije automatskim metrikama, sustav 8 se pokazao boljim u odnosu na sustav treniran na općoj domeni za englesko-hrvatski smjer. To ponovno potvrđuje hipotezu H2.

Tablica 36. Performanse **hibridnog sustava** za statističko strojno prevođenje temeljeno na frazama: tablice prijevoda fraza, tj. fraznih struktura iz **obje domene**, a preostale značajke modela iz **opće domene**, **englesko-hrvatski smjer (sustav 8)**.

metrika	**hibridni sustav**
BLEU	0.0870
NIST	2.8805
METEOR	0.1227
GTM	0.3405
WER	0.7877
TER	0.7822

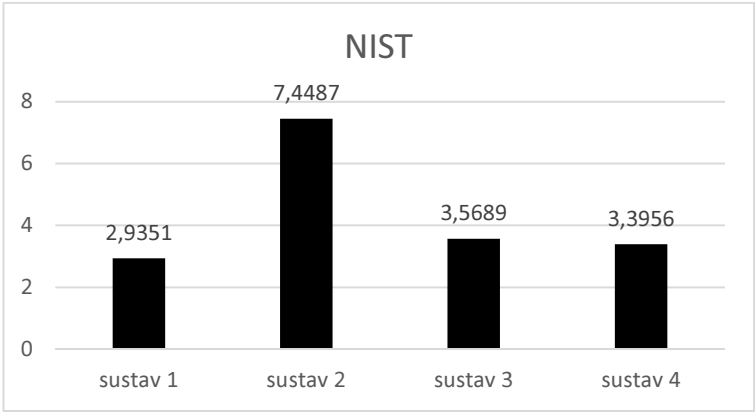
Ukoliko se usporede rezultati BLEU metrika prva četiri sustava, tj. onih koji pokrivaju hrvatsko-engleski smjer, vidljivo je da je daleko najbolji rezultat BLEU metrike postigao drugi sustav, tj. sustav treniran na specijaliziranom korpusu, tj. domeni računalnog softvera (Slika 22). To je očekivano, s obzirom da je sustav u potpunosti orijentiran prevođenju specifične domene, tj. sve značajke sustava su optimizirane za prevođenje teksta iz domene računalnog softvera. Interesantno je što su oba hibridna sustava postigla bolje rezultate u odnosu na sustav koji je treniran na podatkovnom skupu iz opće domene, pri čemu je sustav broj 3 (koristi značajke

modela iz specifične domene) bio bolji u usporedbi s hibridnim sustavom koji pri dekodiranju koristi preostale značajke iz opće domene.



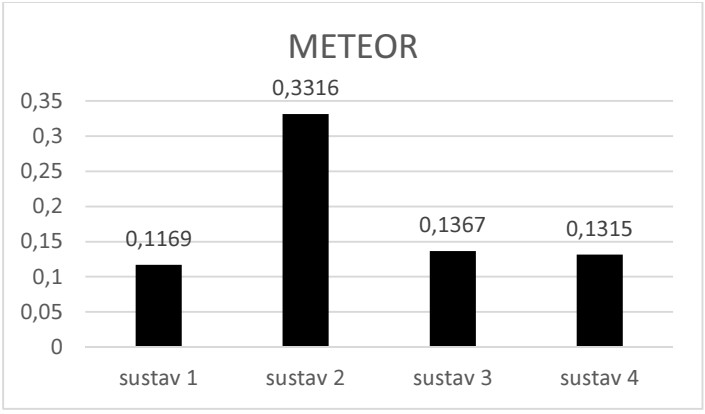
Slika 22. Usporedba BLEU rezultata za različite sustave za statističko strojno prevođenje za hrvatsko-engleski smjer (više je bolje).

Slika 23 prikazuje rezultate usporedbe NIST rezultata sustava za prevođenje s hrvatskog na engleski jezik. Budući da se NIST metrika temelji na BLEU-u, i ovaj ishod evaluacije metrikom NIST je očekivan. Sustav treniran na specifičnoj domeni polučio je najbolje rezultate, a slijede hibridni sustavi 3 i 4, te na začelju sustav 1, tj. sustav treniran na korpusu opće domene. Sustav treniran na domeni računalnog softvera ostvario je dvostruko bolje rezultate u odnosu na najlošiji sustav.



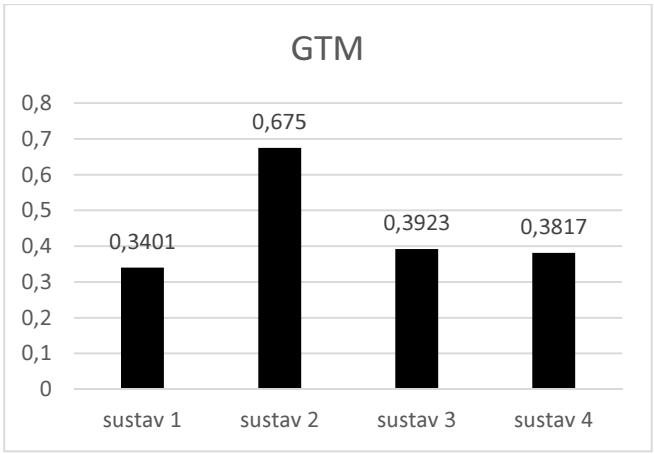
Slika 23. Usporedba NIST rezultata za različite sustave za statističko strojno prevođenje za hrvatsko-engleski smjer (više je bolje).

Slika u nastavku prikazuje METEOR-ove rezultate evaluacije sustava za statističko strojno prevođenje za hrvatsko-engleski smjer. Trostruko bolje prošao je sustav treniran na domeni računalnog softvera, a i u ovom slučaju ga slijede hibridni sustavi 3 i 4, a najlošije je prošao sustav broj 1 (Slika 24).



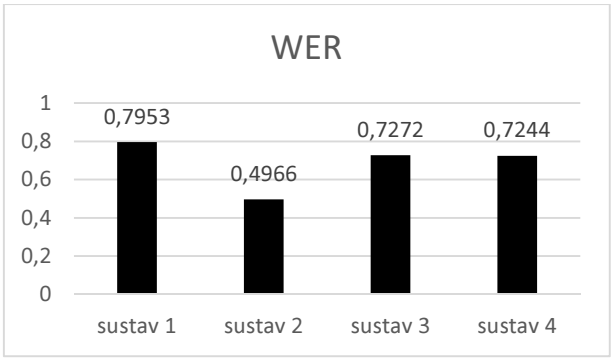
Slika 24. Usporedba METEOR rezultata za različite sustave za statističko strojno prevođenje za hrvatsko-engleski smjer (više je bolje).

Slika 25 upućuje na to da je i u pogledu metrike GTM sustav 2, tj. sustav treniran na domenski specifičnom korpusu prošao najbolje. Najlošije je očekivano prošao sustav broj 1, tj. sustav koji je treniran na općem korpusu. Hibridni sustavi su i u slučaju GTM-a bolji su u odnosu na sustav 1.



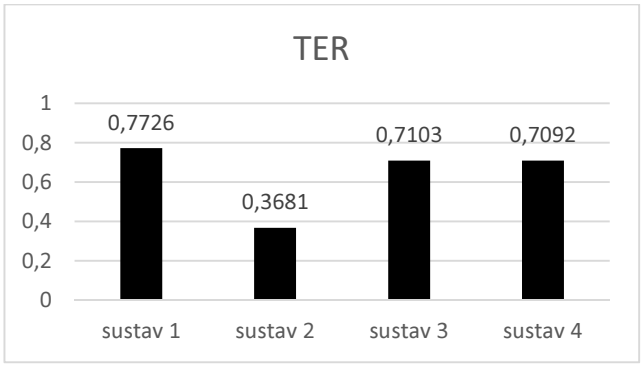
Slika 25. Usporedba GTM rezultata za različite sustave za statističko strojno prevođenje za hrvatsko-engleski smjer (više je bolje).

Slika 26 prikazuje rezultate metrike WER-a za različite sustave za statističko strojno prevođenje za hrvatsko-engleski smjer. Sustav broj 2 i po pitanju WER metrike prošao je najbolje, a slijede ga hibridni sustavi 4 i 3 koji koriste izvandomenske značajke. Sustav koji koristi isključivo domenske značajke (sustav broj 1) prošao je najgore pri evaluaciji kvalitete strojno prevedenog podatkovnog skupa za ispitivanje, tj. testiranje.



Slika 26. Usporedba WER rezultata za različite sustave za statističko strojno prevođenje za hrvatsko-engleski smjer (manje je bolje).

Slika 27 prikazuje rezultate TER metrike za hrvatsko-engleski jezični smjer. I u ovom slučaju sustav treniran na domenski specifičnom podatkovnom skupu postigao je najbolji rezultat, tj. dvostruko bolji rezultat. Očekivano najlošiji rezultat ostvario je sustav treniran na općem korpusu. Kada se promatraju rezultati TER-a, interesantno je što su, unatoč svim ostalim metrikama, hibridni sustavi 3 i 4 samo neznatno bolji od sustava 1 (kao i u slučaju metrike WER).

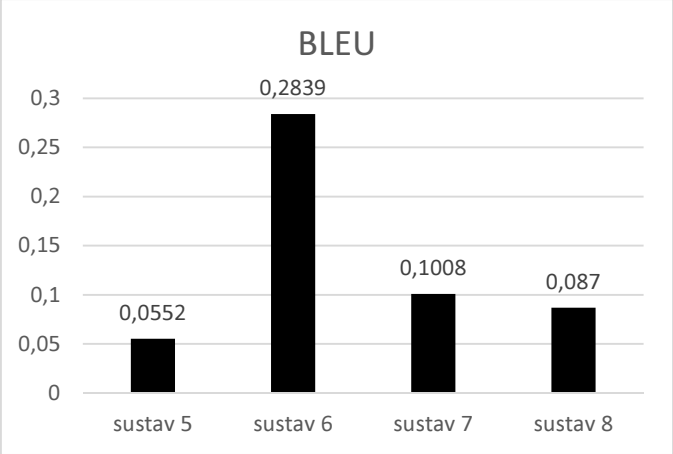


Slika 27. Usporedba TER rezultata za različite sustave za statističko strojno prevođenje za hrvatsko-engleski smjer (manje je bolje).

Time je i potvrđen prvi dio druge hipoteze (H2) doktorskog rada, a koji se odnosi na hrvatsko-engleski jezični smjer. Tj. primjenom hibridnog sustava za statističko strojno prevođenje temeljeno na frazama za hrvatsko-engleski smjer može se postići poboljšanje kvalitete strojnog prijevoda u odnosu na sustave koji koriste značajke svojstvene isključivo jednoj domeni (sustav 1). Međutim, to se ne može tvrditi za sustav 2.

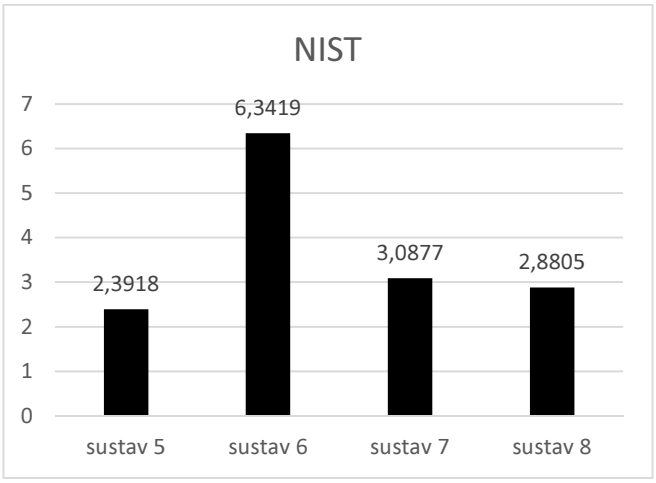
Jednako tako, s obzirom da su automatskim metrikama najbolje performanse evidentirane u sustavu 2, prvi dio prve hipoteze (H1) također se može potvrditi. Drugim riječima, adaptacija, tj. (prilagodba) domene i karakteristike ulaznoga podatkovnog skupa utječu na kvalitetu strojnog prijevoda, tj. na performanse sustava za statističko strojno prevođenje temeljeno na frazama za hrvatsko-engleski jezični par.

Slika 28 prikazuje rezultate evaluacije sustava za statističko strojno prevođenje primjenom metrike BLEU. I u slučaju englesko-hrvatskog smjera trend se nastavlja: najbolji BLEU rezultat postigao je sustav broj 6, tj. sustav treniran na domeni računalnog softvera. Najlošije je bodovan sustav treniran na općoj domeni (sustav 5). Hibridni sustavi 7 i 8 također su bolji u odnosu na sustav 5.



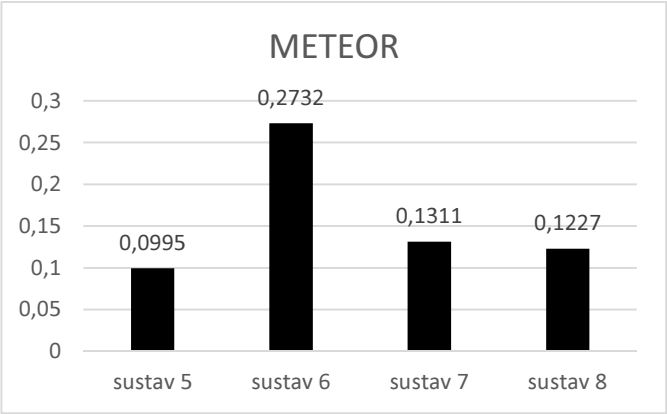
Slika 28. Usporedba BLEU rezultata za različite sustave za statističko strojno prevođenje za engleski-hrvatski smjer (više je bolje).

Slika 29 prikazuje rezultate NIST metrike za englesko-hrvatski smjer. Najlošije je bodovan sustav opće domene, a najbolje sustav specifične domene. Hibridni sustavi 7 i 8 također su bolji u odnosu na sustav 5.



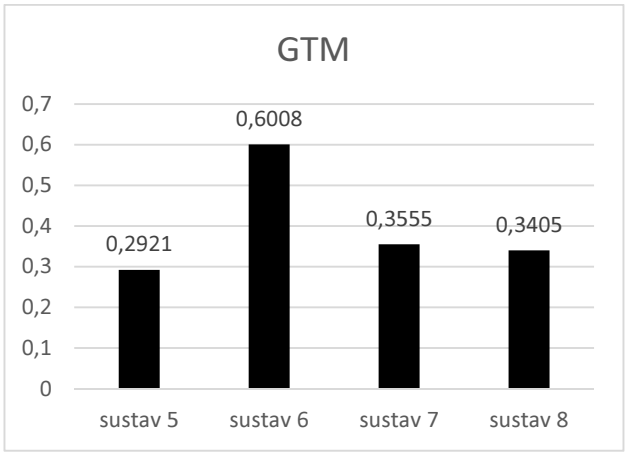
Slika 29. Usporedba NIST rezultata za različite sustave za statističko strojno prevođenje za engleski-hrvatski smjer (više je bolje).

Naredna slika prikazuje rezultate evaluacije metrikom METEOR. Najbolji rezultat postigao je opet sustav 6, a slijede ga hibridni sustavi (Slika 30). Na začelju je sustav treniran na općoj domeni.



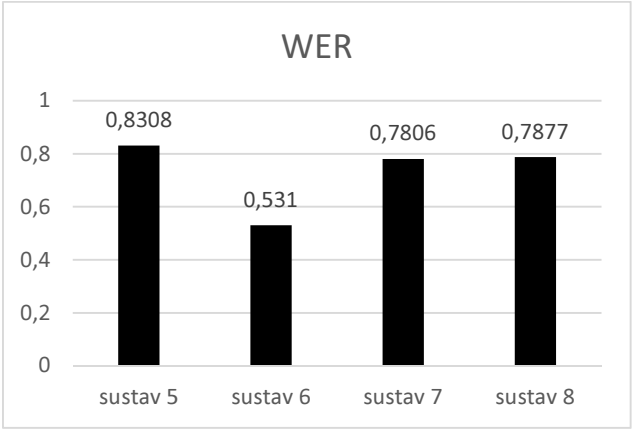
Slika 30. Usporedba METEOR rezultata za različite sustave za statističko strojno prevođenje za engleski-hrvatski smjer (više je bolje).

Slika 31 potvrđuje najbolje performanse sustava koji je treniran na specifičnoj domeni (sustav 6). Hibridni sustavi 7 i 8 također su i u slučaju englesko-hrvatskog smjera bolji u odnosu na sustav koji je treniran na općem podatkovnom skupu.



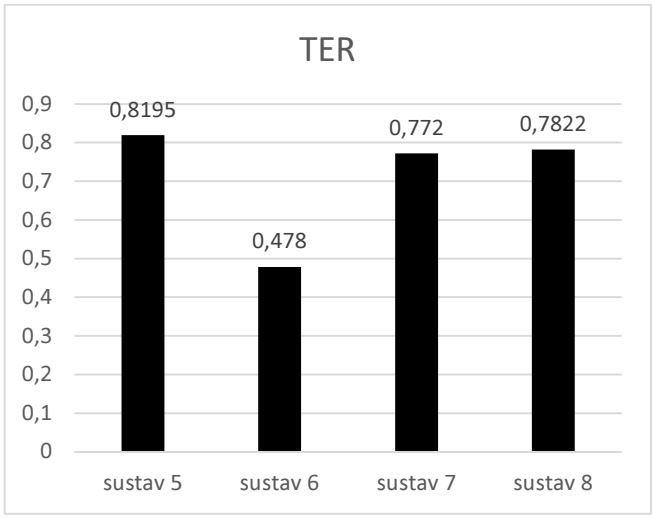
Slika 31. Usporedba GTM rezultata za različite sustave za statističko strojno prevođenje za engleski-hrvatski smjer (više je bolje).

Slika 32 prikazuje rezultate WER metrike primijenjene na sustavima za statističko strojno prevođenje koji prevode s engleskog jezika na hrvatski. WER metrika također ukazuje na to da je sustav koji je treniran na specifičnoj domeni bolji u odnosu na ostale sustave. Slijede ga hibridni sustavi 7 i 8, dok je najslabije bodovan sustav treniran na općoj domeni.



Slika 32. Usporedba WER rezultata za različite sustave za statističko strojno prevođenje za engleski-hrvatski smjer (manje je bolje).

Slika 33 prikazuje rezultate TER metrike. Sustav treniran na domeni računalnog softvera ocijenjen je bolje u odnosu na ostale sustave. TER vrijednost sustava 6 znatno je niža u odnosu na sustave 5, 7 i 8. Ipak, hibridni sustavi bolji su u odnosu na sustav 5.



Slika 33. Usporedba TER rezultata za različite sustave za statističko strojno prevođenje za engleski-hrvatski smjer (manje je bolje).

Konačno, komparativnom analizom automatskih metrika utvrđeno je da se i drugi dio druge hipoteze (H2) (koja se odnosi na engleski-hrvatski smjer) može potvrditi. Ispostavilo se da se primjenom hibridnih sustava (sustavi 3, 4, 7 i 8) koji koristi značajke iz obje domene, tj. opće domene i domene računalnog softvera, za hrvatsko-engleski i englesko-hrvatski jezični par može postići poboljšanje kvalitete strojnog prijevoda u odnosu na sustave koji koriste značajke svojstvene isključivo jednoj domeni, u ovom slučaju sustave 1 i 5. To je dovoljno za potvrđivanje hipoteze H2, međutim, treba naglasiti da to nije slučaj sa sustavima 2 i 6.

Nadalje, s obzirom da su automatskim metrikama najbolje performanse evidentirane u sustavima 2 i 6, a koji su trenirani isključivo na specifičnoj domeni i koji ne koriste izvandomenske značajke, prva hipoteza (H1) također se može potvrditi. To znači da se adaptacijom domene i primjenom karakterističnog, tj. domenski specifičnog ulaznog podatkovnog skupa može utjecati na povećanje kvalitete strojnog prijevoda, tj. na performanse sustava za statističko strojno prevođenje temeljeno na frazama za hrvatsko-engleski i englesko-hrvatski jezični par.

Kako bi se prethodno navedene tvrdnje potvrdile i na statičko značajnoj razini, korištena je posebna metoda za određivanje statističke značajnosti razlike između sustava za strojno prevođenje. Radi se o metodi koja otkriva intervale pouzdanosti koji se zatim mogu međusobno uspoređivati. Odnosno, da bi se utvrdila statistička značajnost razlika među izgrađenim sustavima, primijenjena je metoda ponovnog uzorkovanja s ponavljanjem. Tom metodom moguće je odrediti intervale pouzdanosti, što je i učinjeno u ovom doktorskom radu.

Budući da je pri izgradnji, tj. optimizaciji sustava (ugađanju) korištena metrika BLEU, metoda ponovnog uzorkovanja također je primijenjena na BLEU metrici. Upotrijebljene su BLEU vrijednosti na razini rečenica, a ne na razini korpusa. Korišteni su strojno prevedeni podatkovni skupovi za ispitivanje, tj. testiranje (njih 8, jedan po sustavu) koji su se sastojali od po 1000 rečenica, tj. segmenata.

Metodom ponovnog uzorkovanja s ponavljanjem generiran je velik broj „umjetnih" skupova BLEU vrijednosti koji su se također sastojali od po 1000 nasumičnih BLEU vrijednosti, a koje su proizlazile iz izvornog skupa BLEU vrijednosti. Vrlo je malena vjerojatnost da su novonastali nasumični „umjetni" skupovi bili identični izvornom skupu BLEU vrijednosti. Taj postupak se u ovom doktorskom radu ponavljao 10000 puta, zadana razina značajnosti iznosila je 95% te je svaki put računata aritmetička sredina BLEU vrijednosti.

Stoga su intervali pouzdanosti predstavljali intervale koji sadrže najmanje 95% vrijednosti iz uzoraka. Tablica u nastavku prikazuje rezultate analize *bootstrapanih* podatkovnih skupova (Tablica 37). Najveća aritmetička sredina u *bootstrapanim* podatkovnim skupovima zabilježena je u sustavu 2 za hrvatsko-engleski smjer, te u sustavu 6 za englesko-hrvatski smjer. Kada se analiziraju intervali pouzdanosti, sustavi 1 i 2, 1 i 3, 2 i 3, 2 i 4 međusobno su različiti na statistički značajnoj razini. Očekivano, za sustave 1 i 4 te 3 i 4 to se ne može tvrditi sa sigurnošću. Razlike su također statistički značajne za sustave 5 i 6, 5 i 7, 6 i 7 te 6 i 8 (analogno hrvatsko-engleskom smjeru), dok se za sustave 5 i 8 te 7 i 8 to ne može sa sigurnošću tvrditi.

Tablica 37. Rezultati analize *bootstrapanih* podatkovnih skupova.

hrvatsko-engleski sustavi	**sustav 1**	**sustav 2**	**sustav 3**	**sustav 4**
aritmetička sredina	0.2038	0.4615	0.2439	0.2347
standardna devijacija	0.0076	0.0096	0.0087	0.0088
granica pogreške	0.0153	0.0192	0.0175	0.0175
interval pouzdanosti	[0.1885, 0.2191]	[0.4423, 0.4807]	[0.2264, 0.2614]	[0.2172, 0.2522]
englesko-hrvatski sustavi	**sustav 5**	**sustav 6**	**sustav 7**	**sustav 8**
aritmetička sredina	0.1940	0.4153	0.2302	0.2216
standardna devijacija	0.0075	0.0100	0.0087	0.0085
granica pogreške	0.0151	0.0200	0.0174	0.0170
interval pouzdanosti	[0.1789, 0.2090]	[0.3954, 0.4354]	[0.2128, 0.2476]	[0.2046, 0.2386]

Interesantno je što se i za sustave 5 (treniran na općenitoj domeni) i 8 to ne može tvrditi (koristi preostale značajke iz općenite domene). Bez obzira na to, metodom ponovnog uzorkovanja još jednom je potvrđena druga hipoteza (H2) u ovom doktorskom radu pa makar i samo na jednom paru sustava. Tj. primjenom hibridnog sustava za statističko strojno prevođenje temeljeno na frazama koji koristi različite domenske značajke može se postići poboljšanje kvalitete strojnog prijevoda u odnosu na sustave koji koriste značajke svojstvene isključivo jednoj domeni. To je utvrđeno za oba jezična smjera (hrvatsko-engleski i englesko-hrvatski jezični par) na sustavima 1, 3 i 4 te 5, 7 i 8. Naime, računalna adaptacija domene u hibridnim sustavima ostvarena je pomoću metode alternativnog puta dekodiranja koja je zaslužna za veće vrijednosti evaluacijske metrike BLEU u odnosu na sustave koji koriste isključivo domenske značajke.

No, treba konstatirati da se hipoteza ne može potvrditi u slučaju sustava 2, 3 i 4 te 6, 7 i 8, budući da su sustavi 2 i 6 (domenski specifični sustavi) demonstrirali daleko veću kvalitetu strojnih prijevoda i bolje performanse.

Nadalje, s obzirom da je i na statističko značajnoj razini utvrđeno da su sustavi 2 i 6, tj. sustavi trenirani na domenski specifičnom korpusu (računalni softver) različiti u odnosu na preostale sustave prva hipoteza ovog doktorskog rada (H1) također je u potvrđena za oba smjera. Adaptacijom domene, koja je u ovom slučaju ostvarena pomnim odabirom te karakteristikama pretprocesiranog ulaznog podatkovnog skupa utječe se na kvalitetu strojnog prijevoda, tj. na performanse sustava za statističko strojno prevođenje temeljeno na frazama za hrvatsko-engleski i englesko-hrvatski jezični par.

Treba napomenuti da se pod pojmom granica pogreške (eng. *margin of error, MOE*) podrazumijeva mjera koja odaje disperziju dobivenih vrijednosti (podataka) u okolini očekivane vrijednosti uzorka, pri čemu je polovina njene vrijednosti jednaka standardnoj devijaciji. Grafički prikazi, tj. histogrami BLEU vrijednosti dobivenih metodom ponovnog uzorkovanja također su prikazani u Dodatku A.

S obzirom da se radi o relativno velikim uzorcima, za provjeru razdiobe, tj. testiranje normalne distribucije primijenjeni su testovi Lilliefors (Kolmogorov-Smirnov), Anderson-Darling, Cramer-von Mises te Pearson chi-kvadrat (Dodatak A). Treba napomenuti da vrijednosti D, A, W i P predstavljaju testne statistike izračunate na temelju uzorka i njegovih mjera (aritmetička sredina, varijanca) koje za pojedini test imaju funkciju projiciranja mjera uzorka u pretpostavljenu distribuciju s ciljem određivanja pripadnosti uzorka toj distribuciji (Dodatak A). Ukoliko testna statistika upadne u kritično područje specificirano za svaki test, odbacuje se nul-hipoteza. Nul-hipoteza u ovim slučajevima je pripadnost normalnoj distribuciji.

P vrijednost predstavlja najmanju vrijednost pri kojoj se nul-hipoteza (distribucija je normalna) može odbaciti. Ako je p vrijednost veća ili jednaka 0.05 (≥0.05), nul-hipoteza (distribucija je normalna) se prihvaća. Ako je p vrijednost manja od 0.05 (<0.05) nul-hipoteza se odbacuje, tj. ne radi se o normalnoj distribuciji.

S obzirom da su samo *bootstrapani* uzorci sustava 1, 2, 6, 7, i 8 sigurno normalno distribuirani (Dodatak A), nad njima je provedena dodatna analiza skupova pomoću z-testa (Tablica 38). Uzorci sustava 3, 4 i 5 nisu normalno distribuirani i stoga nisu podvrgnuti z-testu. Kako se radi o vrlo velikim uzorcima, nije upotrijebljen t-test, već je korišten asimptotski normalni z-test, jer dobro aproksimira t-test za velike uzorke (Bluman, 2009).

Ukoliko su dana dva nezavisna uzorka duljina $\boldsymbol{n_1}$ i $\boldsymbol{n_2}$ iz normalno distribuiranih populacija, može se testirati $\boldsymbol{H_0}$ kao što je prikazano u nastavku (4.1 i 4.2) (Bluman, 2009; Huzak, 2006).

$$H_0: \mu_1 - \mu_2 = \delta_0 \tag{4.1}$$

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{4.2}$$

Radi se o asimptotskom testu, s obzirom da z-test pretpostavlja varijance iz izvornog uzorka, tj. standardne devijacije. No, umjesto stvarnih varijanci iz izvornog uzorka, korištene su procijenjene (*bootstrapane*), tzv. uzoračke varijance. One su samo aproksimacija, međutim, kada je uzorak velik, uzoračka standardna devijacija (iz *bootstrapanog* uzorka) aproksimira potrebnu populacijsku standardnu devijaciju. U z-testu upotrijebljen je $\boldsymbol{\delta_0 = 0}$, s obzirom da se testira jednakost očekivanih vrijednosti metrika za dva sustava koja se ispituju. Z-test proveden je na sustavima 1 i 2, 6 i 7, 6 i 8, te 7 i 8. Rezultati z-testa dani u nastavku (Tablica 39).

Tablica 38. Rezultati z-testa na izgrađenim sustavima.

hrvatsko-engleski		
sustavi: 1 i 2	-2098.9746	nul-hipoteza odbačena: sustavi su statistički značajno različiti jer su im očekivane vrijednosti različite

englesko-hrvatski		
sustavi: 6 i 7	1396.5796	nul-hipoteza odbačena: sustavi su statistički značajno različiti jer su im očekivane vrijednosti različite
sustavi: 6 i 8	1476.4055	nul-hipoteza odbačena: sustavi su statistički značajno različiti jer su im očekivane vrijednosti različite
sustavi: 7 i 8	70.6428	nul-hipoteza odbačena: sustavi su statistički značajno različiti jer su im očekivane vrijednosti različite

Z-testom se provjerava jednakost očekivanih vrijednosti za dva sustava. Ukoliko testna statistika upadne u kritično područje <-∞,-1.96> i <1.96,∞> (očitano u tablici vrijednosti funkcije distribucije jedinične normalne razdiobe), nul-hipoteza (da sustavi imaju jednaku očekivanu vrijednost) se odbacuje te iz toga proizlazi da je razlika među sustavima statistički značajna, tj. da sustavi nisu jednaki.

Gornja tablica potvrđuje da su sustavi 1 i 2, 6 i 7, 6 i 8 te 7 i 8 različiti na statistički značajnoj razini (Tablica 39). Z-testom nije potvrđena statistička značajnost razlike sustava 1 i 3, 2 i 3, 2 i 4, 5 i 6 te 5 i 7, a koje su potvrđene analizom intervala pouzdanosti, s obzirom da *bootstrapane* BLEU vrijednosti sustava 3, 4 i 5 nisu bile normalno distribuirane. No, zato se razlika može potvrditi za 7 i 8 što nije bilo moguće analizom intervala pouzdanosti.

5.7. Evaluacija kvalitete strojnog prijevoda

U ovom poglavlju dani su rezultati evaluacije kvalitete strojnih prijevoda generiranih pomoću izgrađenih sustava za statističko strojno prevođenje temeljeno na frazama. Nadalje, provedena je komparativna analiza kvalitete strojnih prijevoda s obzirom na postojeće web servise za statističko strojno prevođenje. Naime, preostalo je još provjeriti treću hipotezu (H3).

Ljudskom evaluacijom analizirani su prvo različiti tipovi pogrešaka u strojnom prijevodu. Pogreške su načelno podijeljene u 5 kategorija: točnost, jezik, terminologija, stil i specifični standardi[21] (Tablica 39). Pomoću 100 referentnih rečenica evaluirano je 100 prijevoda po izgrađenom sustavu, tj. ukupno je evaluirano 800 strojnih prijevoda. Referente rečenice proizlaze iz skupa za ispitivanje (testiranje) te su odabrane nasumično i ekstrahirane pomoću skriptnog jezika sed.

Tablica 39. Klasifikacija pogrešaka u strojnom prijevodu.

točnost	• netočan prijevod, tj. netočna interpretacija izvorne rečenice • nerazumijevanje koncepta, tj. poruke izvorne rečenice • dvoznačan prijevod • ispuštanje (bitni dijelovi iz izvorne rečenice nedostaju u prijevodu) • dodavanje (nepotrebni elementi u prijevodu, a koji nužno ne postoje u izvornoj rečenici) • neprikladan prijevod (iako prevedene sve riječi u rečenici) • neprevedeni dijelovi
jezik	• gramatika: sintaksa, nepridržavanje pravila ciljnog jezika • interpunkcijski znakovi: nepridržavanje pravila ciljnog jezika • pravopis: pogreške, naglasci, dijakritički znakovi, veliko ili malo slovo
terminologija	• nepridržavanje standardne terminologije • nepridržavanje ostalih propisanih terminoloških standarada • nekonzistentna terminologija

[21] adaptacija DQF (eng. *Dynamic Quality Evaluation Framework*) klasifikacije pogrešaka u strojnom prijevodu

stil	• nepridržavanje standarada struke • nekonzistentnost s izvornom rečenicom • nekonzistentnost unutar prijevoda • doslovni prijevod • neobična sintaksa • nedostatak idiomatskog značenja u prijevodu • prizvuk
specifični standardi	• datumi • mjerne jedinice • valuta • granične oznake • adrese • brojevi telefona • poštanski brojevi • tipkovni prečaci • kulturne razlike

U nastavku su rezultati analize pogrešaka u strojnim prijevodima (Tablica 40 i 41). Vidljivo je da u sustavima 1-4 jasno dominiraju pogreške „točnosti", a koje se uglavnom odnose na netočne prijevode, ispuštanje ili dodavanje riječi. Jednako tako, sustavi za strojno prevođenje određene riječi ostavljaju neprevedene, budući da uslijed nedostatka unosa u tablici prijevoda, tj. fraznih struktura za takve riječi ne postaji statistički opis (eng. *sparse data*). Od ukupno 2536 grešaka u sustavima 1-4 evidentirano je 1570 grešaka tipa „točnost" (62%), a u sustavima 5-8 evidentirano je ukupno 2725 grešaka, pri čemu greške tipa „točnost" čine oko 58% (njih 1558). Pogreške tog tipa ponajviše utječu na ljudsku percepciju kvalitete strojnog prijevoda, s obzirom da neprevedene riječi izrazito umanjuju mogućnost razumijevanja izvorne poruke. Pod pogreškom tipa „jezik" podrazumijeva se manjkavost gramatike, tj. konstrukcija rečenica nije prema pravilima standardnog jezika. Tip „jezik" na drugom je po brojnosti grešaka u izgrađenim sustavima za strojno prevođenje.

Greške u interpunkcijskim znakovima bile su manje zastupljene, a pravopisne pogreške skoro i nepostojeće. Tip pogrešaka „terminologija" zastupljen je u manjoj mjeri u sustavima za prevođenje s hrvatskog na engleski, a odnosi se na nepridržavanje standardne terminologije ili općih tehnoloških standarada. Očekivano, sustav treniran na specifičnoj domeni sadržavao je znatno manje pogrešaka u usporedbi sa sustavom 1. Oko 3 puta manje pogrešaka u terminologiji bilo je u hibridnim sustavima (3 i 4) koji su pokazali svoju snagu kada se usporede sa sustavom 1, tj. sustavom treniranim na općenitoj domeni koja ne sadrži domenski specifične konstrukcije i vokabular.

Naime, primjena obje tablice prijevoda fraza, tj. fraznih struktura zaslužna je za daleko manji broj pogrešaka u specifičnoj terminologiji iz domene računalnog softvera. Nekonzistentnost u terminologiji nije toliko izražena u statističkom pristupu strojnom prevođenju. Tip pogreške „stil" odnosi se uglavnom na nekonzistentnost s izvornom rečenicom, doslovne prijevode ili vrlo neobičnu sintaksu, odnosno poredak riječi u rečenici. Stilske pogreške najrjeđe su u sustavu 2, a slijede ga općeniti sustav te hibridni sustav. Tip pogreške „specifični standardi" odnosi se na pogreške u interpretaciji (i doslovnom prijevodu) tipkovnih prečaca. Ukupno gledajući, broj pogrešaka u sustavu 2 cca. dvostruko je manji u odnosu na sustav 1 i hibridne sustave 3 i 4. Treba uočiti da je i ukupan broj pogrešaka u hibridnim sustavima manji u odnosu na sustav treniran na općenitoj domeni (sustav 1). Time su hipoteze H1 i H2 ponovno potvrđene za hrvatsko-engleski smjer strojnog prevođenja.

Tablica 40. Rezultati analize pogrešaka u strojnim prijevodima generiranim pomoću hrvatsko-engleskih sustava za strojno prevođenje.

hrvatsko-engleski sustavi	**sustav 1** BLEU: 0.0732	**sustav 2** BLEU: 0.3734	**sustav 3** BLEU: 0.1229	**sustav 4** BLEU: 0.1064
točnost	456 (57%)	184 (58%)	467 (67%)	463 (64%)
jezik	178 (22%)	105 (33%)	133 (19%)	144 (20%)
terminologija	95 (12%)	18 (6%)	29 (4%)	36 (5%)
stil	67 (8%)	9 (3%)	69 (10%)	81 (11%)
specifični standardi	2 (<1%)	0	0	0
suma	798 (31%)	**316 (12%)**	698 (28%)	724 (29%)
ukupno	2536			

Kada se analiziraju rezultati u nastavku sličan trend može se uočiti i u slučaju englesko-hrvatskih strojnih prijevoda (Tablica 41), pri čemu su sustavi 5-8 ukupno generirali dvjestotinjak pogrešaka više u odnosu na sustave 1-4. Sustav treniran na specifičnom korpusu (sustav 6) sadrži dvostruko manje pogrešaka tipa „točnost". Razlika u broju pogrešaka tipa „točnost" između sustava 6 i hibridnih sustava (sustavi 7 i 8) je oko 80%. Pogreške tipa „jezik" također su prema zastupljenosti u ukupnom broju pogrešaka u strojnim prijevodima zauzele drugo mjesto u sustavima za engleski-hrvatski smjer. „Terminologija" je očekivano najlošija u općenitom sustavu (sustav 5), a najbolja u sustavu treniranom na korpusu iz specifične domene (3 puta manje pogrešaka u odnosu na sustav općenite domene).

Kada se usporede sustavi 5, 7 i 8, može se konstatirati da su stilske pogreške nešto su brojnije u hibridnim sustavima, a odnose se na uglavnom na doslovne prijevode i neobičnu sintaksu (oko 20 pogrešaka više). Pogreške u prijevodu tipkovnih prečaca klasificirane su kao pogreške tipa „specifični standardi" i daleko su najbrojniji u sustavu 1, tj. sustavu koji koristi isključivo domenske značajke. Ta razlika je vrlo izražena (čak 20 puta više u odnosu na hibridne sustave 7 i 8). Isto tako treba naglasiti, da su hibridni sustavi i u slučaju englesko-hrvatskog smjera generirali ukupno manji broj pogrešaka, što također potvrđuje drugu hipotezu (H2). S obzirom da je sustav 6 prema svim tipovima pogrešaka generirano najmanji broj pogrešaka, H1 također je još jednom potvrđena.

Tablica 41. Rezultati analize pogrešaka u strojnim prijevodima generiranim pomoću englesko-hrvatskih sustava za strojno prevođenje.

englesko-hrvatski sustavi	**sustav 5** BLEU: 0.0552	**sustav 6** BLEU: 0.2839	**sustav 7** BLEU: 0.1008	**sustav 8** BLEU: 0.087
točnost	476 (55%)	176 (44%)	451 (62%)	455 (62%)
jezik	192 (22%)	178 (45%)	148 (20%)	151 (20%)
terminologija	101 (12%)	29 (7%)	38 (5%)	37 (5%)
stil	71 (8%)	15 (4%)	89 (12%)	95 (<13%)
specifični standardi	21 (>2%)	0	1 (1%)	1 (<1%)
suma	861 (32%)	**398 (<15%)**	727 (<27%)	739 (27%)
ukupno	2725			

Tablica u nastavku prikazuje primjere strojnih prijevoda generiranih pomoću 3 različita sustava za englesko-hrvatski smjer (Tablica 42). Odabrani su sustavi 5 (općenita domena), 6 (specifična domena) te 7 (hibridni sustav). Izvorne rečenice predstavljaju rečenice koje su bile sastavni dio podatkovnog skupa za ispitivanje (testiranje) te koje su korištene kao ulazni podatkovni skup za generiranje strojnih prijevoda. Referente rečenice predstavljaju „zlatni standard", tj. točne/željene prijevode. Strojni prijevodi su na hrvatskom jeziku, po 3 za svaku izvornu rečenicu na engleskom jeziku. Iz tablice je vidljivo da je sustav treniran na specifičnoj domeni (sustav 6) proizveo najkvalitetnije, tj. najupotrebljivije prijevode kada se usporede s pripadajućim referentnim prijevodima. Na temelju prikazanih primjera strojnih prijevoda vrlo je teško i nepouzdano procijeniti kvalitetu hibridnih sustava u usporedbi s drugim sustavima.

Tablica 42. Primjer strojnih prijevoda englesko-hrvatskih sustava.

englesko-hrvatski sustavi			**rečenice**
1	izvorna rečenica (engleski)		move to next search result and highlight it in the document
	referentni prijevod (hrvatski)		premještanje na sljedeći rezultat pretraživanja i njegovo isticanje u dokumentu
	strojni prijevod	sustav 5	se sljedeći pretražiti rezultat i highlight to u dokument
		sustav 6	prešli na sljedeći rezultat pretraživanja i ga istaknuli unutar dokumenta
		sustav 7	sljedeći pretražiti da move rezultat i highlight ga u dokument
2	izvorna rečenica (engleski)		move bookmarks out of a nested position
	referentni prijevod (hrvatski)		premještanje knjižnih oznaka iz ugniježđenog položaja
	strojni prijevod	sustav 5	se bookmarks iz nested položaj
		sustav 6	premještanje knjižne oznake izvan ugniježđenog položaj
		sustav 7	move van bookmarks od nested položaj
3	izvorna rečenica (engleski)		you can use http , ftp , and mailto protocols to define your link
	referentni prijevod (hrvatski)		da biste definirali vezu , možete koristiti protokole http , ftp i mailto .
	strojni prijevod	sustav 5	moћeљ koristiti http , ftp , i mailto protocols da definirali tvoj kariku .
		sustav 6	možete koristiti http , FTP i mailto protocols da biste definirali vaš vezu .
		sustav 7	možete upotrijebiti http , FTP i mailto protocols da definirali vaš link .

Ukupno je više pogrešaka zabilježeno u sustavima koji prevode s engleskog na hrvatski, tj. s morfološki manje bogatog jezika na morfološki bogat jezik. Nadalje, općenito su sustavi 1 i 5 generirali najveći broj pogrešaka. Uzrok tome može biti i inicijalna opservacija da su podatkovni skupovi za treniranje općenitih sustava sadržavali i manju količinu šuma (srodne jezike, ćirilično pismo itd.) (Tablica 42). Međutim, takav scenarij je čest pri izgradnji sustava za slavenske jezike, budući da je količina dostupnih korpusa ograničena. Isto tako, različitost podatkovnih skupova za treniranje (općenita domena), ugađanje i testiranje (domena računalnog softvera) implicira velik udio neviđenih riječi (eng. *out-of-vocabulary*), umanjenu efikasnost sravnjivanja riječi, nesrazmjer u

vokabularu te razlike u duljini i konstrukciji rečenica. Hrvatski i engleski jezik imaju znatne strukturne razlike, što također utječe na kvalitetu prijevoda s engleskog na hrvatski.

Naime, hrvatski jezik dozvoljava izvjesnu slobodu u poretku riječi, dok engleski jezik prati karakterističan uzorak „SVO" (subjekt-glagol-objekt, eng. *subject-verb-object*) s relativno fiksnim poretkom (Lopez, 2008). Uzorak „SVO" temeljni je poredak članova rečenice u hrvatskom jeziku i najčešći, međutim, estetski i arhaični uzorci su također vrlo česti, pogotovo u poeziji i određenim književnim rodovima (Seljan, 2004). Npr. „SVO" konstrukcija „Marko čita novu knjigu" se relativno jednostavno može pretvoriti u rječu „OSV" (objekt-subjekt-glagol, eng. *object-subject-verb*) konstrukciju „Novu knjigu Marko čita", s obzirom da se cijele frazne kategorije (NP, VP, PP itd.) mogu relativno slobodno premještati bez utjecaja na značenje rečenice. Takve inverzije i ostali tipovi permutacija mogu se okarakterizirati kao stilske promjene, međutim, moguće su uslijed morfološkog bogatstva hrvatskog jezika. Hrvatski jezik vrlo je flektivan jezik i primjenjuje deklinacijske završetke koji ukazuju na rod, broj i padež, izravne i neizravne objekte itd. Ipak, sintaktičke kategorije, tj. vrste riječi (imenice, glagoli, prijedlozi itd.) koje se pojavljuju na početku rečenice više su naglašene.

Time je i u ovom doktorskom radu potvrđeno da prevođenje s manje složenog jezika (engleski) na više složeni jezik (hrvatski) također bitno utječe na kvalitetu strojnog prijevoda. To se odražava u rezultatima automatskih metrika i broju pogrešaka u strojnim prijevodima, s obzirom da sustavi za statističko strojno prevođenje općenito imaju izvjesne poteškoće s potrebnim premještanjem riječi, tj. generiranjem ispravnog poretka riječi u rečenici.

Ljudsku evaluaciju prema kriterijima adekvatnosti i fluentnosti provela su 3 evaluatora različitih struka (informatika, ekonomija, politologija). Radi se o evaluatorima kojima ja hrvatski jezik materinji te koji tečno govore engleski jezik. Evaluacija je za oba kriterija provedena na skali od 1 (najlošije) do 4 (najbolje) na rečenicama koje su navedene u Dodatku B. Treba ponoviti, da fluentnost ukazuje na razinu gramatičke ispravnosti, pravopis, pridržavanje općih jezičnih standarada te ustaljenju uporabu specifičnih termina i imena. Intuitivno je prihvatljiva i izvorni govornici ju vrlo lako mogu interpretirati. Adekvatnost ukazuje na količinu prenesenog značenja iz izvornog jezika u ciljni. Drugim riječima, fluentnošću se ispituje koliko „prirodno" jedan strojni prijevod zvuči izvornom govorniku ciljnog jezika, dok se adekvatnošću mjeri količina informacije koja je prenesena iz izvorne rečenice u prijevod.

U nastavku su prikazani rezultati ljudske evaluacije strojnih prijevoda (Tablica 43 i 44). Iz tablice proizlazi da je sustav 2 ocijenjen kao najbolji sustav za prevođenje s hrvatskog na engleski jezik, s prosječnom ocjenom oko 3. Ukupna ocjena (zbroj prosjeka ocjena adekvatnosti i

fluentnosti, pri čemu je 8 maksimum) cca. 2.5 puta je veća u odnosu na sustav 1 i hibridne sustave 3 i 4.

Interesantno je što su hibridni sustavi po pitanju ljudske evaluacije prošli nešto slabije u odnosu na sustav 1, međutim, ta razlika je neznatna (oko 2%). I ljudskom evaluacijom potvrđena je hipoteza H1 za hrvatsko-engleski smjer, dok hipotezu H2 nije moguće sa sigurnošću ni potvrditi ni odbaciti.

Tablica 43. Ljudska evaluacija sustava za strojno prevođenje: hrvatsko-engleski smjer.

Ljudska evaluacija	**hrvatsko-engleski**			
	sustav 1	**sustav 2**	**sustav 3**	**sustav 4**
adekvatnost evaluator 1	1.2800	3.3200	1.3600	1.3200
adekvatnost evaluator 2	1.1200	2.9100	1.0500	1.0400
adekvatnost evaluator 3	1.2800	3.3200	1.2929	1.2626
fluentnost evaluator 1	1.2600	3.2100	1.1900	1.2300
fluentnost evaluator 2	1.1400	2.9300	1.1100	1.0800
fluentnost evaluator 3	1.2000	3.1300	1.1414	1.2020
adekvatnost prosjek	1.2267	**3.1833**	1.2343	1.2075
fluentnost prosjek	1.2000	**3.0900**	1.1471	1.1707
ukupno: adekvatnost + fluentnost (prosjeci)	2.4267	**6.2733**	2.3814	2.3782

I u slučaju englesko-hrvatskih sustava za strojno prevođenje, sustav treniran na specifičnom korpusu iz domene računalnog softvera ocijenjen je daleko najbolje (oko 2.5 puta), dok je najlošije ocijenjen sustav 5 (Tablica 44). Evaluatori su, međutim, u ovom slučaju strojne prijevode generirane pomoću hibridnih sustava 7 i 8 ocijenili bolje u odnosu na sustav 5, tj. sustav treniran na općenitoj domeni. Ponovno su i za engleski-hrvatski smjer potvrđene druga i treća hipoteza.

Tablica 44. Ljudska evaluacija sustava za strojno prevođenje: englesko-hrvatski smjer.

Ljudska evaluacija	**englesko-hrvatski**			
	sustav 5	**sustav 6**	**sustav 7**	**sustav 8**
adekvatnost evaluator 1	1.2100	3.2400	1.3300	1.3000

adekvatnost evaluator 2	1.0000	2.7600	1.0100	1.0200
adekvatnost evaluator 3	1.2500	3.2400	1.2700	1.2700
fluentnost evaluator 1	1.0300	2.8300	1.1500	1.1800
fluentnost evaluator 2	1.0100	2.5700	1.0000	1.0200
fluentnost evaluator 3	1.1100	2.8800	1.1400	1.1700
adekvatnost prosjek	1.1533	**3.0800**	1.2033	1.1967
fluentnost prosjek	1.0500	**2.7600**	1.0967	1.1233
ukupno: adekvatnost + fluentnost (prosjeci)	2.2033	**5.8400**	2.300	2.3200

Nakon provedbe ljudske evaluacije izvršena je analiza pomoću Cronbach alphe radi mjerenja razine (ne)slaganja evaluatora (Tablica 45), s obzirom da je jedno od nepoželjnih obilježja ljudske evaluacije subjektivna percepcija kvalitete prijevoda. Tom mjerom ispituje se konzistentnost među evaluatorima. Treba ponoviti da vrijednosti $\alpha \geq 0.9$ upućuje na visoku konzistentnost, $0.8 \leq \alpha < 0.9$ dobru konzistentnost, $0.7 \leq \alpha < 0.8$ prihvatljivu, $0.6 \leq \alpha < 0.7$ upitnu, $0.5 \leq \alpha < 0.6$ slabu i $\alpha < 0.5$ neprihvatljivu konzistentnost.

Kada se analiziraju hrvatsko-engleski sustavi (sustavi 1-4), vidljivo je da je najveća konzistentnost u ocjenjivanju adekvatnosti zabilježena u sustavu 2, a za fluentnost u hibridnom sustavu 3. Općenito, vrijednosti Cronbach alphe za hrvatsko-engleske sustave upućuju na dobru ili visoku konzistentnost. Razina slaganja evaluatora slabija je za englesko-hrvatske sustave. Naime, iako je za sustav 6 izmjerena najveća vrijednost Cronbach alphe (za oba kriterija), ostali sustavi su ocijenjeni s relativno slabom konzistentnošću. Najniža konzistentnost zabilježena je za kriterij fluentnosti za sustav 5, tj. sustav koji je treniran na općenitoj domeni. Što je uzrok tome potrebno je detaljnije analizirati u budućim istraživanjima.

Tablica 45. Vrijednosti Cronbach alphe.

Cronbach alpha	**adekvatnost**	**fluentnost**
hrvatsko-engleski		
sustav 1	0.8724	0.8960
sustav 2	**0.9244**	0.9303
sustav 3	0.8691	**0.9380**
sustav 4	0.8328	0.8995

englesko-hrvatski		
sustav 5	0.7135	0.4048
sustav 6	**0.9385**	**0.9446**
sustav 7	0.6782	0.6661
sustav 8	0.6599	0.7919
ukupno		
	sustavi (1-4)	**sustavi (5-8)**
adekvatnost	0.9692	0.9650
fluentnost	0.9769	0.9709

Treba ponoviti kako su u ovom doktorskom radu upotrijebljene sljedeće metrike: *Word Error Rate* (WER), *Translation Error/Edit Rate* (TER), *BiLingual Evaluation Understudy* (BLEU), *National Institute of Standards and Technology* (NIST), *Metric for Evaluation of Translation with Explicit ORdering* (METEOR) i *General Text Matcher* (GTM). Važna odlika automatskih metrika jest mogućnost rangiranja sustava za strojno prevođenje na približno isti način kao što bi to učinili ljudski evaluatori, no znatno brže. Međutim, treba spomenuti da su automatske metrike ipak relativno nepouzdane za kraće prijevode (Callison-Burch et al., 2006).

Iako je BLEU jedna od najčešće korištenih metrika, postoji mogućnost relativno slabe korelacije s ljudskom evaluacijom koja se provodi s obzirom na kriterije kojima se procjenjuje prikladnost i gramatička ispravnost strojnog prijevoda. Glavni razlog tome jest što BLEU ispuštanje sadržajnih riječi ne penalizira dodatno, tj. sve riječi se tretiraju jednako važnima (Koehn, 2011). Nadalje, BLEU evaluacija je efikasnija ukoliko se koristi veći broj referentnih prijevoda, što je posebno pogodno u slučaju prevođenja sinonima ili idiomatskih konstrukcija (Callison-Burch et al., 2006).

Metrike WER i TER spadaju u metrike za izračun broja pogrešaka (eng. *error measures*), dok ostale navedene metrike pripadaju skupini za izračun točnosti (eng. *accuracy measures*). Automatske metrike međusobno se razlikuju prema načinu mjerenja sličnosti i prema varijablama koje ispituju. Međutim, zajedničko svim metrikama jest da bolje rangiraju automatske prijevode, što je podudarnost s referentnim (ljudskim) prijevodom veća (Jurafsky et al., 2013.).

S obzirom da je za izračun korelacije metrika i ljudske evaluacije predložen Pearsonov koeficijent korelacije (Callison-Burch et al., 2010; Koehn, 2010), u nastavku je ispitana korelacija između automatskih evaluacijskih metrika te ljudske evaluacije za hrvatsko-engleski te englesko-hrvatski smjer. Pearsonova korelacija izračunata je na razini svih 8 sustava i ukupnog zbroja

prosjeka adekvatnosti i fluentnosti. Vrijednosti Pearsonovog koeficijenta korelacije dane su u Tablici 46.

Tablica 46. Pearsonova korelacija automatskih metrika s ljudskom evaluacijom.

Korelacija automatskih metrika s ljudskom evaluacijom	
BLEU-ljudska evaluacija	0.8009
NIST-ljudska evaluacija	0.7876
METEOR-ljudska evaluacija	0.7986
GTM-ljudska evaluacija	0.7797
WER-ljudska evaluacija	-0.7474
TER-ljudska evaluacija	-0.7892

Interpretacija vrijednosti Pearsonovog koeficijenta korelacije može se izvršiti prema tablici u nastavku (Tablica 47) (Udovičić et al., 2007). Na temelju izračuna korelacije može se tvrditi da automatske metrike i ljudska evaluacija vrlo dobro koreliraju. Najviše koreliraju BLEU i ljudska evaluacija, a najmanje WER i ljudska evaluacija. S porastom vrijednosti BLEU, NIST, METEOR i GTM metrika rastu i ocjene evaluatora, dok s porastom WER-a i TER-a opadaju ocjene evaluatora.

Tablica 47. Interpretacija vrijednosti Pearsonovog koeficijenta korelacije.

Interpretacija koeficijenta korelacije	
nikakva ili neznatna korelacija	r od ± 0.00 do ± 0.25
slaba korelacija	r od ± 0.25 do ± 0.50
umjerena korelacija	r od ± 0.50 do ± 0.75
vrlo dobra korelacija	r od ± 0.75 do ± 1
savršena pozitivna korelacija	r=1
savršena negativna korelacija	r=-1

U Dodatku C dana je deskriptivna statistika rezultata ljudske evaluacije kvalitete strojnih prijevoda, koja će se detaljnije analizirati u budućim istraživanjima. Dane su aritmetička sredina, standardna pogreška aritmetičke sredine (eng. *standard error of the mean, SEM*), medijan, mod,

standardna devijacija, varijanca, mjera spljoštenosti (eng. *kurtosis*), mjera asimetrije (eng. *skewness*), raspon (eng. *range*), minimum i maksimum.

Pritom treba napomenuti da mjera standardne pogreške aritmetičke sredine ($\boldsymbol{SEM = \frac{\sigma}{\sqrt{n}}}$) predstavlja standardnu devijaciju aritmetičkih sredina podijeljenu s duljinom uzorka (Ahn i Fessler, 2003), i ona se smanjuje kako $\boldsymbol{n}$ teži u beskonačnost (što ne vrijedi u slučaju same varijance).

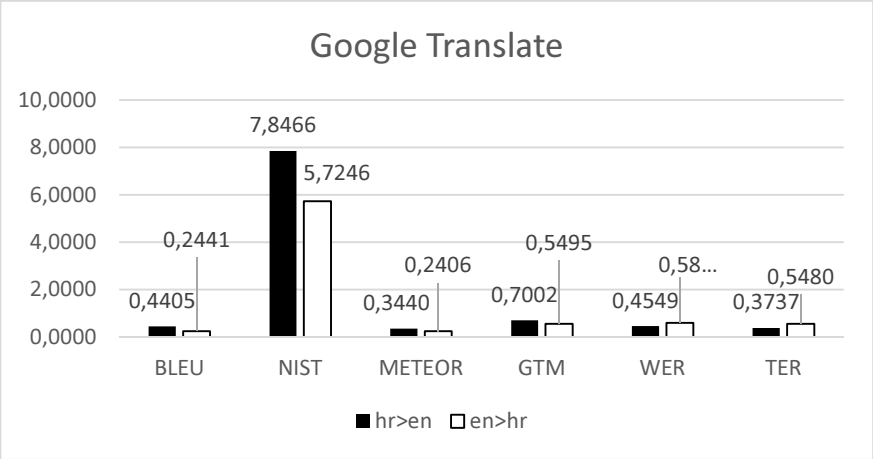
Nadalje, mjera spljoštenosti ukazuje na to da li distribucija vrijednosti odgovara normalnoj (Gaussovoj) razdiobi. Normalna distribucija ima spljoštenost 0, a spljoštenija distribucija ima negativnu vrijednost mjere spljoštenosti. Ukoliko je jedna distribucija šiljastija u odnosu na normalnu razdiobu, tada je i vrijednost mjere spljoštenosti pozitivna.

Simetrična distribucija ima asimetriju jednaku 0. Asimetrična distribucija s dugačkim repom prema desno (veće vrijednosti) ima pozitivnu asimetriju, dok distribucija s dugačkim repom prema lijevo (manje vrijednosti) ima negativnu asimetriju. Ukoliko je asimetrija veća od 1 ili -1 distribucija je uvelike asimetrična.

Nakon provedene ljudske evaluacije strojnih prijevoda dobivenih pomoću vlastitih sustava za statističko strojno prevođenje, generirani su i strojni prijevodi pomoću online prevodilačkih alata kako bi se međusobno usporedila kvaliteta strojnih prijevoda. Strojni prijevodi za oba jezična smjera generirani su pomoću dva web servisa: Google Translate i Yandex Translate u ožujku 2015.

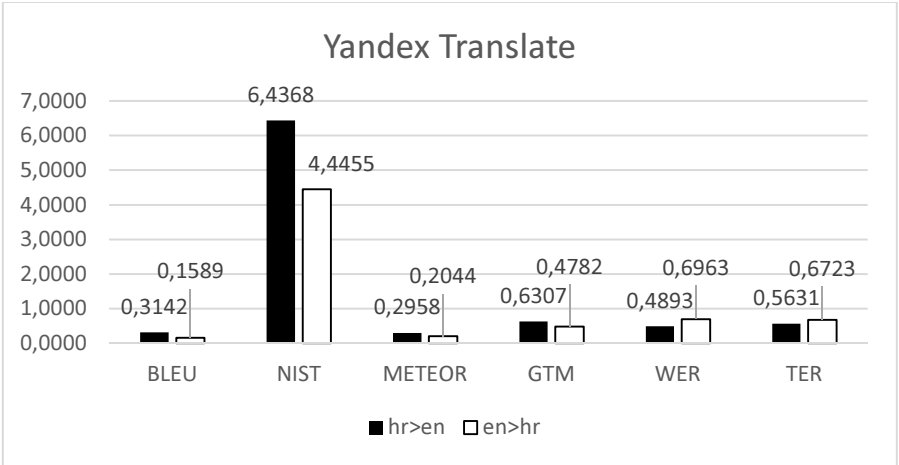
Radi se o vrlo različitim arhitekturama modela, no, s obzirom da se oba web servisa također temelje na statističkom strojnom prevođenju (Tohmetov et al., 2014; Koehn, 2010), usporedba strojnih prijevoda generiranih pomoću vlastitih sustava i prijevoda generiranih pomoću popularnih web servisa uvijek daje vrijedne povratne informacije koje mogu biti od velike koristi pri izgradnji sustava za strojno prevođenje. Pored toga, za usporedbu sličnih sustava ili različitih inačica istog sustava ponajviše se preporučuje primjena BLEU metrike (Callison-Burch et al., 2006).

Nakon generiranja strojnih prijevoda pomoću web servisa provedena je evaluacija kvalitete s obzirom na metrike BLEU, NIST, METEOR, GTM, WER i TER, kao i u slučaju evaluacije 8 ranije opisanih sustava za statističko strojno prevođenje. Za metrike BLEU, NIST, METEOR i GTM vrijedi „više je bolje“, dok za WER i TER vrijedi suprotno. U nastavku su prikazani rezultati evaluacije automatskim metrikama za Google Translate za oba smjera (Slika 34).



Slika 34. Rezultati evaluacije automatskim metrikama za sustav Google Translate za oba smjera.

Iz gornjeg grafikona proizlazi da je Google Translate pri prevođenju podatkovnog skupa za testiranje (ispitivanje) generirao prijevode relativno visoke kvalitete (Slika 34). I Google Translate postigao je bolje rezultate pri prevođenju s hrvatskog na engleski, tj. s kompleksnijeg na manje kompleksan jezik. BLEU od čak 0.4405 za hrvatsko-engleski upućuje na zaista kvalitetne i upotrebljive prijevode, a to potvrđuje i NIST od skoro 8. GTM vrijednost od 0.7 također je vrlo visoka, a predstavlja vrlo povoljan omjer preciznosti i odziva, s obzirom da ova metrika favorizira dulja preklapanja riječi s ispravnim poretkom. U nastavku su prikazani rezultati evaluacije automatskim metrikama za Yandex Translate za oba smjera (Slika 35).



Slika 35. Rezultati evaluacije automatskim metrikama za sustav Yandex Translate za oba smjera.

Rezultati ispitivanja strojnih prijevoda generiranih pomoću web servisa Yandex Translate slabiji su u odnosu na Google Translate za sve metrike (Slika 35). I u slučaju ovog web servisa rezultati su bolji za prevođenje s hrvatskog na engleski. BLEU od 0.31 upućuje također na relativno dobre strojne prijevode za hrvatsko-engleski smjer. To potvrđuje i NIST vrijednost koja je veća od 6. WER od cca. 0.7 za englesko-hrvatski ukazuje ipak na relativno malen broj riječi s ispravnim poretkom u rečenici, a to je potvrđeno i vrijednošću TER metrike (0.67).

Tablica 48 komparativno prikazuje performanse sustava Google Translate i Yandex Translate u usporedbi s izgrađenim sustavima za hrvatsko-engleski smjer (sustavi 1-4). Iz rezultata proizlazi da je Google Translate bodovan bolje, i to s obzirom na sve korištene automatske evaluacijske metrike. Sustav 2, tj. sustav treniran na specifičnom korpusu iz domene računalnog softvera ostvario je neznatno bolji rezultat za metriku TER. Ako se usporede rezultati svih metrika, sustav 2 se nalazi na drugom mjestu sa slabijim rezultatima za oko 3.5-18%.

Rezultate slične sustavu 2 postiže i Yandex Translate, koji ne zaostaje mnogo za sustavom 2. Hibridni sustavi 3 i 4 u ovom poretku zauzimaju četvrto, odnosno 5. mjesto. U ovoj usporedbi sustav 1, tj. sustav treniran na općenitoj domeni zauzeo je posljednje 6. mjesto. Isto tako, Google Translate i Yandex Translate ostvarili su bolje rezultate u odnosu na hibridne sustave, s obzirom da oba web servisa raspolažu daleko većim jezičnim i prijevodnim modelima.

Time je treća hipoteza (H3) dijelom odbačena, s obzirom da izgrađeni sustavi za statističko strojno prevođenje temeljeno na frazama za hrvatsko-engleski smjer u komparaciji s web servisima za statističko strojno prevođenje (Google Translate i Yandex Translate) nisu polučili bolje rezultate u odnosu na web servise u određenoj domeni. Tj. sustav 2 je demonstrirao bolje rezultate u odnosu na Yandex Translate, međutim, to nije uspio učiniti i za Google Translate. No, potrebno je provjeriti i statističku značajnost rezultata evaluacijskih metrika.

Tablica 48. Performanse sustava Google Translate i Yandex Translate u usporedbi s izgrađenim sustavima za hrvatsko-engleski smjer

metrika	**hrvatsko-engleski**					
	Google Translate	**Yandex Translate**	**sustav 1**	**sustav 2**	**sustav 3**	**sustav 4**
BLEU	**0.4405**	0.3142	0.0732	0.3734	0.1229	0.1064
NIST	**7.8466**	6.4368	2.9351	7.4487	3.5689	3.3956
METEOR	**0.3440**	0.2958	0.1169	0.3316	0.1367	0.1315
GTM	**0.7002**	0.6307	0.3401	0.6750	0.3923	0.3817

WER	**0.4549**	0.4893	0.7953	0.4966	0.7272	0.7244
TER	0.3737	0.5631	0.7726	**0.3681**	0.7103	0.7092

Tablica 49 komparira rezultate evaluacijskih metrika za sustave koji prevode s engleskog na hrvatski jezik. Interesantno je što je sustav 6, tj. sustav treniran za specifičnu domenu (računalni softver) ostvario najbolje rezultate. Ostvareni rezultati bolji su za oko 10%-16% u usporedbi sa sustavom Google Translate, odnosno 25% do 80% u usporedbi s treće plasiranim Yandex Translate sustavom.

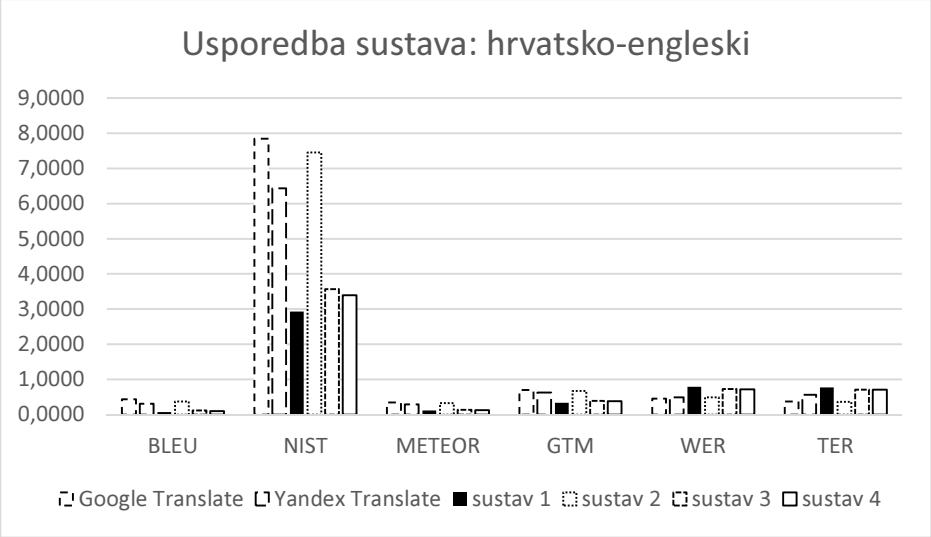
Hibridni sustavi 7 i 8 bili su lošiji u odnosu na web servise te u poretku za englesko-hrvatski smjer zauzimaju 4. i 5. mjesto. BLEU vrijednost sustava 5, tj. sustava treniranog na općoj domeni manja je u odnosu na najbolje rangirani sustav 6 oko 5.1 puta, oko 4.4 puta u odnosu na Google Translate, te oko 2.9 puta u odnosu na Yandex Translate.

S obzirom da je za englesko-hrvatski smjer najbolje rezultate postigao sustav 6, tj. sustav treniran na specifičnoj domeni, tj. domeni računalnog softvera, može se tvrditi da je treća hipoteza (H3) dijelom i prihvaćena. Naime, izgrađen sustav za statističko strojno prevođenje temeljeno na frazama za englesko-hrvatski smjer u komparaciji s postojećim web servisima za statističko strojno prevođenje (Google Translate i Yandex Translate) polučio je bolje rezultate za određenu domenu. I za englesko-hrvatski smjer potrebno je odrediti statističku značajnost rezultata evaluacijskih metrika.

Tablica 49. Performanse sustava Google Translate i Yandex Translate u usporedbi s izgrađenim sustavima za englesko-hrvatski smjer.

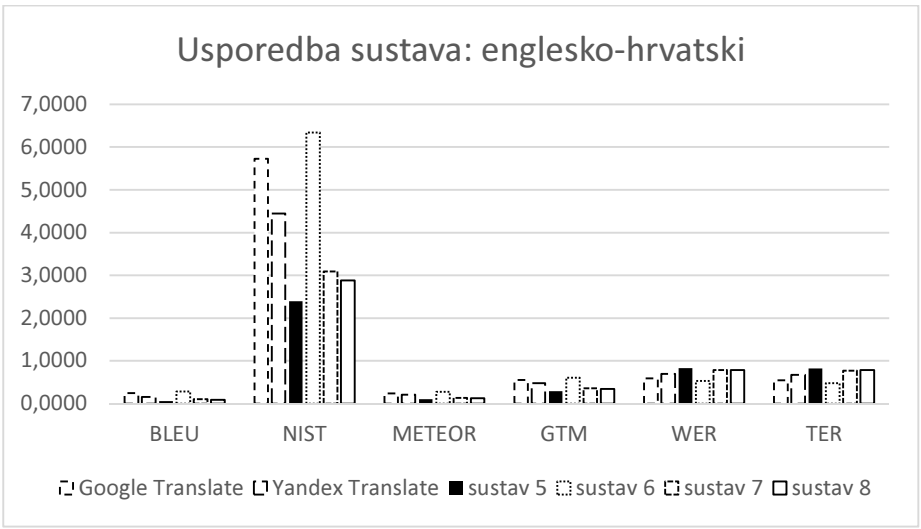
metrika	**englesko-hrvatski**					
	Google Translate	**Yandex Translate**	**sustav 5**	**sustav 6**	**sustav 7**	**sustav 8**
BLEU	0.2441	0.1589	0.0552	**0.2839**	0.1008	0.087
NIST	5.7246	4.4455	2.3918	**6.3419**	3.0877	2.8805
METEOR	0.2406	0.2044	0.0995	**0.2732**	0.1311	0.1227
GTM	0.5495	0.4782	0.2921	**0.6008**	0.3555	0.3405
WER	0.5883	0.6963	0.8308	**0.5310**	0.7806	0.7877
TER	0.5480	0.6723	0.8195	**0.4780**	0.7720	0.7822

U nastavku je dan grafički prikaz rezultata evaluacijskih metrika za sve korištene sustave za statističko strojno prevođenje, za hrvatsko-engleski smjer (Slika 36).



Slika 36. Prikaz rezultata evaluacijskih metrika za sve korištene sustave za statističko strojno prevođenje, za hrvatsko-engleski smjer.

Slika 37 prikazuje odnos korištenih sustava za englesko-hrvatski smjer i za sve korištene evaluacijske metrike.



Slika 37. Prikaz rezultata evaluacijskih metrika za sve korištene sustave za statističko strojno prevođenje, za englesko-hrvatski smjer.

I u slučaju online prevodilačkih alata na strojno prevedenim podatkovnim skupovima za testiranje (ispitivanje) primijenjena je metoda ponovnog uzorkovanja s ponavljanjem radi određivanja intervala pouzdanosti (n=10000). Intervali su određeni na temelju „umjetnih", tj. *bootstrapanih* podatkovnih skupova te pomoću njihovih aritmetičkih sredina vrijednosti BLEU metrike (na razini rečenica) i standardnih devijacija. U nastavku dani su rezultati metode ponovnog uzorkovanja (Tablica 50).

Tablica 50. Rezultati analize *bootstrapanih* podatkovnih skupova sustava Google Translate i Yandex Translate.

hrvatsko-engleski	**Google Translate**	**Yandex Translate**
aritmetička sredina	0.4791	0.3783
standardna devijacija	0.0096	0.0094
granica pogreške	0.0192	0.0187
interval pouzdanosti	[0.4599, 0.4983]	[0.3596, 0.3970]
englesko-hrvatski	**Google Translate**	**Yandex Translate**
aritmetička sredina	0.3477	0.2755
standardna devijacija	0.0090	0.0079
granica pogreške	0.0179	0.0158
interval pouzdanosti	[0.3298, 0.3656]	[0.2597, 0.2913]

Nakon analize intervala pouzdanosti (Tablica 37 i 50) može se zaključit da su razlike među izgrađenim sustavima statistički značajne za: sustav 1 i Google Translate, sustav 1 i Yandex Translate, sustav 2 i Yandex Translate, sustav 3 i Google Translate, sustav 3 i Yandex Translate, sustav 4 i Google Translate te sustav 4 i Yandex Translate za hrvatsko-engleski smjer prevođenja. Analizom intervala nije potvrđeno da je Google Translate statistički značajno bolji u odnosu na sustav 2, tj. na sustav treniran na domeni računalnog softvera.

Za englesko-hrvatski smjer, razlike su značajne između: sustava 5 i Google Translate, sustava 5 i Yandex Translate, sustava 6 i Google Translate, sustava 6 i Yandex Translate, sustava 7 i Google Translate, sustava 7 i Yandex Translate, sustava 8 i Google Translate te sustava 8 i Yandex Translate. Može se zaključiti da je sustav 6, treniran na domenski specifičnom korpusu, bio zaista i značajno bolji u odnesu na web prevodilačke servise.

Zatim se provjerila razdioba rezultata BLEU metrike na razini rečenica, tj. segmenata iz 10000 *bootstrapanih* uzoraka. Tj. testiralo se da li se radi o normalnim distribucijama vrijednosti rezultata

metrike BLEU (Dodatak A). Svi testirani uzorci web servisa Google Translate i Yandex Translate bili su normalno distribuirani te se stoga pristupilo z-testu. Od izgrađenih sustava upotrijebljeni su sustavi koji su također zadovoljili test normalne razdiobe, tj. sustave 1 i 2 (za hrvatsko-engleski smjer) te sustave 6, 7 i 8 za englesko-hrvatski smjer. Rezultati z-testa dani u nastavku (Tablica 51)

Tablica 51. Rezultati z-testa na izgrađenim sustavima i online prevodilačkim alatima.

hrvatsko-engleski	**Google Translate**	**Yandex Translate**
sustav 1	2243.2839	1443.147
sustav 2	129.4351	-620.488
englesko-hrvatski	**Google Translate**	**Yandex Translate**
sustav 6	-504.1973	-1097.39
sustav 7	940.3464	385.6966
sustav 8	951.2094	464.8288

Iz gornje tablice proizlazi da se nul-hipoteza u svim testiranim slučajevima može odbaciti, budući da vrijednosti upadaju u kritično područje <-∞,-1.96> i <1.96.∞>. Drugim riječima, sve testirane kombinacije sustava su statistički značajno različite, s obzirom da su im i očekivane vrijednosti različite (Tablica 51).

Z-test potvrđuje rezultate analize intervala pouzdanosti, međutim, ovdje je ipak uočena statistički značajna razlika između sustava 2 i Google Translate te se stoga može tvrditi da je Google Translate za hrvatsko-engleski smjer generirao bolje prijevode. S obzirom da vrijednosti uzorkovanih BLEU metrika sustava 3, 4 i 5 nisu normalno distribuirane, z-testom se nije mogla utvrditi značajnost razlike, u usporedbi s online prevodilačkim servisima. Z-testom je dijelom potvrđena treća hipoteza u ovom radu.

6. ZAKLJUČAK

Radom su analizirane postojeće teorije i metodologije izgradnje sustava za statističko strojno prevođenje temeljeno na frazama te istraženi novi pristupi povećanju kvalitete automatskog strojnog prijevoda u općoj domeni i domeni računalnog softvera za englesko-hrvatski i hrvatsko-engleski jezični par. Za vrijeme provođenja doktorskog istraživanja primijenjene su različite znanstvene metode, kao što su kvalitativna analiza znanstvenog dosega i metoda na području statističkog strojnog prevođenja, eksperiment, opservacija, mjerenje i kvantitativna analiza istraživačkih rezultata, komparativna analiza rezultata, induktivno i deduktivno zaključivanje te dokazivanje i opovrgavanje hipoteza.

U središtu istraživanja bili su statističko strojno prevođenje temeljeno na frazama te računalna adaptacija domene. Adaptacija domene u strojnom prevođenju u širem smislu predstavlja prilagodbu modela sustava za statističko strojno prevođenje određenoj domeni prevođenja, tj. namjeni sustava. Ona se može ostvariti na velik broj načina, npr. pomoću manipulacije ulaznoga podatkovnog skupa, tj. tekstualnih korpusa s ciljem postizanja veće kvalitete strojnog prijevoda u određenoj domeni ili primjenom složenijih metoda.

U radu su korišteni različiti ulazni podatkovni skupovi, ovisno o smjeru prevođenja i namjeni u procesu izgradnje sustava za strojno prevođenje. Za treniranje jezičnih modela ciljnog jezika upotrijebljeni su jednojezični tekstualni korpusi različitih veličina. Nadalje, za treniranje prijevodnih modela korišteni su paralelni korpusi, također različitih veličina. Primjerice, paralelni korpus koji je upotrijebljen u procesu treniranja sustava za prevođenje teksta iz domene računalnog softvera bio je 16 puta manji u odnosu na podatkovni skup za treniranje sustava općenite domene. Podatkovni skup koji je pak korišten za ugađanje svih izgrađenih sustava za statističko strojno prevođenje temeljeno na frazama sastojao se od 1000 sravnjenih rečenica, tj. segmenata koji proizlaze iz specifičnog korpusa u domeni računalnog softvera.

Podatkovni skup za ispitivanje, tj. testiranje svih izgrađenih sustava za statističko strojno prevođenje i web servisa sastojao se također od 1000 sravnjenih rečenica, tj. segmenata (dotičnog jezičnog smjera) iz specijaliziranog korpusa u domeni računalnog softvera. Korišteni podatkovni skupovi za treniranje, ugađanje te ispitivanje, tj. testiranje bili su disjunktni. Prije nego li se

započelo s treniranjem statističkih modela, bilo je potrebno pretprocesirati podatkovne skupove, što je uključivalo tokenizaciju teksta, pretvaranje veličine prvog slova tokena u najfrekventniji oblik te čišćenje korpusa. Naime, da bi adaptacija domene u modelu sustava za strojno prevođenje bila uspješna, tj. da bi se postigla veća kvaliteta strojnog prijevoda tekstualni korpusi trebaju biti što konzistentniji, domenom homogeniji te visoke kvalitete.

Skup alata koji su omogućili razvoj modela sustava za statističko strojno prevođenje temeljeno na frazama bili su GIZA++, koji omogućuje učenje IBM-ovih modela i sravnjivanje riječi, Moses, koji je prvenstveno korišten u postupku treniranja prijevodnih modela i dekodiranja prijevoda te IRSTLM, koji je upotrijebljen za treniranje i izglađivanje jezičnog modela, što je posebno važno, budući da se, zbog Zipfovog zakona, dodavanjem sve veće i veće količine podataka, ne mogu nužno pokriti sve kombinacije n-grama i specifične rečenične konstrukcije. Međutim, treba napomenuti da je sustav za strojno prevođenje jak samo onoliko koliko je jak njegov najslabiji element, tj. efikasnost sustava ovisi o najslabijoj komponenti u modelu sustava. Drugim riječima, ukoliko se u postupku treniranja statističkih modela rabe nekvalitetni ili domenski raznovrsni korpusi malenog opsega, to će izravno utjecati na karakteristike statističkih modela te time i na kvalitetu strojnog prijevoda.

Ostali alati koji su bili potrebni za pretprocesiranje i analizu podatkovnih skupova, ekstrakciju korpusa, osiguravanje disjunktnosti podatkovnih skupova za statističko strojno prevođenje izrađeni su pomoću programskih jezika Python, sed i Perl.

Automatske metrike ocjenjuju ekvivalentnost izvornog i ciljnog jezika, tj. vrše izračun kvalitete strojnog prijevoda, no pored toga imaju još jednu vrlo važnu ulogu u izgradnji sustava za statističko strojno prevođenje – služe za optimiziranje (ugađanje) sustava u odnosu na određenu metriku. Radi povećanja kvalitete strojnog prijevoda s obzirom na određenu metriku (najčešće BLEU), nad log-linearnim modelom sustava za statističko strojno prevođenje provodi se metoda diskriminativnog učenja čime se žele otkriti optimalne vrijednosti težina komponenti modela, tj. značajki. Tim pristupom nastoji se osigurati točnost prijevoda te usklađenost s modelom sustava za statističko strojno prevođenje temeljeno na frazama. Učinak podešavanja težina mjeri se odabranom metrikom - takav postupak pripada metodi učenja s minimalnom stopom pogreške (MERT), a provodi se kroz više iteracija.

U ovom istraživanju eksperimentirano je na ukupno 8 sustava za strojno prevođenje: 4 za hrvatsko-engleski te 4 za englesko-hrvatski smjer. Od toga su ukupno četiri sustava tzv. hibridni sustavi. Hibridni sustav za statističko strojno prevođenje, u ovom doktorskom radu, podrazumijeva primjenu alternativnog puta dekodiranja koja dozvoljava izgradnju skupa mogućih

prijevoda pomoću više prijevodnih modela. Takav hibridni sustav za strojno prevođenje ne sadrži isključivo podmodele svojstvene samo jednoj domeni (tj. ne koristi samo modele trenirane na jednoj domeni) već kombinira statističke modele iz različitih domena. Drugim riječima, prvo su trenirana i ugođena dva sustava, od čega bi se jedan od izvandomenskih prijevodnih modela u hibridnom sustavu zatim koristio kao alternativni put dekodiranja. Skup mogućih prijevoda pretražuje se prvo u prijevodnom modelu treniranom na specifičnom podatkovnom skupu, a tek zatim u prijevodnom modelu treniranom na nekoj drugoj (općenitoj) domeni ukoliko mogući prijevodi nisu pronađeni u prvom, tj. preferiranom modelu. U tom slučaju, prijevodni model treniran na općenitoj domeni predstavlja *back-off* prijevodni model za riječi i fraze koje nisu viđene u prvoj, tj. preferiranoj tablici prijevoda fraza, tj. fraznih struktura. Ovom metodom računalne adaptacije domene nastojale su se poboljšati performanse sustava za statističko strojno prevođenje za hrvatsko-engleski i englesko-hrvatski jezični par.

Sustavi 1 (hrvatsko-engleski) i 5 (englesko-hrvatski) trenirani su na korpusu opće domene, sustavi 2 (hrvatsko-engleski) i 6 (englesko-hrvatski) trenirani su na specijaliziranom korpusu, tj. domeni računalnog softvera. Sustavi 3 (hrvatsko-engleski) i 7 (englesko-hrvatski) su hibridni sustavi koji koriste tablice prijevoda fraza, tj. fraznih struktura iz obje domene, a preostale značajke modela (leksičko premještanje, jezični model itd.) iz domene računalnog softvera. Sustavi 4 (hrvatsko-engleski) i 8 (englesko-hrvatski) su hibridni sustavi koji koristi tablice prijevoda fraza, tj. fraznih struktura iz obje domene, a preostale značajke modela (leksičko premještanje, jezični model itd.) iz općenite domene.

Nakon izgradnje sustava za strojno prevođenje pristupilo se evaluaciji kvalitete strojnih prijevoda pomoću automatskih metrika. Rezultati metrika na razini rečenica, tj. segmenata iz 10000 *bootstrapanih* uzoraka korišteni su u metodi ponovnog uzorkovanja s ponavljanjem, radi otkrivanja intervala pouzdanosti koji su zatim međusobno uspoređivani radi ispitivanja statističke značajnosti razlika među sustavima za strojno prevođenje. Nakon primjene automatskih metrika za evaluaciju kvalitete prijevoda, izvršena je analiza pogrešaka u strojnim prijevodima prema pet kriterija: točnost, jezik, terminologija, stil i specifični standardi. Zatim je izvršena i ljudska evaluacija na svih 8 sustava (100 rečenica po sustavu) prema kriterijima točnosti/adekvatnosti i tečnosti/fluentnosti. Ljudsku evaluaciju izvršila su tri evaluatora na skali od 1-4. Nakon toga ispitana je razina (ne)slaganja evaluatora pomoću Cronbach alphe kako bi se ustanovila razina konzistentnosti u ocjenjivanju strojnih prijevoda. Potom su ispitane i razine korelacija između rezultata ljudske evaluacije i rezultata automatskih metrika pomoću Pearsonovog koeficijenta korelacije. Sustavi su, s obzirom na ishode evaluacije, odgovarajuće rangirani.

Izračunom Pearsonovog koeficijenta korelacije potvrđena je općenito vrlo dobra korelacija rezultata ljudske evaluacije i rezultata automatskih metrika. Za mjerenje stupnja međusobne (ne)složnosti ljudskih evaluatora izračunate se vrijednosti Cronbach alphe, koje za hrvatsko-engleske sustave upućuju na dobru i visoku konzistentnost među evaluatorima. Razina slaganja evaluatora slabija je za englesko-hrvatske sustave.

Radom su ispitane tri glavne hipoteze. Prva je ispitivala utjecaj adaptacija, tj. prilagodbe domene i karakteristika ulaznoga podatkovnog skupa na kvalitetu strojnog prijevoda, tj. na performanse sustava za statističko strojno prevođenje temeljeno na frazama za hrvatsko-engleski i englesko-hrvatski jezični par. Ova hipoteza (H1) se u potpunosti može potvrditi, s obzirom da su za oba jezična smjera, tj. hrvatsko-engleski i englesko-hrvatski, sustavi trenirani na korpusu specifične domene (sustavi 2 i 6) generirali najbolje, tj. najprihvatljivije strojne prijevode. To je potvrđeno i automatskim evaluacijskim metrikama BLEU, NIST, METEOR, GTM, WER i TER. Znatno bolje performanse sustava 2 i 6 dokazane su i na statističko značajnoj razini. Visoka kvaliteta domenski specifičnih podatkovnih skupova, tj. paralelnih i jednojezičnih korpusa korištenih za treniranje statističkih modela, doprinijela je evaluacijskim rezultatima strojnih prijevoda sustava 2 i 6, tj. sustava treniranih za domenu računalnog softvera. To ukazuje i na činjenicu da se i s vrlo ograničenom količinom podatkovnih skupova za treniranje u konačnici mogu generirati dobri prijevodi, što je potvrđeno i analizom te ukupnim brojem pogrešaka. U ovom slučaju sustavi 1 i 5, tj. sustavi trenirani na 16 puta većem korpusu nisu uspjeli demonstrirati veću kvalitetu u odnosu na domenski specifične sustave 2 i 6. Ljudska evaluacija potvrđuje dominaciju sustava 2 i 6 u usporedbi sa sustavima treniranim na općenitoj domeni (1 i 5) te hibridnim sustavim (3, 4, 7 i 8). Bolje performanse domenski specifičnih sustava potvrđuje i metoda ponovnog uzorkovanja s ponavljanjem, primijenjena na BLEU metrici te z-test.

Drugom hipotezom istražile su se mogućnosti primjene hibridnih sustava za statističko strojno prevođenje temeljeno na frazama za hrvatsko-engleski i englesko-hrvatski jezični par te mogućnosti poboljšanja kvalitete strojnog prijevoda u odnosu na sustave koji koriste značajke svojstvene isključivo jednoj domeni. Interesantno je što se u slučaju oba jezična smjera pomoću evaluacije automatskim metrikama ispostavilo da se primjenom hibridnog sustava koji kombinira domenske i izvandomenske značajke mogu postići bolji rezultati u odnosu na sustave trenirane na samo jednoj domeni (tj. sustave 1 i 5). Razlog leži u tome što izvandomenski statistički modeli osiguravaju relativno visok odziv, dok istovremeno domenski specifičan podatkovni skup pruža preciznost. Snaga hibridnih sustava dokazana je u slučaju sustava 1, 3 i 4 te 5, 7 i 8. No, u slučaju komparacije sustava 2, 3, i 4 te 6, 7 i 8 hipoteza nije potvrđena. No, neovisno o tome, hipoteza H2 se može potvrditi budući da je dokazano da je metodom alternativnog puta dekodiranja, kao

oblikom računalne adaptacije domene, moguće postići poboljšanja kvalitete strojnog prijevoda u odnosu na sustave koji koriste isključivo domenske značajke. Ljudska evaluacija doduše ne potvrđuje rezultate automatskih metrika na hibridnim sustavima za hrvatsko-engleski, tj. kada se kompariraju sustavi 1, 3 i 4, međutim, ta razlika je neznatna (oko 2%). Za englesko-hrvatski smjer, ljudska evaluacija pak potvrđuje bolje performanse hibridnih sustava u odnosu na sustav 5. Kvalitativna i kvantitativna analiza pogrešaka u strojnim prijevodima također potvrđuje da su hibridni sustavi 3, 4, 7 i 8 ukupno generirali manji broj pogrešaka u odnosu na sustave 1 i 5. Metoda ponovnog uzorkovanja potvrdila je značajnu razliku između sustava treniranih na općenitoj domeni i hibridnih sustava: 1 i 3, 1 i 4, 5 i 7 te 5 i 8. Z-testom je, međutim, zbog ograničenja na normalne distribucije bilo moguće potvrditi samo razliku među hibridnim sustavima 7 i 8 te između sustava 1 i 2, 6 i 7 te 6 i 8.

Treća hipoteza provjerila je tvrdnju da vlastiti, tj. izgrađeni sustavi za statističko strojno prevođenje temeljeno na frazama u komparaciji s postojećim web servisima za statističko strojno prevođenje mogu postići bolje rezultate za određenu domenu. Kada se analiziraju sustavi za prevođenje hrvatsko-engleskog smjera, može se konstatirati da je Google Translate ostvario bolje rezultate evaluacijskih metrika u odnosu na hibridne sustave te sustave 1 i 2. No, sustav 2 ipak je nadmašio sustav Yandex Translate. Međutim, sustav 6 nadjačao je oba web servisa u slučaju englesko-hrvatskog smjera. Time je H3 ipak samo djelomična potvrđena. Ljudska evaluacija, Cronbach alpha, izračun korelacije te analiza pogrešaka u strojnim prijevodima nisu izvršeni za prijevode generirane pomoću Google Translate i Yandex Translate online prevodilačkih servisa. No, analizom *bootstrapanih* intervala pouzdanosti potvrđene su statistički značajne razlike među sustavima, pri čemu je Google Translate demonstrirao bolje performanse u odnosu na sustav 2, no, lošije u odnosu na sustav 6. Hibridni sustavi bili su značajno lošiji u odnosu na online prevodilačke alate. Z-test potvrđuje nadmoć Google Translate alata u odnosu na ostale sustava za strojno prevođenje i najbolje rangirani sustav za hrvatsko-engleski smjer prevođenja (sustav 2). Z-test isto tako potvrđuje da je izgrađeni sustav 6 bio statistički značajno bolji u odnosu na analizirane web servise.

Može se zaključiti da je ovim doktorskim radom istražena problematika međuovisnosti domene i performansi sustava za statističko strojno prevođenje temeljeno na frazama na praktičnom primjeru hrvatsko-engleskog jezičnog para za oba smjera, te na općenitoj domeni i domeni računalnog softvera. Opaženo je da statistički sustavi trenirani na izvandomenskom podatkovnom skupu nisu uspjeli generirati strojne prijevode prihvatljive kvalitete za domenu računalnog softvera. Glavni razlog tome jesu velike razlike između općenite domene i specifične domene, u smislu vokabulara, tj. specifične terminologije, stila, gramatike, diskursa itd. Razvijeni

su i sustavi pomoću korpusa koji pripada vrlo specifičnom području. Zatim su primijenjene određene tehnike adaptacije domene kako bi se poboljšala kvaliteta strojnih prijevoda u domeni računalnog softvera. Hibridnim sustavima nastojalo se utjecati na povećanje performansi sustava za statističko strojno prevođenje.

Nadalje, ovim radom dokazane su dvije hipoteze, a treća je potvrđena samo djelomično, s obzirom da istraživanja na jednom jezičnom smjeru nisu potvrdila očekivanja u potpunosti. Karakteristike ulaznoga podatkovnog skupa i hibridni pristup izgradnji sustava za strojno prevođenje uvelike utječu na kvalitetu englesko-hrvatskih i hrvatsko-engleski strojnih prijevoda. Ukupno manji broj pogrešaka generirali su sustavi trenirani na specifičnoj domeni, a slijede ga hibridni sustavi. Na začelju su sustavi trenirani na općenitoj domeni, što dokazuje da se sustavi za strojno prevođenje trebaju razvijati za posebnu namjenu. Nema sumnje da su u ovom istraživanju sustavi trenirani na domenski specifičnim korpusima generirali znatno bolje rezultate i kvalitetnije strojne prijevode. U ovom radom istraženo je i koliko je domenski specifičnog korpusa potrebno za razvoj sustava koji prevode specifičnu domenu (računalni softver). Rezultati istraživanja upućuju na to da se i relativno malenim, ali visoko kvalitetnim podatkovnim skupovima mogu izgraditi vrlo kvalitetni sustavi. Naime, iako su u ovom istraživanju domensko specifični sustavi trenirani na višestruko manjim podatkovnim skupovima, ispostavilo se da su u stanju nadmašiti ne samo vlastite sustave trenirane na općenitom korpusu, već i online prevodilačke alate koji raspolažu znatno većim resursima. Ta razlika potvrđena je i metodom ponovnog uzorkovanja s ponavljanjem te z-testom.

Budući da se prilagodba ulaznog podatkovnog skupa prevođenju specifične domene također može smatrati jednim oblikom adaptacije domene, rezultati istraživanja jasno ukazuju na to da se pomnim ugađanjem parametara sustava pomoću specifičnih korpusa mogu ostvariti značajne razlike u rezultatima, a to potvrđuje i ljudska evaluacija te evaluacija automatskim metrikama. Metoda alternativnog puta dekodiranja pokazala se također vrlo efikasnom metodom adaptacije domene, budući da su za oba jezična smjera hibridni sustavi demonstrirali poboljšanje performansi sustava za strojno prevođenje. Time je dokazano da se postojeći, tj. već trenirani sustavi za statističko strojno prevođenje mogu prilagoditi sasvim različitim domenama. Određene razlike između hibridnih i nehibridnih sustava su i statistički značajne (npr. sustavi 1 i 3 te 5 i 7). Drugim riječima, hibridni sustavi bili su efikasniji u odnosu na sustave trenirane na općenitoj domeni, unatoč činjenici da komponente hibridnih sustava nisu posebno ugađane. Već naprotiv, upotrijebljene su težine značajki iz domenski specifičnih ili općenitih sustava. Ipak, hibridni sustavi niti jednom nisu uspjeli nadmašiti rezultate sustava treniranih na specifičnom korpusu.

7. LITERATURA

1. ACCEPT. ACCEPT Automated Community Content Editing PorTal: Analysis of existing metrics and proposal for a task-oriented metric. Seventh Framework Programme, project deliverable, p. 9, 2012.
2. Adeyanju, I.; Wiratunga, N.; Lothian, R.; Craw, S. Applying Machine Translation Evaluation Techniques to Textual CBR. Proceedings of 18th International Conference on Case-Based Reasoning, ICCBR 2010. In Case-Based Reasoning. Research and Development - Lecture Notes in Computer Science, vol. 6176, Springer, pp 21-35, 2010.
3. Agarwal, A.; Lavie, A. METEOR, M-BLEU and M-TER: Evaluation Metrics for High-Correlation with Human Rankings of Machine Translation Output. ACL 2008: Third Workshop on Statistical Machine Translation (StatMT '08), pp. 115-118, 2008.
4. Ahn, S.; Fessler, J. A. Standard errors of mean, variance, and standard deviation estimators. Technical Report 413, Comm. and Sign. Proc. Lab., Dept. of EECS, Univ. of Michigan, p. 2, 2003.
5. Al-Onaizan, Y.; Curin, J.; Jahr, M.; Knight, K.; Lafferty, J.; Melamed, D.; Och, F.-J.; Purdy, D.; Smith, N. A.; Yarowsky, D. Statistical Machine Translation. Final Report of the JHU Summer Workshop, p. 42, 1999.
6. ALPAC. Language and Machines: Computers in Translation and Linguistics - A Report by the Automatic Language Processing Advisory Committee Division of Behavioral Sciences National Academy of Sciences. National Research Council National Academy of Sciences, National Research Council, p. 138, 1966.
7. Arun, A.; Koehn, P. Online Learning Methods for Discriminative Training of Phrase Based. Statistical Machine Translation. Proceedings of the MT Summit XI, pp. 6, 2007.
8. Auli, M.; Galley, M.; Gao, J. Large-scale Expected BLEU Training of Phrase-based Reordering Models. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, pp. 1250-1260, 2014.

9. Axelrod, A.; He, X.; Gao, J. Domain Adaptation via Pseudo In-Domain Data Selection. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics (ACL), pp. 355-362, 2011.
10. Aziz, W.; Sousa, S. C. M. de; Specia, L. PET: a Tool for Post-editing and Assessing Machine Translation. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), pp. 3982-3987, 2012.
11. Aziz, W.; Special, L. PET: A Standalone Tool for Assessing Machine Translation through Post-editing. [Aslib 2012] Translating and the Computer Conference 34, p. 5, 2012.
12. Baldwin, T. Translation Memory Engines: A Look under the Hood and Road Test Proceedings of the 15th International Japanese/English Translation Conference, p. 19, 2004.
13. Banerjee, P.; Kumar Naskar, S.; Roturier, J.; Way, A.; van Genabith, J. Domain Adaptation in SMT of User-Generated Forum Content Guided by OOV Word Reduction: Normalization and/or Supplementary Data? Proceedings of the 16th EAMT Conference, pp. 169-176, 2012.
14. Banerjee, P.; Kumar Naskar, S.; Roturier, J.; Way, A.; van Genabith, J. Domain Adaptation in Statistical Machine Translation of User-Forum Data using Component-Level Mixture Modelling. Proceedings of Machine Translation Summit XIII, pp. 285-292, 2011.
15. Banerjee, S.; Lavie, A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. ACL 2005: Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, pp. 65-72, 2005.
16. Berger, A. L.; Brown, P. F.; Della Pietra, S. A.; Della Pietra, V. J. ; Gillett, J. R.4 ; Lafferty. J. D.; Mercer, R. L.; Printz, H.; Ureš, L. The Candide System for Machine Translation. Proceedings of the HLT '94 Workshop on Human Language Technology, Association for Computational Linguistics, pp. 157-162, 1994.
17. Bertoldi, N. A tutorial on the IRSTLM library. Machine Translation Marathon 2008: Open Source Convention, tool demonstration documentation, p. 34, 2008.
18. Bertoldi, N.; Federico, M. Domain Adaption for Statistical Machine Translation with Monolingual Resources. ACL 2009: Fourth Workshop on Statistical Machine Translation (StatMT '09), pp. 182-189, 2009.

19. Birch, A.; Osborne, M.; Koehn, P. CCG Supertags in Factored Statistical Machine Translation. Proceedings of the Second Workshop on Statistical Machine Translation Pages (StatMT '07), pp. 9-16, 2007.
20. Black, A. W.; Brown, R. D.; Frederking, R.; Singh, R.; Moody, J.; Steinbrecher, E. TONGUES: rapid development of a speech-to-speech translation system. Proceedings of the second international conference on Human Language Technology Research (HLT '02), pp. 183-186, 2002.
21. Bird, S.; Klein, E.; Loper, E. Natural Language Processing with Python - Analyzing Text with the Natural Language Toolkit. O'Reilly Media, p. 504, 2009.
22. Bisazza, A.: Ruiz, N.; Federico, M. Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation. Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2011), pp. 136-143, 2011.
23. Bluman, A. G. Elementary Statistics: A Step by Step Approach. McGraw-Hill, 7th Edition, p. 897, 2009.
24. Bojar, O.; Buck, C.; Callison-Burch, C.; Federmann, C.; Haddow, B.; Koehn, P.; Monz, C.; Post, M.; Soricut, R.; Specia, L. Findings of the 2013 Workshop on Statistical Machine Translation. Proceedings of the Eighth Workshop on Statistical Machine Translation, Association for Computational Linguistics, pp. 1-44, 2013.
25. Bojar, O.; Buck, C.; Federmann, C.; Haddow, B.; Koehn, P.; Leveling, J.; Monz, C.; Pecina, P.; Post, M., Saint-Amand, H.; Soricut, R.; Specia, L.; Tamchyna, A. Findings of the 2014 Workshop on Statistical Machine Translation. Proceedings of the Ninth Workshop on Statistical Machine Translation, pp. 12-58, Association for Computational Linguistics, 2014.
26. Boyd-Graber, J. Machine Translation: Phrase-Based Models. Department of Computer Science at University of Colorado Boulder, course material, p. 27, 2014.
27. Brkić, M.; Seljan, S.; Matetić, M. Machine Translation Evaluation for Croatian-English and English-Croatian Language Pairs. Proceedings of the 8th International NLPCS Workshop: Human-Machine Interaction in Translation, pp. 93-104, 2011.
28. Brown, P. F.; Della Pietra, V. J.; Della Pietra S. A.; Mercer, R. L. The mathematics of statistical machine translation: parameter estimation. Computational Linguistics - Special issue on using large corpora: II, vol. 19, no. 2, pp. 263-311, 1993.
29. Buck, C.; Heafield, K.; van Ooyen, B. N-gram Counts and Language Models from the Common Crawl. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), pp. 3579-3584, 2014.

30. Bulyko, I.; Matsoukas, S.; Schwartz, R.; Nguyen, L.; Makhoul, J. Language Model Adaptation in Machine Translation from Speech. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007), vol. 4, pp. 117-120, 2007.
31. Burchardt, A.; Lommel, A. Practical Guidelines for the Use of MQM in Scientific Research on Translation Quality. Preparation and Launch of a Large-scale Action for Quality Translation Technology, report, p. 19, 2014.
32. Callison-Burch, C.; Koehn, P.; Monz, C.; Peterson, K.; Przybocki, M.; Zaidan, O. Findings of 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. Fifth Workshop on Statistical Machine Translation and Metrics (MATR), Association for Computational Linguistics, pp. 17-53, 2010.
33. Callison-Burch, C.; Koehn, P.; Monz, C.; Post, M.; Soricut, R.; Specia, L. Findings of the 2012 Workshop on Statistical Machine Translation. Proceedings of the 7th Workshop on Statistical Machine Translation, Association for Computational Linguistics, pp. 10-51, 2012.
34. Callison-Burch, C.; Osborne, M. Bootstrapping Parallel Corpora. HLT-NAACL: 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond (HLT-NAACL-PARALLEL '03), vol. 3, pp. 44-49, 2003.
35. Callison-Burch, C.; Osborne, M.; Koehn, P. Re-evaluating the Role of BLEU in Machine Translation Research. EACL: 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006), pp. 249-256, 2006.
36. Calude, A. S. Machine Translation of Various Text Genres. Te Reo - the journal of the Linguistic Society of New Zealand, vol. 46, p. 67-94, 2004.
37. Carl, M. Translog-II: a Program for Recording User Activity Data for Empirical Reading and Writing Research. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), pp. 4108-4112, 2012.
38. Carpuat, M.; Daume III, H.; Fraser, A.; Quirk, C.; Braune, F.; Clifton, A.; Irvine, A.; Jagarlamudi, J.; Morgan, J.; Razmara, M.; Tamchyna, A.; Henry, K.; Rudinger, R. Domain Adaptation in Machine Translation: Final Report. JHU-Summer Workshop 2012, p. 60, 2012.
39. Ceauşu, A.; Tinsley, J.; Zhang, J.; Way, A. Experiments on Domain Adaptation for Patent Machine Translation in the PLuTO project. Proceedings of the 15th Conference of the European Association for Machine Translation, pp. 21-28, 2011.

40. Cer, D.; Manning, C. D.; Jurafsky, D. The Best Lexical Metric for. Phrase-Based Statistical MT System Optimization. Proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10), pp. 555-563, 2010.
41. Chatzitheodorou, K.; Poulis, A. Experiments on domain-specific Statistical Machine Translation at the European Parliament. [Aslib 2012] Translating and the Computer Conference 34, p. 6, 2012.
42. Chen, B.; Cherry, C. A Systematic Comparison of Smoothing Techniques for Sentence-Level BLEU. Proceedings of the Ninth Workshop on Statistical Machine Translation, Association for Computational Linguistics pages, pp. 362-367, 2014.
43. Chen, S. F.; Goodman, J. An Empirical Study of Smoothing Techniques for Language Modeling. Technical Report TR-10-98, Computer Science Group at Harvard University, p. 64, 1998.
44. Chen, S. F.; Goodman, J. An Empirical Study of Smoothing Techniques for Language Modeling. Proceedings of the 34th annual meeting on Association for Computational Linguistics (ACL '96), p. 10, 1996.
45. Chen, S. F.; Goodman, J. An Empirical Study of Smoothing Techniques for Language Modeling. Computer Speech and Language, vol. 13, pp. 359-394, 1999.
46. Chéragui, M. A. Theoretical Overview of Machine translation. Proceedings of the 4th International Conference on Web and Information Technologies (ICWIT 2012), pp. 160-169, 2012.
47. Chhatbar, C. D. Improving Statistical Topic Models by Using Ontological Concepts. Department of Computer Science at Australian National University, p. 25, 2010.
48. Chiang, D. A hierarchical phrase-based model for statistical machine translation. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL '05), pp. 263-270, 2005.
49. Chiang, D.; Knight, K.; Wang, W. 11,001 New Features for Statistical Machine Translation. Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL, Association for Computational Linguistics, pp. 218-226, 2009.
50. Civera, J.; Juan, A. Domain Adaptation in Statistical Machine Translation with Mixture Modelling. Proceedings of the Second Workshop on Statistical Machine Translation, Association for Computational Linguistics, pp. 177-180, 2007.

51. Clark, S. Word Alignment Models. Statistical Machine Translation course at University of Cambridge (Clark, S.; de Gispert, A.), lecture 3, p. 27, 2010.
52. Costa-jussa`, M. R.; Banchs, R. E.; Rapp, R.; Lambert, P.; Eberle, K.; Babych, B. Workshop on Hybrid Approaches to Translation: Overview and Developments. Proceedings of the 2nd HyTra Workshop, Association for Computational Linguistics, pp. 1-6, 2013.
53. Crammer, K.; Singer, Y. Ultraconservative Online Algorithms for Multiclass Problems. Journal of Machine Learning Research, vol. 3, pp. 951-991, 2003.
54. De Almeida, G.; O'Brien, S. Analysing Post-Editing Performance: Correlations with Years of Translation Experience. Proceedings of the 14th annual conference of the European association for machine translation (EAMT 2010), pp. 8, 2010.
55. Denkowski, M.; Lavie, A. Extending the METEOR Machine Translation Evaluation Metric to the Phrase Level. NAACL: 2010 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies (HLT-NAACL 2010), pp. 250-253, 2010.
56. Denkowski, M.; Lavie, A. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of. Machine Translation Systems. Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation, pp. 85-91, 2011.
57. Denkowski, M.; Lavie, A. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. Proceedings of the Ninth Workshop on Statistical Machine Translation, Association for Computational Linguistics, pp. 376-380, 2014.
58. Dillinger, M. Introduction to MT. The Ninth Conference of the Association for Machine Translation in the Americas, tutorial documentation, p. 29, 2010.
59. Dillinger, M.; Marciano, J. Introduction to MT. The Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2012), tutorial documentation, p. 47, 2012.
60. Doddington, G. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. Second international conference on Human Language Technology Research (HLT '02), pp. 138-145, 2002.
61. Duma, M.-S.; Vertan, C. Integration of Machine Translation in On-line Multilingual Applications – Domain Adaptation. Translation: Computation, Corpora, Cognition. Special Issue on Language Technologies for a Multilingual Europe; Rehm, G.; Sasaki, F.; Stein, D.; Witt, A. (Eds.), vol. 3, no. 1, pp. 61-74, 2013.

62. Durrani, N.; Haddow, B.; Heafield, K.; Koehn, P. Edinburgh's Machine Translation Systems for European Language Pairs. Proceedings of the Eighth Workshop on Statistical Machine Translation, Association for Computational Linguistics, pp. 114-121, 2013.
63. Dyer, C. Discriminative Training of Translation Models. Seventh Machine Translation Marathon 2012 (MTM2012), lecture documentation, p. 83, 2012.
64. Dyer, C. Statistical Machine Translation. The Sixth Machine Translation Marathon (MT Marathon 2011), lecture material, p. 85, 2011.
65. Eck, M.; Vogel, S.; Waibel, A. Language Model Adaptation for Statistical Machine Translation Based On Information Retrieval. Proceedings of the 4th international conference on Language Resources and Evaluation (LREC 2004), pp. 327-330, 2004.
66. Efron, B.; Tibshirani, R. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. Statistical Science, vol. 1, no. 1, pp. 54-77, 1986.
67. Eisele, A.; Christian F.; Uszkoreit, H.; Saint-Amand, H.; Kay, M.; Jellinghaus, M.; Hunsicker, S.; Herrmann, T.; Chen, Y. Hybrid Architectures for Multi-Engine Machine Translation. Translating and the Computer 30, p. 12, 2008.
68. España-Bonet, C.; Gonzàlez, M. Statistical Machine Translation and Automatic Evaluation. The 9th edition of the Language Resources and Evaluation Conference (LREC 2014), tutorial documentation, p. 308, 2014.
69. Farzindar, A.; Khreich, W. Evaluation of Domain Adaptation Techniques for TRANSLI in a Real-World Environment. Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA 2012), p. 5, 2012.
70. Federico, M.; Bertoldi, N.; Cettolo, M. IRST Language Modeling Toolkit – user manual. Fondazione Bruno Kessler (FBK), p. 8, 2008.
71. Federico, M.; Bertoldi, N.; Cettolo, M. IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models. Ninth Annual Conference of the International Speech Communication Association (INTERSPEECH 2008), pp. 1618-1621, 2008b.
72. Federico, M.; Cattelan, A.; Trombetti, M. Measuring User Productivity in Machine Translation Enhanced. Computer Assisted Translation. Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA 2012), p. 10, 2012.
73. Federmann, C. How can we measure machine translation quality? Tralogy I - Translation Careers and Technologies: Convergence Points for the Future, p. 9, 2011.

74. Fishel, M.; Sennrich, R. Handling Technical OOVs in SMT. Proceedings of The Seventeenth Annual Conference of the European Association for Machine Translation (EAMT), pp. 159-162, 2014.
75. Folajimi, Y. O.; Omonayin, I. Using Statistical Machine Translation (SMT) as a Language Translation Tool for Understanding Yoruba Language. EIE's 86 2nd Intl' Conf. Comp., Energy, Net., Robotics and Telecom. (eieCon2012), pp. 86-91, 2012.
76. Foster, G.; Kuhn, R. Mixture-Model Adaptation for SMT. Proceedings of the Second Workshop on Statistical Machine Translation, Association for Computational Linguistics pp. 128-135, 2007.
77. Foster, G.; Kuhn, R. Stabilizing Minimum Error Rate Training. Proceedings of the 4th EACL Workshop on Statistical Machine Translation (StatMT'09/ACL 2009), Association for Computational Linguistics, pp. 242-249, 2009.
78. Gale, W. A.; Church, K. W. A program for aligning sentences in bilingual corpora. Proceedings of the 29th annual meeting ACL '91, Association for Computational Linguistics, pp. pages 177-184, 1991.
79. Garcia, I. Training Quality Evaluators. Revista Tradumàtica: tecnologies de la traducció: Traducció i qualitat, no. 12, pp. 430-436, 2014.
80. Gaspari, F.; Hutchins, J. Online and free! Ten years of online machine translation: Origins, developments, current use and future prospects. Proceedings of the Machine Translation Summit XI, pp. 199-206, 2007.
81. Giménez, J. Empirical Machine Translation and its Evaluation. Talk at Yahoo Research in Barcelona, p. 187, 2008.
82. Giménez, J.; Gonzàlez, M. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation – Technical Manual version 1.0. TALP Research Center, LSI Department, Universitat Politècnica de Catalunya - Barcelona, p. 43, 2011.
83. Giménez, J.; Màrquez, L. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. The Prague Bulletin of Mathematical Linguistics, no. 94, pp. 77-86, 2010.
84. Gong, L.; Max, A.; Yvon, F. Towards Contextual Adaptation for Any-text Translation. Proceedings of the IWSLT 2012 Conference, pp. 292-299, 2012.
85. Gonzàlez, M. Automatic MT Evaluation. The 9th edition of the Language Resources and Evaluation Conference (LREC 2014), tutorial documentation, p. 76, 2014.

86. Goutte, C.; Carpuat, M.; Foster, G. The Impact of Sentence Alignment Errors on Phrase-Based Machine Translation Performance. Conference of the Association for Machine Translation in the Americas (AMTA 2012), p. 8, 2012.
87. Görög, A. Translation and Quality Editorial. Revista Tradumàtica: tecnologies de la traducció: Traducció i qualitat, no. 12, pp. 388-391, 2014.
88. Graham, Y.; Baldwin, T. Testing for Significance of Increased Correlation with Human Judgment. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, pp. 172-176, 2014.
89. Green, S.; Cer. D.; Manning, C. D. An Empirical Comparison of Features and Tuning for Phrase-based Machine Translation. Proceedings of the Ninth Workshop on Statistical Machine Translation, Association for Computational Linguistics, pp. 466-476, 2014.
90. Habash, N.; Olive, J.; Christianson, C.; McCary, J. Machine Translation from Text. In Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation; Olive, J.; Christianson, C.; McCary, J. (Eds.), Springer, pp. 133-397, 2011.
91. Haddow, B. Word Alignment. Third Machine Translation Marathon (MT Marathon 2009), Institute of Formal and Applied Linguistics Faculty of Mathematics and Physics at Charles University, p. 3, 2009.
92. Hampshire, S.; Porta Salvia, C. Translation and the Internet: Evaluating the Quality of Free Online Machine Translators. Quaderns. Rev. trad. 17, pp. 197-209, 2010.
93. Hardt, D; Elming, J. Incremental Re-training for Post-editing SMT. Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010), p. 10, 2010.
94. Hasler, E.; Haddow, B.; Koehn, P. Combining Domain and Topic Adaptation for SMT. Proceedings of AMTA 2014, vol. 1, pp. 139-151, 2014.
95. Haque. R.; Kumar Naskar, S.; van Genabith, J.; Way, A. Experiments on Domain Adaptation for English—Hindi SMT. Proceedings of PACLIC 23: the 23rd Pacific Asia Conference on Language, Information and Computation, pp. 670-677, 2009.
96. Hutchins, J. ALPAC: the (in)famous report. MT News International, no. 14, pp. 9-12, 1996.

97. Hutchins, J. Current commercial machine translation systems and computer-based translation tools: system types and their uses. International Journal of Translation, vol.17, no.1-2, pp. 5-38, 2005b.
98. Hutchins, J. Has machine translation improved? some historical comparisons. Proceedings of the MT Summit IX, Association for Machine Translation in the Americas, p. 8, 2003.
99. Hutchins, J. Fifty Years of the Computer and Translation. Machine Translation Review, no. 6, pp. 22-24, 1997.
100. Hutchins, J. Machine Translation and Human Translation: In Competition or in Complementation? International Journal of Translation, vol.13, no.1-2, pp. 5-20, 2001c.
101. Hutchins, J. Machine translation over fifty years. Histoire, Epistémologie, Langage: Le traitement automatique des langues, vol. 23, no. 1, pp. 7-31, 2001b.
102. Hutchins, J. Multiple Uses of Machine Translation and Computerised Translation Tools. Proceedings of the International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages (ISMTCL 2009), pp. 13-20, 2009.
103. Hutchins, J. The Development and Use of Machine Translation Systems and Computer-based Translation Tools. International Journal Of Translation, vol. 15, no. 1, pp. 5-26, 2003.
104. Hutchins, J. The Georgetown-IBM experiment demonstrated in January 1954. Machine translation: from real users to research: 6th conference of the Association for Machine Translation in the Americas (AMTA 2004), pp. 102-114, 2004b.
105. Hutchins, J. Towards a definition of example-based machine translation. Proceedings of the Tenth Machine Translation Summit (MT Summit X), Asia-Pacific Association for Machine Translation, Thai Computational Linguistics Laboratory (NICT), p. 8, 2005.
106. Hutchins, J. Towards a new vision for MT. MT Summit VIII - Machine Translation in the Information Age, European Association for Machine Translation, introductory speech, p. 6, 2001.
107. Hutchins, J. Two precursors of machine translation: Artsrouni and Trojanskij. International Journal of Translation, vol. 16, no. 1, pp. 11-31, 2004.
108. Huzak, M. Vjerojatnost i matematička statistika. Sveučilište u Zagrebu, PMF-Matematički odjel, predavanja s Poslijediplomskog specijalističkog sveučilišnog studija aktuarske matematike, p. 106, 2006.

109. Irvine, A.; Callison-Burch, C. Combining Bilingual and Comparable Corpora for Low Resource Machine Translation. Proceedings of the Eighth Workshop on Statistical Machine Translation, Association for Computational Linguistics, pp. 262-270, 2013.
110. Jiang J.; Way, A.; Haque, R. Translating User-Generated Content in the Social Networking Space. Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA 2012), p. 9, 2012.
111. Junczys-Dowmunt, M.; Pouliquen, B. SMT of German Patents at WIPO: Decompounding and Verb Structure Pre-reordering. Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT2014), pp. 217-220, 2014.
112. Jurafsky, D. Language Modeling. CS 124: From Languages to Information course at Stanford University, course material, p. 88, 2015.
113. Jurafsky, D.; Martin, J. Speech and Language Processing: Pearson New International Edition. Pearson Education Limited, 2nd edition, p. 944, 2013.
114. Kanavos, P.; Kartsaklis, D. Integrating Machine Translation with Translation Memory: A Practical Approach. Proceedings of the Second Joint EM+/CNGL Workshop "Bringing MT to the User: Research on Integrating MT in the Translation Industry" (JEC '10), pp. 11-20, 2010.
115. Khalilov, M.; Choudhury, R. Building English-Chinese and Chinese-English MT engines for the computer software domain. Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT 2012), pp. 7-11, 2012.
116. Klaper, D.; Ebling, S.; Volk, M. Building a German/Simple German Parallel Corpus for Automatic Text Simplification. ACL 2013: The Second Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR 2013), pp. 11-19, 2013.
117. Knight, K. A Statistical MT Tutorial Workbook. JHU summer workshop, p. 36, 1999.
118. Knight, K.; Koehn, P. What's New in Statistical Machine Translation. 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL 2003), tutorial documentation, p. 89, 2003.
119. Koehn, P. Challenges in Statistical Machine Translation. Presentation at PARC, Google, ISI, MITRE, BBN, University of Montreal, p. 51, 2004.

120. Koehn, P. Computer Aided Translation. Seventh Machine Translation Marathon (MT Marathon 2012), lecture documentation, p. 118, 2012.
121. Koehn, P. Europarl: A Parallel Corpus for Statistical Machine Translation. AAMT: The Tenth Machine Translation Summit, pp. 79-86, 2005.
122. Koehn, P. Introduction to Statistical Machine Translation. Chinese Workshop for Machine Translation, tutorial documentation, p. 214, 2008.
123. Koehn, P. Moses - Statistical Machine Translation System: User Manual and Code Guide. Statmt, University of Edinburgh, p. 353, 2015.
124. Koehn, P. The Foundation for Statistical Machine Translation at MIT. DARPA/TIDES MT Evaluation Workshop, p. 20, 2004b.
125. Koehn, P. Statistical Machine Translation. Cambridge University Press. ISBN-13: 978-0521874151, 2010.
126. Koehn, P. Statistical Machine Translation: the basic, the novel, and the speculative. EACL: 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006), tutorial documentation, p. 81, 2006.
127. Koehn, P. Statistical Significance Tests for Machine Translation Evaluation. Proceedings of EMNLP 2004, Association for Computational Linguistics, pp. 388-395, 2004c.
128. Koehn, P. What is a Better Translation? Reflections on Six Years of Running Evaluation Campaigns. Proceedings for TRALOGY 2011 Conference, p. 1-9, 2011.
129. Koehn, P.; Axelrod, A.; Birch Mayne, A.; Callison-Burch, C.; Osborne, M.; Talbot, D. Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation. International Workshop on Spoken Language Translation 2005, p. 8, 2005.
130. Koehn, P.; Callison-Burch, C. Statistical Machine Translation. 20th European Summer School in Logic, Language and Information (ESSLLI 2008), course material, p. 149, 2008.
131. Koehn, P.; Germann, U. The Impact of Machine Translation Quality on Human Post-Editing. Proceedings of the EACL 2014 Workshop on Humans and Computer-assisted Translation, Association for Computational Linguistics, p. 38-46, 2014.
132. Koehn, P.; Haddow, B. Interpolated backoff for factored translation models. Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA), p. 10, 2012b.

133. Koehn, P.; Haddow, B. Towards Effective Use of Training Data in Statistical Machine Translation. Proceedings of the 7th Workshop on Statistical Machine Translation, Association for Computational Linguistics, pp. 317-321, 2012.
134. Koehn, P.; Hoang, H. Machine Translation with Open Source Software. Machine Translation Summit XIV, tutorial documentation, p. 147, 2013.
135. Koehn, P.; Hoang, H.; Birch, A.; Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; Zens, R.; Dyer, C.; Bojar, O.; Constantin, A.; Herbst, E. Moses: Open Source Toolkit for Statistical Machine Translation. Proceedings of the ACL 2007Annual Meeting of the Association for Computational Linguistics (ACL) - Demo and Poster Sessions, pages 177-180, 2007.
136. Koehn, P.; Knight, K. Empirical Methods for Compound Splitting. Proceedings of the Tenth conference on European Chapter of the Association for Computational Linguistics (EACL), vol. 1, pp. 187-193, 2003.
137. Koehn, P.; Och, F. J.; Marcu, D. Statistical Phrase-Based Translation. Proceedings of the 2003 Human Language technology Conference - North American Chapter of the Association for Computational Linguistics (HLT/NAACL 2003), p. 7, 2003.
138. Koehn, P.; Schroeder, J. Experiments in Domain Adaptation for Statistical Machine Translation. ACL 2007: Second Workshop on Statistical Machine Translation (StatMT '07), Association for Computational Linguistics, pp. 224-227, 2007.
139. Koletnik Korošec, M. The Internet, Google Translate and Google Translator Toolkit - Nuisance or Necessity in Translator Training? Tralogy I - Translation Careers and Technologies: Convergence Points for the Future, p. 17, 2011.
140. Latour, J. Evaluating Statistical Machine Translation from English to Dutch. Proceedings of the 1st Twente Student Conference on IT: Intelligent Interaction, p. 5, 2004.
141. Lavie, A. Evaluating the Output of Machine Translation Systems. The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010), tutorial documentation, p. 86, 2010.
142. Lavie, A.; Agarwal, A. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. Proceedings of the ACL 2007 Workshop on Statistical Machine Translation, pp. 228-231, 2007.
143. Lavie, A.; Sagae, K.; Jayaraman, S. The Significance of Recall in Automatic Metrics for MT Evaluation. In Machine Translation: From Real Users to Research - Lecture Notes in Computer Science, vol. 3265, Springer, pp 134-143, 2004.

144. Lavergne, T.; Allauzen, A.; Le, H.-S. ; Yvon, F. LIMSI's experiments in domain adaptation for IWSLT11. Proceedings of the IWSLT 2011, pp. 62-67, 2011.
145. Läubli, S.; Fishel, M.; Weibel, M.; Volk, M. Statistical Machine Translation for Automobile Marketing Texts. Proceedings of the XIV Machine Translation Summit, p. 265-272, 2013.
146. Liang, P.; Bouchard-Côté, A.; Klein, D.; Taskar, B. An End-to-End Discriminative Approach to Machine Translation. Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (ACL-44), pp. 761-768, 2006.
147. Lin, C.-Y. ROUGE: A Package For Automatic Evaluation Of Summaries. Proceedings of the ACL-04 Workshop: Text Summarization Branches Out, pp. 74-81, 2004
148. Liu, C.; Dahlmeier, D.; Ng, H. T. Better Evaluation Metrics Lead to Better Machine Translation. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 375-384, 2011.
149. Liu, D.; Gildea, D. Improved Tree-to-string Transducer for Machine Translation. Proceedings of the ACL Workshop on Statistical Machine Translation (ACL08-SMT), Association for Computational Linguistics, pp. 62-69, 2008.
150. Lommel, A. Multidimensional Quality Metrics – MQM. Meta-Forum 2013, p. 18, 2013.
151. Lommel, A.; Burchardt, A.; Popovic, M.; Harris, K.; Avramidis, E.; Uszkoreit, H. Using a new analytic measure for the annotation and analysis of MT errors on real data. Proceedings of the 17th Annual Conference of the European Association for Machine Translation, European Association for Machine Translation, Croatian Language Technologies Society, pp. 165-172, 2014.
152. Lommel, A.; Burchardt, A.; Uszkoreit, H. Multidimensional Quality Metrics: A Flexible System for Assessing Translation Quality. [Aslib 2013] Translating and the Computer Conference 35, p. 7, 2013.
153. Lommel, A.; Uszkoreit, H.; Burchardt, A. Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. Revista Tradumàtica: tecnologies de la traducció: Traducció i qualitat, no. 12, pp. 455-463, 2014b.
154. Lopez, A. Statistical Machine Translation. ACM Computing Surveys, vol. 40, no. 3, article 8, p. 49, 2008.

155. Lu, Y.; Wang, L.; Wong, D. F.; Chao, L. S.; Wang, Y.; Oliveira, F. Domain Adaptation for Medical Text Translation Using Web Resources. Proceedings of the Ninth Workshop on Statistical Machine Translation, Association for Computational Linguistics, pp. 233-238, 2014.
156. MacCartney, B. Smoothing. NLP Lunch Tutorial at Stanford University, p. 33, 2005.
157. Madnani, N. iBLEU Interactively Debugging and Scoring Statistical Machine Translation Systems. 2011 Fifth IEEE International Conference on Semantic Computing (ICSC), pp. 213-214, 2011.
158. Madnani, N. Language Models. INFM718G/CMSC838G course on Data-Intensive Information Processing Applications (Lin, J.; Madnani, N.) at University of Maryland, course material, p. 63, 2010.
159. Manning, C. D.; Schütze, H. Foundations of Statistical Natural Language Processing. The MIT Press, p. 620, 1999.
160. Mansour, S.; Ney, H. Simple and Effective Weighted Phrase Extraction for Machine Translation Adaptation. Proceedings of the 9th International Workshop on Spoken Language Translation, pp. 193-200, 2012.
161. Matsoukas, S.; Rosti, A.-V. I.; Zhang, B. Discriminative Corpus Weight Estimation for Machine Translation. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, ACL and AFNLP, pp. 708-717, 2009.
162. Matusov, E.; Leusch, G.; Bender, O.; Ney, H. Evaluating Machine Translation Output with Automatic Sentence Segmentation. Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2005), pp. 138-144, 2005.
163. Mauser, A.; Hasan, S.; Ney, H. 2008. Automatic Evaluation Measures for Statistical Machine Translation System Optimization. Sixth International Conference on Language Resources and Evaluation (LREC'08), pp. 3089-3092, 2008.
164. Melamed D. I.; Green R.; Joseph P. Precision and recall of machine translation. NAACL: 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL 2003), vol. 2, pp. 61-63, 2003.
165. Mohammadi, M.; GhasemAghaee, N. Building Bilingual Parallel Corpora Based on Wikipedia. Proceedings of the 2010 Second International Conference on Computer Engineering and Applications (ICCEA '10), vol. 2, pp. 264-268, 2010.

166. Mohit, B.; Liberato, F.; Hwa, R. Language Model Adaptation for Difficult to Translate Phrases. Proceedings of the 13th Annual Conference of the EAMT, pp. 160-167, 2009.
167. Mooney, R. J. N-Gram Language Models. CS 388: Natural Language Processing course at University of Texas at Austin, course material, p. 22, 2015.
168. Morado Vázquez, L.; Rodríguez Vázquez, S.; Bouillon, P. Comparing forum data post-editing performance using translation memory and machine translation output: a pilot study. Proceedings of the XIV Machine Translation Summit, pp. 249-256, 2013.
169. Munteanu, D. S.; Marcu, D. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. Computational Linguistics Journal, vol. 31, no. 4, pp. 477-504, 2005.
170. Nakov, P. Improving English-Spanish statistical machine translation: experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing. Proceedings of the Third Workshop on Statistical Machine Translation, Association for Computational Linguistics, pp. 147-150, 2008.
171. Nakov, P.; Ng, H. T. Improving Statistical Machine Translation for a Resource-Poor Language Using Related Resource-Rich Languages. Journal of Artificial Intelligence Research, vol. 44, pp. 179-222, 2012.
172. Niehues, J. Adaptation in Machine Translation. Fakultät für Informatik des Karlsruher Instituts für Technologie (KIT), doctoral dissertation, p. 208, 2014.
173. Niehues, J.; Mediani, M.; Herrmann, T.; Heck, M.; Herff, C.; Waibel, A. The KIT Translation system for IWSLT 2010. Proceedings of IWSLT 2010, pp. 93-98, 2010.
174. Niehues, J.; Waibel, A. Detailed Analysis of different Strategies for Phrase Table Adaptation in SMT. Proceedings of the American Machine Translation Association (AMTA2012), p. 10 , 2012.
175. Niehues, J.; Waibel, A. Domain Adaptation in Statistical Machine Translation using Factored Translation Models. Proceedings of the 14th Annual Conference of the European Association for Machine Translation - EAMT 2010, p. 7, 2010.
176. Nießen, S.; Och, F.; Leusch, G.; Ney, H. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. Second International Conference on Language Resources and Evaluation (LREC-2000), pp. 39-45, 2000.
177. O'Brien, S. Introduction to Post Editing: Who, What, How and Where to Next? The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010), tutorial documentation, p. 30, 2010.

178. Och, F. J. Minimum error rate training in statistical machine translation. ACL 2003: 41st Annual Meeting on Association for Computational Linguistics, vol. 1, pp. 160-167, 2003.
179. Och, F. J.; Ney, H. A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, vol. 29, no. 1, pp. 19-51, 2003.
180. Och, F. J.; Ney, H. Improved Statistical Alignment Models. Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, pp. 440-447, 2000.
181. Offersgaard, L.; Hansen, D. H. SMT systems for less-resourced languages based on domain-specific data. Proceedings of The 5th Workshop on Building and Using Comparable Corpora (LREC'12), p. 75-80, 2012.
182. Ohashi, K.; Yamamoto, K.; Saito, K.; Nagata, M. NUT-NTT Statistical Machine Translation System for IWSLT 2005. Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2005), pp. 128-133, 2005.
183. Okpor, M. D. Machine Translation Approaches: Issues and Challenges. IJCSI International Journal of Computer Science Issues, vol. 11, no. 5/2, pp. 159-165, 2014.
184. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. BLEU: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002), pp. 311-318, 2002.
185. Pecina, P.; Toral, A.; van Genabith, J. Simple and Effective Parameter Tuning for Domain Adaptation of Statistical Machine Translation. Proceedings of COLING 2012: Technical Papers, pp. 2209-2224, 2012.
186. Pecina, P.; Toral, A.; Way, A.; Papavassiliou, V.; Prokopidis, P.; Giagkou, M. Towards using web-crawled data for domain adaptation in statistical machine translation. Proceedings of the 15th conference of the European Association for Machine Translation (EAMT 2011), pp. 297-304, 2011.
187. Phillips, J. H.; Van Ess-Dykema, C.; Allison, T.; Gerber, L. Parallel Corpus Development at NVTC. Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010), p. 7, 2010.
188. Plitt, M.; Masselot, F. A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. The Prague Bulletin of Mathematical Linguistics, no. 93, pp. 7-16, 2010.
189. Popović, M.; Ney, H. Towards Automatic Error Analysis of Machine Translation Output. Computational Linguistics Journal, vol. 37, no. 4, Association for Computational Linguistics, pp. 657-688, 2011.

190. Post, M.; Callison-Burch, C.; Osborne, M. Constructing Parallel Corpora for Six Indian Languages via Crowdsourcing. Proceedings of the 7th Workshop on Statistical Machine Translation, Association for Computational Linguistics, pp. 401-409, 2012.
191. Pouliquen, B.; Elizalde, C.; Junczys-Dowmunt, M.; Mazenc, C.; García-Verdugo, J. Large-scale multiple language translation accelerator at the United Nations. Proceedings of the XIV Machine Translation Summit, p. 345-352, 2013.
192. Pouliquen, B.; Mazenc, C.; Elizalde, C.; García-Verdugo, J. Statistical Machine Translation prototype using UN parallel documents. Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT 2012), pp. 12-19, 2012.
193. Rapp, R. Back-translation Score: Automatic MT Evaluation at the Sentence Level without Reference Translations. Proceedings of the ACL-IJCNLP 2009 Conference, Association for Computational Linguistics, pp. 133-136, 2009.
194. Razmara, M.; Foster, G.; Sankaran, B.; Sarkar, A. Mixing Multiple Translation Models in Statistical Machine Translation. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics pp. 940-949, 2012.
195. Reddy, M. V.; Hanumanthappa, M. NLP challenges for machine translation from English to Indian languages. International Journal of Computer Science and Informatics, vol. 3, no. 1, pp. 35, 2013.
196. Reinke, U. State of the Art in Translation Memory Technology. Translation: Computation, Corpora, Cognition. Special Issue on Language Technologies for a Multilingual Europe, vol. 3, no. 1, pp. 27-48, 2013.
197. Roturier, J.; Bensadoun, A. Evaluation of MT Systems to Translate User Generated Content. Proceedings of the 13th Machine Translation Summit (MT Summit XIII), pp. 244-251. 2011.
198. Ruopp, A.; Xia, F. Finding parallel texts on the web using cross-language information retrieval. Proceedings of the 2nd International Workshop on „Cross Lingual Information Access (CLIA) Addressing the Information Need of Multilingual Societies“, pp. 18-25, 2008.
199. Schaefer, F.; Van de Walle, J.; Van den Bogaert, J. Moses SMT as an Aid to Translators in the Production Process. Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT2014), pp. 89-92, 2014.

200. Schwenk, H. Investigations on Large-Scale Lightly-Supervised Training for Statistical Machine Translation. Proceedings of IWSLT 2008, pp. 182-189, 2008.
201. Schwenk, H.; Koehn, P. Large and Diverse Language Models for Statistical Machine Translation. Proceedings of the International Joint Conference on Natural Language Processing, pp. 661-666, 2008.
202. Schwenk, H.; Senellart, A. Translation Model Adaptation for an Arabic/French News Translation System by Lightly-Supervised Training. Proceedings of the MT Summit XII, p. 8, 2009.
203. Seljan, S. Lexical-Functional Grammar of the Croatian Language: Theoretical and Practical Models. Proceedings of the Interlinguistica: XIX International Conference of the Association of Young Linguists, p. 10, 2004.
204. Seljan, S.; Brkić, M.; Kučiš, V. Evaluation of Free Online Machine Translations for Croatian-English and English-Croatian Language Pairs. INFuture2011: The Future of Information Sciences - Information Sciences and e-Society, pp. 331-344, 2011.
205. Seljan, S.; Pavuna, D. Why Machine-Assisted Translation (MAT) Tools for Croatian? Proceedings of 28th International Information Technology Interfaces Conference – ITI, pp. 469-475, 2006.
206. Seljan, S.; Tucaković, M.; Dunđer, I. Human Evaluation of Online Machine Translation Services for English/Russian-Croatian. New Contributions in Information Systems and Technologies - Advances in Intelligent Systems and Computing, vol. 353, Springer, pp. 1089-1098, 2015.
207. Seljan, S.; Vičić, T.; Brkić, M. BLEU Evaluation of Machine-Translated English-Croatian Legislation. Eight International Conference on Language Resources and Evaluation (LREC'12), pp. 2143-2148, 2012.
208. Sennrich, R. Mixture-modeling with unsupervised clusters for domain adaptation in statistical machine translation. Proceedings of the 16th EAMT Conference, European Association for Machine Translation, 185-192, 2012.
209. Sennrich, R. Perplexity minimization for translation model domain adaptation in statistical machine translation. Proceedings of the 13th Conference of the European Chapter of the ACL, Association for Computational Linguistics, pp. 539-549, 2012b.
210. Sennrich, R. The UZH system combination system for WMT 2011. Proceedings of the Sixth Workshop on Statistical Machine Translation, pp. 166-170, 2011.
211. Servan, C.; Schwenk, H. Optimising Multiple Metrics with MERT. The Prague Bulletin of Mathematical Linguistics, vol. 96, no. 1, pp. 109-117, 2011.

212. Shannon, C. E. A Mathematical Theory of Communication. Bell System Technical Journal, vol. 27, no. 3, pp. 379-423.
213. Sharoff, S. Beyond Translation Memories: finding similar documents in comparable corpora. Proceedings of the Translating and the Computer Conference, p. 7, 2012.
214. Sima'an, K. Statistical Machine Translation. Elements of Language Processing and Learning (2013-14) course (Titov, I.), Institute for Logic, Language and Computation at University of Amsterdam, course material, p. 66, 2013.
215. Simard, M.; Fujita, A. A Poor Man's Translation Memory Using Machine Translation Evaluation Metrics. Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA 2012), p. 10, 2012.
216. Sinhal, R. A.; Chandak, M. B. Design Aspects in Machine Translation. Proceedings of the National Conference on Emerging Trends in Computer Science and Information Technology (NCETSIT-2011), pp. 27-34, 2011.
217. Skadiņš, R.; Skadiņa, I.; Pinnis, M.; Vasiļjevs, A.; Hudík, T. Application of Machine Translation in Localization into low-resourced languages. Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT2014), pp. 209-216, 2014.
218. Smith, J. R.; Quirk, C.; Toutanova, K. Extracting parallel sentences from comparable corpora using document level alignment. Proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10), pp. 403-411, 2010.
219. Specia, L. Fundamental and New Approaches to Statistical Machine Translation. Propor 2010 - International Conference on Computational Processing of the Portuguese Language, tutorial documentation, p. 32, 2010.
220. Snover, M.; Dorr, B.; Schwartz, R. Language and Translation Model Adaptation using Comparable Corpora. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 857-866, 2008.
221. Snover, M.; Dorr, B.; Schwartz, R.; Makhoul, J.; Micciula, L.; Weischedel, R. A Study of Translation Error Rate with Targeted Human Annotation. Technical Report: LAMP-TR-126/CS-TR-4755/UMIACS-TR-2005-58, p. 17, 2005.
222. Snover, M., Dorr, B.; Schwartz, R.; Micciulla, L.; Makhoul, J. A Study of Translation Edit Rate with Targeted Human Annotation. Association for Machine Translation in the Americas (AMTA), pp. 223-231, 2006.

223. Søgaard, A.; Johannsen, A.; Plank, B.; Hovy, D.; Martinez, H. What's in a p-value in NLP? Proceedings of the Eighteenth Conference on Computational Language Learning, Association for Computational Linguistics, pp. 1-10, 2014.
224. Sousa, S. C. M. de; Aziz, W.; Specia, L. Assessing the Post-Editing Effort for Automatic and Semi-Automatic. Translations of DVD Subtitles. Proceedings of the International Conference Recent Advances in Natural Language Processing, pp. 97-103, 2011.
225. Specia, L.; Turchi, M.; Cancedda, N.; Dymetman, M.; Cristianini, N. Estimating the Sentence-Level Quality of Machine Translation Systems. Proceedings of the 13th Annual Conference of the EAMT, pp. 28-35, 2009.
226. Stein, D. Machine Translation - Past, Present, and Future. Translation: Computation, Corpora, Cognition (TC3), vol. 3, no. 1, pp. V-XII, 2013.
227. Stüker, S.; Waibel, A. Towards human translations guided language discovery for ASR systems. The first International Workshop on Spoken Languages Technologies for Under-resourced languages (SLTU - 2008), pp. 76-79, 2008.
228. Stymne, S. BLAST: A Tool for Error Analysis of Machine Translation Output. Proceedings of the ACL-HLT 2011 System Demonstrations, Association for Computational Linguistics, pp. 56-61, 2011.
229. Stymne, S. Machine translation evaluation. GSLT/NGSLT course in Machine Translation at Linköping University, course material, p. 11, 2008.
230. Stymne, S. Pre- and Postprocessing for Statistical Machine Translation into Germanic Languages. Proceedings of the ACL-HLT 2011 Student Session, Association for Computational Linguistics, pp. 12-17, 2011.
231. Sun, Y.; Liu, J.; Li, Y. Deploying MT into a Localisation Workflow: Pains and Gains. Proceedings of the 13th Machine Translation Summit, pp. 236-243, 2011.
232. Tiedemann, J. Machine Translation - Phrase-Based Statistical MT. 5LN426 course in Machine translation at Uppsala University, course material, p. 11, 2009.
233. Tiedemann, J.; Stymne, S.; Hardmeier, C. Machine Translation Evaluation. 5LN426/ 5LN711 course in Machine Translation at Uppsala University, course material, p. 10, 2014.
234. Thurmair, G.; Aleksić, V. Creating Term and Lexicon Entries from Phrase Tables. Proceedings of the 16th EAMT Conference, pp. 253-260, 2012.
235. Tillmann, C.; Zhan, T. A Discriminative Global Training Algorithm for Statistical MT. Proceedings of the 21st International Conference on Computational Linguistics and

44th Annual Meeting of the ACL, Association for Computational Linguistics, pp. 721-728, 2006.

236. Tohmetov, T. A.; Ushakov, A. O.; Vanushin, I. S. The Problems Of Machine Translation. Proceedings of the XII All-Russian scientific-practical conference of students, graduate students and young scientists, pp. 267-268, 2014.
237. Tomás, J.; Mas, J.-À.; Casacuberta, F. A Quantitative Method for Machine Translation Evaluation. Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: are evaluation methods metrics and resources reusable?, p. 8, 2003.
238. Tripathi, S.; Sarkhel, J. K. Approaches to machine translation. Annals of Library and Information Studies (ALIS), vol. 57, pp. 388-393, 2010.
239. Tufiş, D.; Barbu, A.-M. Computational bilingual lexicography: automatic extraction of translation dictionaries. Journal of Information Science and Technology, Romanian Academy, vol. 4, no. 3, pp. 325-352, 2001.
240. Turian, J. P.; Shen, L.; Melamed, D. I. Evaluation of Machine Translation and its Evaluation. Proceedings of the MT Summit IX, pp. 386-393, 2003.
241. Turchi, M.; De Bie, T.; Cristianini, N. Learning Performance of a Machine Translation System: a Statistical and Computational Analysis. Proceedings of the Third Workshop on Statistical Machine Translation, Association for Computational Linguistics, pp. 35-43, 2008.
242. Turchi, M.; De Bie, T.; Goutte, C.; Cristianini, N. Learning to Translate: A Statistical and Computational Analysis. Advances in Artificial Intelligence, vol. 2012, p. 15, 2012.
243. Turchi, M.; Goutte, C.; Cristianini, N. Learning Machine Translation from In-domain and Out-of-domain Data. Proceedings of 16th Annual Conference of the European Association for Machine Translation, pp. 305-312, 2012b.
244. Tyers, F. M.; Sánchez-Martínez, F.; Sánchez-Martínez, O.; Forcada, M. L. Free/Open-Source Resources in the Apertium Platform for Machine Translation Research and Development. The Prague Bulletin of Mathematical Linguistics, no. 93 pp. 67-76, 2010.
245. Udovičić, M.; Baždarić, K.; Bilić-Zulle, L.; Petrovečki, M. What we need to know when calculating the coefficient of correlation? Biochemia Medica, vol. 17, no. 1, pp. 10-15, 2007.
246. Ueffing, N.; Haffari, G.; Sarkar, A. Semi-supervised model adaptation for statistical machine translation. Machine Translation Journal, vol. 21, no. 2, pp. 77-94, 2007.

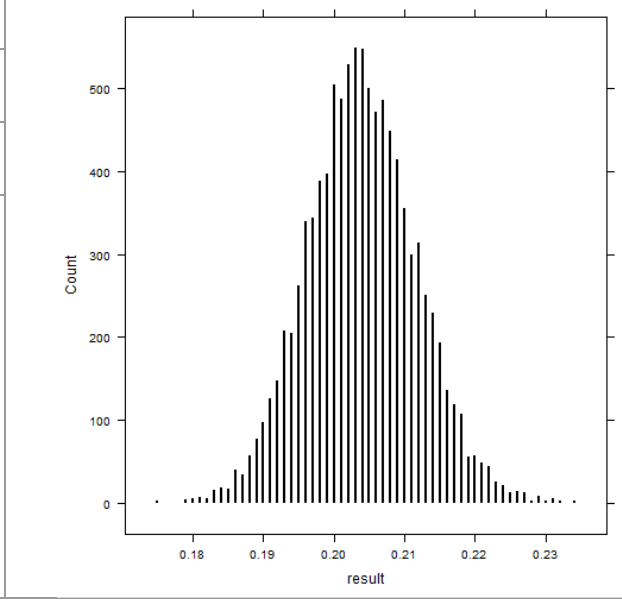
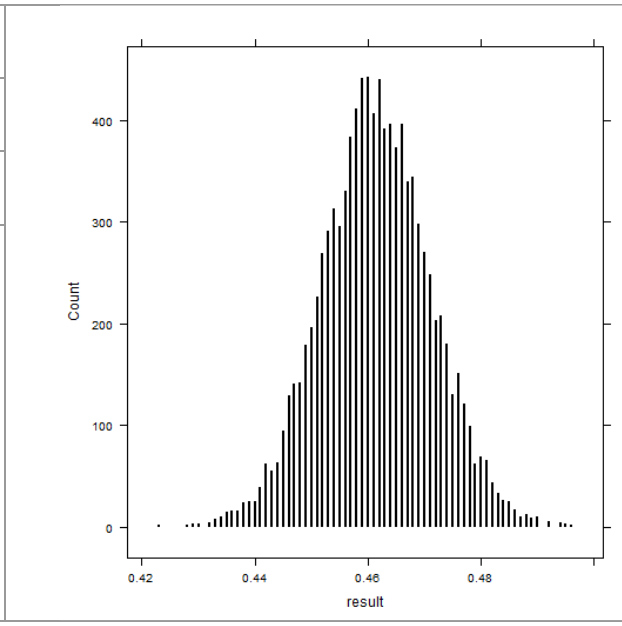
247. Unnikrishnan, P.; Antony, P. J.; Soman, K. P. A Novel Approach for English to South Dravidian Language Statistical Machine Translation System. International Journal on Computer Science and Engineering (IJCSE), vol. 2, no. 8, pp. 2749-2759, 2010.
248. Vertan, C.; Duma, M.-S. Domain Adaptation in Machine Translation. Machine Translation Summit XIV, tutorial documentation, p. 112, 2013.
249. Vilar, D.; Schneider, M.; Burchardt, A.; Wedde, T. Towards the Integration of MT into a LSP Translation Workflow. Proceedings of the 16th EAMT Conference, pp. 73-76, 2012.
250. Vogel, S.; Ney, H.; Tillmann, C. HMM-Based Word Alignment in Statistical Translation. Proceedings of the 16th conference on Computational linguistics (COLING '96), vol. 2, pp. 836-841, 1996.
251. Vogel, S.; Tribble, A. Improving Statistical Machine Translation for a Speech-to-Speech Translation Task. Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP-2002), Workshop on Speech-to-Speech Translation, p. 4, 2002.
252. Wandmacher, T. Adaptive word prediction and its application in an assistive communication system. Neuphilologische Fakultät der Universität Tübingen, doctoral dissertation, p. 191, 2009.
253. Wang, K.; Zong, C.; Su, K.-Y. Integrating Translation Memory into Phrase-Based Machine Translation during Decoding. Proceedings of the 51st Annual Meeting of the ACL, Association for Computational Linguistics, pp. 11-21, 2013.
254. Wang, W.; Macherey, K.; Macherey, W.; Och, F.; Xu, P. Improved Domain Adaptation for Statistical Machine Translation. The Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2012), p. 10, 2012.
255. Watanabe, T.; Sumita, E. Statistical machine translation decoder based on phrase. Proceedings of the INTERSPEECH 2002 Conference, pp. 95-102, 2002.
256. Way, A.; Hassan, H. Statistical Machine Translation: Trends & Challenges. Second International Conference on Arabic Language Resources & Tools, tutorial documentation, p. 174, 2009.
257. Weller, M.; Fraser, A.; Heid, U. Combining Bilingual Terminology Mining and Morphological Modeling for Domain Adaptation in SMT. Proceedings of the 17th Annual Conference of the European Association for Machine Translation, European Association for Machine Translation, Croatian Language Technologies Society, pp. 11-18, 2014.

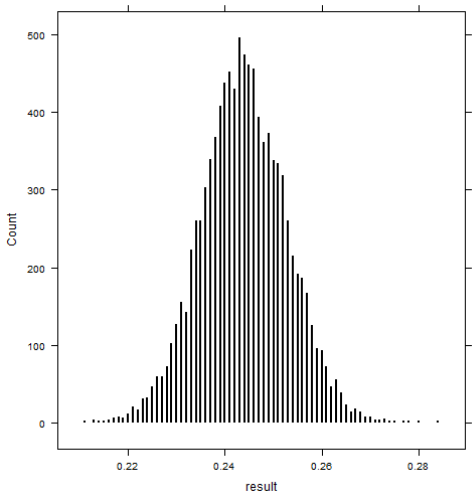
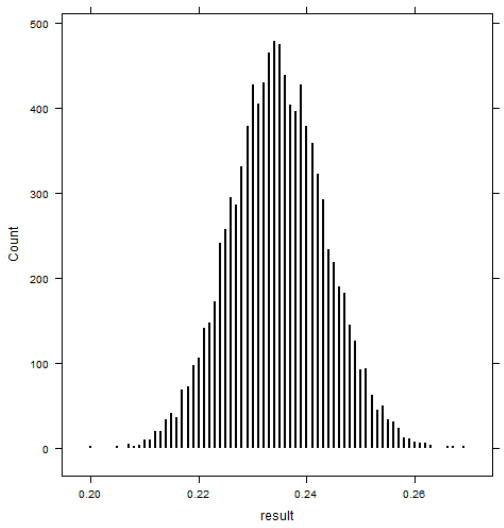
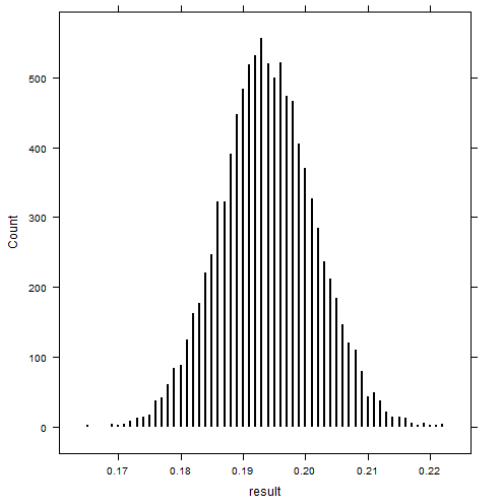
258. Wetzel, D.; Bond, F. Enriching parallel corpora for statistical machine translation with semantic negation rephrasing. ACL 2012: Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-6 '12), pp. 20-29, 2012.
259. Wu, H.; Wang, H. Comparative study of word alignment heuristics and phrase-based SMT. Proceedings of MT Summit XI , p. 8, 2007.
260. Wu, H.; Wang, H.; Liu, Z. Alignment Model Adaptation for Domain-Specific Word Alignment. Proceedings of the 43rd Annual Meeting of the ACL, Association for Computational Linguistics, pp. 467-474, 2005.
261. Wu, H.; Wang, H.; Zong, C. Domain Adaptation for Statistical Machine Translation with Domain. Dictionary and Monolingual Corpora. Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008), pp. 993-1000, 2008.
262. Yamada, K.; Knight, K. A Syntax-based Statistical Translation Model. Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL '01), pp. 523-530, 2001.
263. Yasuda, K.; Zhang, R.; Yamamoto, H.; Sumita, E. . Method of Selecting Training Data to Build a Compact and Efficient Translation Model. International Joint Conference on Natural Language Processing (IJCNLP 2008), pp. 655-660, 2008.
264. Ye, Y.; Zhou, M.; Lin, C.-Y. Sentence Level Machine Translation Evaluation as a Ranking Problem: one step aside from BLEU. Proceedings of the Second Workshop on Statistical Machine Translation, Association for Computational Linguistics, pp. 240-247, 2007.
265. Yıldırım, E.; Tantuğ, A. C. Evaluation of Domain Adaptation Approaches to Improve the Translation Quality. In New Trends in Computational Collective Intelligence: Studies in Computational Intelligence; Camacho, D.; Sang-Wook, K.; Trawiński, B. (Eds.), vol. 572, Springer, pp. 15-26, 2015.
266. Yu, Q.; Max, A.; Yvon, F. Revisiting sentence alignment algorithms for alignment visualization and evaluation. Proceedings of the Fifth Workshop on Building and Using Comparable Corpora (BUCC 2012), pp. 10-16, 2012.
267. Zhang, J.; Zong, C. Learning a Phrase-based Translation Model from Monolingual Data with Application to Domain Adaptation. Proceedings of the 51st Annual Meeting of the ACL, Association for Computational Linguistics, pp. 1425-1434, 2013.

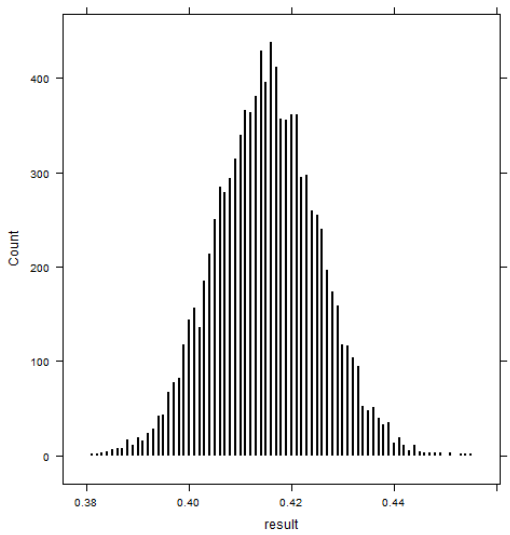
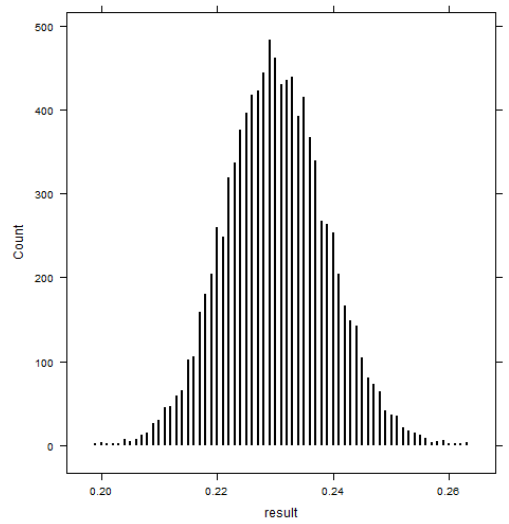
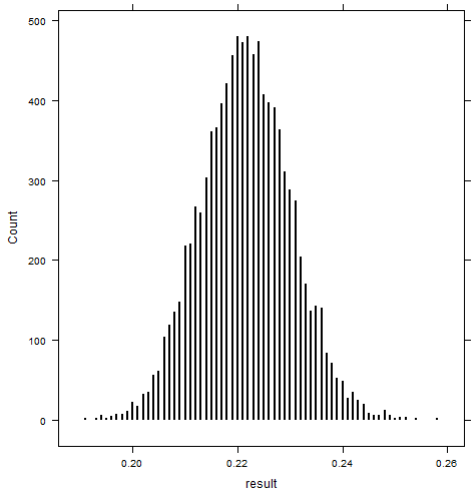
268. Zhang, Y.; Vogel, S. Measuring Confidence Intervals for the Machine Translation Evaluation Metrics. Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-2004), p. 10, 2004.

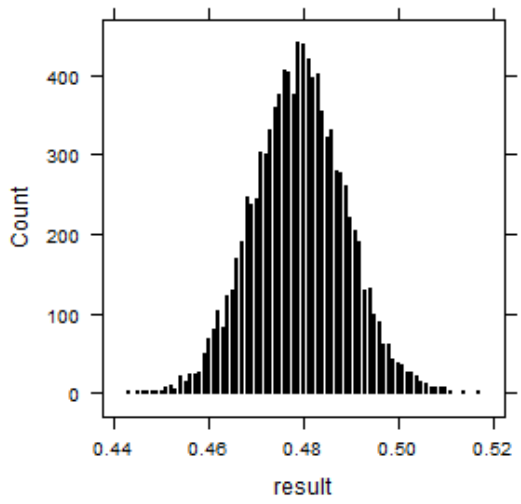
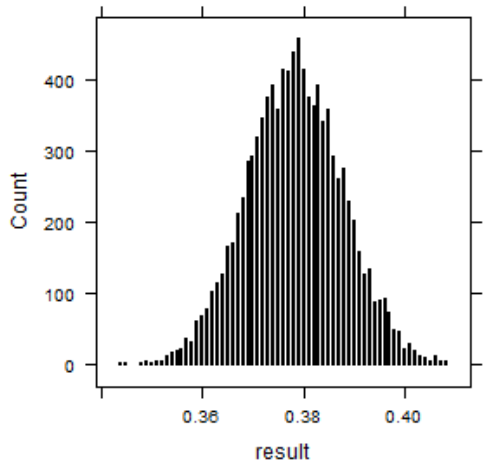
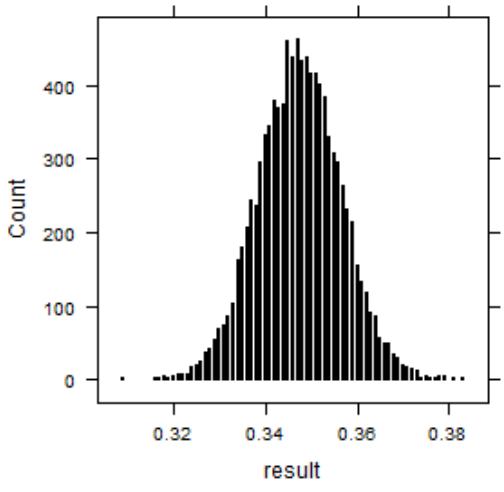
269. Zhang, Y.; Vogel, S.; Waibel, A. Interpreting Bleu/NIST scores: How Much Improvement Do We Need to Have a Better System? Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), p. 4, 2004.

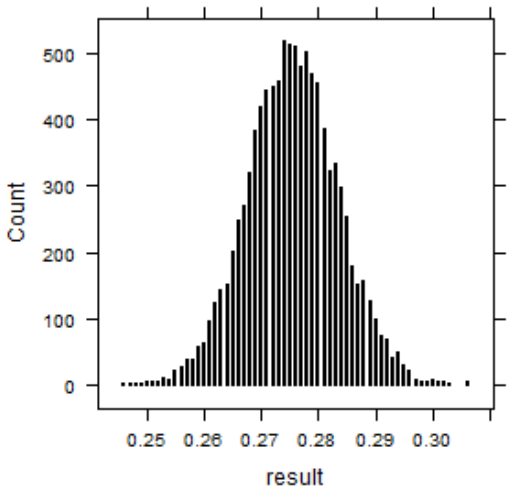
DODATAK A Rezultati metode ponovnog uzorkovanja

hrvatsko-engleski sustavi	**sustav 1**	
Lilliefors (Kolmogorov-Smirnov) test	D=0.0053224, p=0.7048	
Anderson-Darling test	A=0.36405, p=0.4391	
Cramer-von Mises test	W=0.0427, p=0.6322	
Pearson chi-kvadrat test	P=65.824, p=0.8143	
hrvatsko-engleski sustavi	**sustav 2**	
Lilliefors (Kolmogorov-Smirnov) test	D=0.0076774, p=0.1638	
Anderson-Darling test	A=0.51972, p=0.1865	
Cramer-von Mises test	W=0.070173, p=0.2782	
Pearson chi-kvadrat test	P=76.688, p=0.4886	

hrvatsko-engleski sustavi	**sustav 3**	
Lilliefors (Kolmogorov-Smirnov) test	D=0.012046, p=0.002351	
Anderson-Darling test	A=2.1169, p=2.234e-05	
Cramer-von Mises test	W=0.37493, p=4.898e-05	
Pearson chi-kvadrat test	P=90.288, p=0.1428	
hrvatsko-engleski sustavi	**sustav 4**	
Lilliefors (Kolmogorov-Smirnov) test	D=0.0091602, p=0.0569	
Anderson-Darling test	A=0.87786, p=0.02462	
Cramer-von Mises test	W=0.12674, p=0.04866	
Pearson chi-kvadrat test	P=92.768, p=0.1064	
englesko-hrvatski sustavi	**sustav 5**	
Lilliefors (Kolmogorov-Smirnov) test	D=0.010878, p=0.009748	
Anderson-Darling test	A=1.2075, p=0.003799	
Cramer-von Mises test	W=0.19588, p=0.006071	
Pearson chi-kvadrat test	P = 127.22, p=0.0002796	

englesko-hrvatski sustavi	**sustav 6**	
Lilliefors (Kolmogorov-Smirnov) test	D=0.006822, p=0.3131	
Anderson-Darling test	A=0.45854, p=0.2632	
Cramer-von Mises test	W=0.070034, p=0.2794	
Pearson chi-kvadrat test	P = 82.896, p=0.3026	
englesko-hrvatski sustavi	**sustav 7**	
Lilliefors (Kolmogorov-Smirnov) test	D=0.0059528, p=0.53	
Anderson-Darling test	A=0.30388, p=0.5717	
Cramer-von Mises test	W=0.052236, p=0.479	
Pearson chi-kvadrat test	P = 59.296, p=0.933	
englesko-hrvatski sustavi	**sustav 8**	
Lilliefors (Kolmogorov-Smirnov) test	D=0.0052837, p=0.7153	
Anderson-Darling test	A=0.22366, p=0.8252	
Cramer-von Mises test	W=0.02597, p=0.8969	
Pearson chi-kvadrat test	P=47.44, p=0.9968	

hrvatsko-engleski sustavi	**Google Translate**	
Lilliefors (Kolmogorov-Smirnov) test	D=0.0052674, p=0.7197	
Anderson-Darling test	A 0.26448, p=0.6963	
Cramer-von Mises test	W=0.031884, p=0.8207	
Pearson chi-kvadrat test	P=91.68, p=0.1214	
hrvatsko-engleski sustavi	**Yandex Translate**	
Lilliefors (Kolmogorov-Smirnov) test	D=0.0067626, p=0.3261	
Anderson-Darling test	A=0.41399, p=0.3361	
Cramer-von Mises test	W=0.060157, p=0.3765	
Pearson chi-kvadrat test	P=82.288, p=0.3191	
englesko-hrvatski sustavi	**Google Translate**	
Lilliefors (Kolmogorov-Smirnov) test	D=0.006031, p=0.5087	
Anderson-Darling test	A=0.26205, p=0.7045	
Cramer-von Mises test	W=0.029195, p=0.8581	
Pearson chi-kvadrat test	P=61.712, p= 0.8979	

englesko-hrvatski sustavi	**Yandex Translate**	
Lilliefors (Kolmogorov-Smirnov) test	D=0.0073233, p=0.2169	
Anderson-Darling test	A=0.53177, p=0.1741	
Cramer-von Mises test	W=0.078067, p=0.2195	
Pearson chi-kvadrat test	P=77.488, p=0.463	

DODATAK B Ljudska evaluacija sustava za statističko strojno prevođenje

U nastavku su prikazani rezultati ljudske evaluacije sustava za statističko strojno prevođenje na skali od 1 do 4. Zadnji stupci (A i F) predstavljaju prosječne vrijednosti prosudbe evaluatora za adekvatnost i fluentnost.

sustav 1				
Rbr.	**referentan prijevod**	**strojni prijevod**	**A**	**F**
1	Command + Q	Command + Q	4,00	4,00
2	Keyboard shortcuts	Tipkovni prečaci	1,00	1,00
3	this section lists common shortcuts for moving around a document .	in this section popisani najčešći prečaci for kretanje by dokumentu .	1,67	1,00
4	marquee Zoom tool	Označivač for zumiranje	1,00	1,00
5	select Object tool	tools to select objekata	2,33	2,33
6	crop tool	tools for obrezivanje	1,67	1,67
7	cycle through drawing markup tools : arrow , Line , Rectangle , Oval , Polygon Line , Polygon , pencil tool , Eraser tool	Kruženje tools for označavanje crteža : Strelica , Linija , Pravokutnik , Elipsa , Linija poligona , Poligon , pencil , Gumica	1,00	1,00
8	cycle through attach tools : attach file , Record Audio Comment	kružno kretanje through alate for prilaganje : Priloži datoteku , tools for a picture zvučnih komentara	1,00	1,00
9	keys for navigating a PDF	tipke for navigaciju within PDF-a	1,00	1,00
10	move focus to menus (Windows , UNIX) ; expand first menu item (UNIX)	Premještanje žarišta on izbornike (Windows , UNIX) ; proširenje the first party izbornika (UNIX)	1,00	1,00
11	move focus to toolbar in browser	Premještanje žarišta on alatnu ribbons in pregledniku	1,00	1,00
12	move focus to next tab in a tabbed dialog box	Premještanje žarišta on next phonecard in kartičnom dijaloškom frame	1,00	1,33
13	move to next search result and highlight it in the document	Premještanje on next result pretraživanja and his isticanje in dokumentu	1,00	1,00
14	keys for working with navigation panels	tipke for working with pločama for navigaciju	1,67	1,67
15	move among the elements of the active navigation panel	kretanje among elementima aktivne records for navigaciju	1,67	1,00

16	up arrow or Down arrow	Strelica up or Strelica down	1,67	1,67
17	move focus to previous item in a navigation panel	Premještanje žarišta on prethodnu stavku on navigacijskoj record	1,00	1,00
18	move to previous pane	Premještanje on prethodno okno	1,00	1,00
19	Reflow a tagged PDF , and return to unreflowed view	change prijeloma PDF-a with strukturnim oznakama and taking a on a without changes prijeloma	1,00	1,00
20	activate and deactivate Read Out Loud	Aktiviranje and deaktiviranje Reading aloud	1,00	1,00
21	after you create page thumbnails , you can embed them in the PDF .	after stvaranja , sličice pages can ugraditi in PDF .	1,00	1,00
22	Embedding prevents the page thumbnails from redrawing each time you click the Pages button , often a time-consuming process .	Ugradnja sprječava again iscrtavanje sličica pages every time push the play button pages because that can at all to insist upon much time .	1,00	1,00
23	in the Pages panel , choose Embed All Page Thumbnails or Remove Embedded Page Thumbnails from the options menu .	on the record pages with izbornika Opcije choose your Ugradi all sličice pages or Ukloni ugrađene sličice pages .	1,00	1,00
24	Embed or unembed page thumbnails in a PDF Portfolio	Ugradnja or removing sličica pages of PDF portfelja	1,00	1,00
25	to embed page thumbnails , click Embed Page Thumbnails , and then click Run Sequence .	you would ugradili sličice pages , kliknite Ugradi sličice pages , and then Pokreni slijed .	1,00	1,00
26	follow the instructions provided .	follow dobivene instructions .	2,67	2,67
27	choose one of the following file formats : BMP , GIF , JPEG , PCX , PNG , EPS / PS , or TIFF .	choose your one of next formata datoteka : BMP , GIF , JPEG , PCX , PNG , EPS Musashi PS or TIFF .	2,33	2,00
28	Page thumbnails are miniature previews of the pages in a document	Sličice pages they minijaturni prikazi pages papers .	1,00	1,00
29	you can use page thumbnails to jump quickly to a selected page or to adjust the view of the page .	Sličice pages in can use for quick jump on odabranu page or prilagodbu presenting pages .	1,00	1,00
30	in adobe Reader ® , when you move a page thumbnail , you move the corresponding page .	when the program Adobe Reader do premještate sličicu pages , actually premještate odgovarajuću page .	1,00	1,33
31	in acrobat , when you move , copy , or delete a page thumbnail , you move , copy , or delete the corresponding page .	when the program Acrobat premještate , kopirate or brišete sličice pages , actually premještate , kopirate or brišete odgovarajuću page .	1,00	1,67
32	Page thumbnails appear in the navigation pane .	in navigacijskom will oknu whim sličice pages .	1,00	1,00
33	define the tabbing order	Definiranje redoslijeda kretanja	1,00	1,00
34	in the Pages panel , you can set the order in which a user tabs through form fields , links , and comments for each page .	on the record pages can for every page zadati redoslijed which will korisnik strain on tipku Tab move between fields obrasca , tie and .	1,33	1,00
35	select a page thumbnail , and choose Page properties from the options menu .	choose your sličicu pages , and then with izbornika Opcije choose your the qualities of pages .	1,00	1,00
36	moves in the order specified by the	moving redoslijedom definiranim in	1,00	1,00

	authoring application .	autorskoj aplikaciji .		
37	if the document was created in an earlier version of acrobat , the tab order is Unspecified by default .	if he document made in my old verziji Acrobata , zadana value redoslijeda kretanja is not a definite .	1,00	1,00
38	about bookmarks	about knjižnim oznakama	1,00	2,00
39	each bookmark goes to a different view or page in the document .	every knjižna brand pridružuje back različitom ghost or page papers .	1,00	1,00
40	in acrobat , you can set bookmark destinations as you create each bookmark .	occasion stvaranja every knjižne visible in program Acrobat can ask income .so for knjižne visible .	1,00	1,00
41	Bookmarks can also perform actions , such as executing a menu item or submitting a form .	Knjižne visible can and izvršavati action , like izvršavanja stavki izbornika or of sending obrasca .	1,00	1,00
42	Bookmarks act as a table of contents for some PDFs .	Knjižne get visible for some PDF-ove acts like pregledi sadržaja .	1,33	1,00
43	open the page where you want the bookmark to link to , and adjust the view settings .	open the page with which want connect knjižnu tag and prilagodite postavke presenting .	1,67	1,00
44	if you don ' t select a bookmark , the new bookmark is automatically added at the end of the list	if not odaberete knjižnu tag , new knjižna brand automatically will add on end popisa .	1,67	1,00
45	type or edit the name of the new bookmark .	Upišite or Fix the name new knjižne visible	1,00	1,00
46	in Reader , you can make bookmarks easier to read by changing their text appearance .	in program Reader knjižne visible can do čitljivijima if change appearance script .	1,67	1,67
47	wrap text in a long bookmark	Prelamanje script in long knjižnoj the mark	1,00	1,00
48	click the Bookmarks button , and choose Wrap Long Bookmarks from the options menu .	Kliknite button Knjižne visible and with izbornika Opcije choose your break rainbow knjižne visible .	1,00	1,33
49	you can change the appearance of a bookmark to draw attention to it .	you can change appearance knjižne visible to you is istaknuli .	2,00	1,67
50	in the Bookmarks panel , select one or more bookmarks .	on the record Knjižne visible choose your one or more knjižnih brand .	2,00	2,00
51	in the document pane , move to the location you want to specify as the new destination .	trapped in a cave-in papers move on odredište which want give as new odredište .	1,00	1,00
52	in the Bookmark properties dialog box , click Actions .	in dijaloškom frame the qualities of knjižne visible kliknite action .	1,67	1,00
53	Deleting a bookmark deletes any bookmarks that are subordinate to it .	brisanje knjižne visible izbrisat will and everything podređene knjižne visible .	1,67	1,00
54	you can nest a list of bookmarks to show a relationship between topics .	list knjižnih brand can ugnijezditi you would ilustrirali relations between subject .	1,00	1,00
55	Nesting creates a parent / child relationship .	Gniježđenje breeds relationship nadređeno Musashi podređeno .	1,00	1,00
56	Nesting a bookmark (left) , and the result (right)	Gniježđenje knjižne visible (left) and result (starboard)	1,00	1,00
57	move bookmarks out of a nested	Premještanje knjižnih tag number	1,00	1,00

	position	from ugniježđenog position		
58	from the options menu , choose expand Top-Level Bookmarks or Collapse Top-Level Bookmarks .	with izbornika opcija choose your proširi knjižne visible most a or Sažmi knjižne visible most a .	1,00	1,00
59	tagged bookmarks give you greater control over page content than do regular bookmarks .	Strukturirane knjižne visible pružaju you a bigger control sadržajem pages of usual knjižnih brand .	1,00	1,00
60	converted web pages typically include tagged bookmarks .	Pretvorene web-stranice usually sadrže označene knjižne visible .	1,00	1,00
61	select the structure elements you want specified as tagged bookmarks .	choose your element strukture which want give as strukturirane knjižne visible .	1,67	1,67
62	Edit tags with the tags tab	Uređivanje strukturnih brand by a card Strukturne visible	1,00	1,00
63	add multimedia to PDFs	Dodavanje multimedije in PDF-ove	1,33	1,33
64	drag a rectangle where you want to create a link .	Opišite pravokutnik on place on which want to create contact .	1,00	1,00
65	select the destination file and click select .	choose your odredišnu datoteku and kliknite pick .	1,00	1,67
66	select the options you want in the create link dialog box .	choose your željene opcije in dijaloškom frame creating connections .	1,67	2,00
67	changing the properties of an existing link affects only the currently selected link .	mijenjanje svojstava postojeće connections it does only on trenutačno odabranu contact .	1,00	1,00
68	select the link tool and double-click the link rectangle .	choose your tools for connections and dvokliknite pravokutnik connections .	1,00	1,00
69	select the Locked option if you want to prevent users from accidentally changing your settings .	choose your opciju locked if you want to prevent accidently mijenjanje postavki .	1,67	2,00
70	you can attach PDFs and other types of files to a PDF .	PDF-u can prilagati PDF-ove and other kinds datoteka .	1,67	1,00
71	in the Add files dialog box , select the file you want to attach , and click Open .	in dijaloškom frame Give files choose your datoteku ever want priložiti and kliknite Open .	1,67	1,00
72	in the attachments panel , select an attachment , and then choose Delete attachment from the options menu .	on the record Privici choose your privitak and with izbornika opcija choose your Izbriši privitak .	1,00	1,00
73	click Use advanced Search options at the bottom of the window , and then select include attachments .	on the window kliknite She napredne opcije pretraživanja , and then choose your turn on the privitke .	1,00	1,00
74	actions are set in the properties dialog box .	action back postavljaju in dijaloškom frame the qualities of .	1,00	1,00
75	add actions with page thumbnails	Dodavanje action with sličicama pages	1,00	1,00
76	to enhance the interactive quality of a document , you can specify actions , such as changing the zoom value , to occur when a page is opened or closed .	you would poboljšali interaktivnost papers , can you define it action , like mijenjanja 2nd zumiranja , who pokreću when pages open or close .	1,00	1,00

77	Executes a specified menu command as the action .	izvršava navedenu order izbornika as action .	1,00	1,00
78	plays the specified sound file .	Reproducira navedenu zvučnu datoteku .	1,00	1,00
79	plays a specified movie that was created as acrobat 6-compatible .	Reproducira film kompatibilan with Acrobatom 6 .	1,33	1,00
80	before you add this action , specify the appropriate layer settings .	before dodavanja these action navedite appropriate postavke sloja .	1,00	1,33
81	Toggles between showing and hiding a field in a PDF document .	he keeps get between pokazivanja and sakrivanja fields in PDF dokumentu .	1,00	1,00
82	Triggers determine how actions are activated in Media clips , pages , and form fields .	Okidači određuju way pokretanja action in medijskim isječcima you on stranicama and in the fields obrazaca .	1,00	1,00
83	when the page containing the Media clip becomes the current page .	when pages with medijskim isječkom becomes trenutna pages .	1,00	1,00
84	you can also use JavaScript with PDF forms and batch sequences .	with PDF obrascima and skupnim sljedovima can use and JavaScript .	1,00	1,00
85	tagged web bookmarks are initially all at the same level , but you can rearrange them and nest them in family groups to help keep track of the hierarchy of material on the web pages .	Strukturirane knjižne visible for web in the outset on the same razinama , but them can premještati and ugnijezditi in sections you would easier followed hijerarhiju material on web-stranicama .	1,00	1,00
86	you can display a dialog box with the current page ' s URL , title , date and time downloaded , and other information .	can you to turn a dijaloškog frame number with URL-om present pages , naslovom , railway and time preuzimanja those other informacijama .	1,00	1,00
87	the browser opens in a new application window to the page you specify .	Preglednik will open in new aplikacijskom window on navedenoj page .	1,33	1,67
88	drag a rectangle to define the first article box .	Opišite pravokutnik to define first frame column .	1,00	1,00
89	Use the Article tool to create, display, and make changes to an article box in the PDF document.	for creating , pregledavanje and mijenjanje frame number articles in PDF dokumentu use some tools for articles .	1,67	2,00
90	when editing a batch sequence , click output options .	occasion uređivanja skupnog slijeda kliknite on Opcije gate .	1,00	1,33
91	optimizing : fast web View option	optimizacija : opcija brzog presenting for web	1,00	1,00
92	Downsample	Smanjivanje number piksela	1,00	1,00
93	Reduces file size by eliminating unnecessary pixel data .	smaller datoteku uklanjanjem nepotrebnih piksela .	1,00	1,00
94	Disables all actions related to submitting or importing form data , and resets form fields .	Onemogućuje all action with slanjem or uvozom podataka from obrazaca and again postavlja fields obrazaca .	1,00	1,00
95	form data is merged with the page to become page content .	facts obrazaca stapaju with stranicom and station part of her sadržaja .	1,00	1,00
96	removes all versions of an image except the one destined for on-	Uklanja all verzije pictures except those namijenjene prikazivanju on	1,33	1,00

	screen viewing .	zaslonu .		
97	embedded thumbnails , deleting	ugrađene sličice , brisanje	1,00	1,00
98	fragmented images , merging	fragmentirane pictures , spajanje	1,67	1,00
99	removes embedded search indexes , which reduces file size .	Uklanja ugrađena kazala for pretraživanje you grow shorter datoteku .	1,00	1,00
100	removes all bookmarks from the document .	Uklanja all knjižne map from papers .	1,00	1,00
		sustav 2		
Rbr.	**referentan prijevod**	**strojni prijevod**	**A**	**F**
1	Command + Q	Command + Q	4,00	4,00
2	Keyboard shortcuts	keyboard shortcuts	4,00	4,00
3	this section lists common shortcuts for moving around a document .	in this list are section najčešći shortcuts to navigate by the document .	2,67	2,33
4	marquee Zoom tool	the Marquee Zoom tool	4,00	4,00
5	select Object tool	the Select Object tool	4,00	4,00
6	crop tool	Crop Tool	4,00	4,00
7	cycle through drawing markup tools : arrow , Line , Rectangle , Oval , Polygon Line , Polygon , pencil tool , Eraser tool	Kruženje tools to highlight art : arrow , lines , rectangle , polygon , Poligon Elipsa , lines , the Pencil tool , the Pencil	2,67	3,00
8	cycle through attach tools : attach file , Record Audio Comment	Kružno navigate through the tools for attaching : Attach A File , the Record Audio Comment tool	2,67	3,00
9	keys for navigating a PDF	the navigation within a PDF	3,00	4,00
10	move focus to menus (Windows , UNIX) ; expand first menu item (UNIX)	move focus to izbornike (Windows , UNIX) ; expand the first menu items (UNIX)	3,00	3,00
11	move focus to toolbar in browser	move focus to the toolbar in the browser .	4,00	3,67
12	move focus to next tab in a tabbed dialog box	move focus to the next tab in kartičnom dialog box	3,00	3,00
13	move to next search result and highlight it in the document	move the search result in the document and to highlight	3,00	2,67
14	keys for working with navigation panels	the to work with navigation panels	2,67	2,33
15	move among the elements of the active navigation panel	move among elements active navigation panel	3,00	2,67
16	up arrow or Down arrow	arrow up or down arrow	4,00	3,33
17	move focus to previous item in a navigation panel	move focus to the previous item in the navigation panel	4,00	4,00
18	move to previous pane	move to a pane	2,67	4,00
19	Reflow a tagged PDF , and return to unreflowed view	reflow a tagged PDF and return to unreflowed view	4,00	4,00
20	activate and deactivate Read Out Loud	activate and deactivate Read Out Loud	4,00	4,00
21	after you create page thumbnails , you can embed them in the PDF .	after you create , page thumbnails , you can embed in the PDF .	4,00	2,67
22	Embedding prevents the page	embedding prevents redrawing of the	3,00	2,67

	thumbnails from redrawing each time you click the Pages button , often a time-consuming process .	page thumbnails each time you click the Pages button , because that can require a many time .		
23	in the Pages panel , choose Embed All Page Thumbnails or Remove Embedded Page Thumbnails from the options menu .	in the Pages panel , from the Options menu , choose Embed All page thumbnails or Remove embedded the page thumbnails .	2,67	2,67
24	Embed or unembed page thumbnails in a PDF Portfolio	embedding or remove page thumbnails from the PDF Portfolio	3,00	2,67
25	to embed page thumbnails , click Embed Page Thumbnails , and then click Run Sequence .	to embed the page thumbnails , click Embed page thumbnails , and then Run Sequence .	4,00	4,00
26	follow the instructions provided .	follow the resulting instructions .	2,67	4,00
27	choose one of the following file formats : BMP , GIF , JPEG , PCX , PNG , EPS / PS , or TIFF .	choose one of the following file formats : BMP , GIF , JPEG , PCX , PNG , EPS / PS or TIFF .	4,00	4,00
28	Page thumbnails are miniature previews of the pages in a document	page thumbnails are minijaturni views document pages .	2,67	3,00
29	you can use page thumbnails to jump quickly to a selected page or to adjust the view of the page .	page thumbnails in you can use to quickly jump to the selected page or customize the Page Display .	3,33	3,00
30	in adobe Reader ® , when you move a page thumbnail , you move the corresponding page .	when in Adobe Reader ® premještate a page thumbnail , you premještate the appropriate page .	1,67	2,33
31	in acrobat , when you move , copy , or delete a page thumbnail , you move , copy , or delete the corresponding page .	when in Acrobat premještate , copy or delete is page thumbnails , you premještate , copy or delete is the appropriate page .	2,67	2,00
32	Page thumbnails appear in the navigation pane .	in the navigation pane appear the page thumbnails .	3,67	2,67
33	define the tabbing order	define the tabbing order	4,00	4,00
34	in the Pages panel , you can set the order in which a user tabs through form fields , links , and comments for each page .	in the Pages panel , you can for each page zadati the order in which the user pressing Tab move between form fields , links , and comments .	3,00	3,00
35	select a page thumbnail , and choose Page properties from the options menu .	select a page thumbnail , and then choose Properties from the Options menu of the page .	3,33	3,33
36	moves in the order specified by the authoring application .	moves order definiranim in the authoring application .	2,67	2,00
37	if the document was created in an earlier version of acrobat , the tab order is Unspecified by default .	if the document is created in starijoj version of Acrobat , the default tabbing order is Is Not specified .	3,00	2,67
38	about bookmarks	about bookmarks	4,00	4,00
39	each bookmark goes to a different view or page in the document .	each bookmark associated with the različitom view or page of the document .	2,00	3,00
40	in acrobat , you can set bookmark destinations as you create each bookmark .	when you create each bookmarks in Acrobat , you can identify specific destinations to bookmarks .	3,00	3,00
41	Bookmarks can also perform actions	bookmarks can izvršavati actions ,	2,67	3,00

	, such as executing a menu item or submitting a form .	such as izvršavanja items menu or send the form .		
42	Bookmarks act as a table of contents for some PDFs .	bookmarks for some PDFs ponašaju as reviews content .	1,67	2,33
43	open the page where you want the bookmark to link to , and adjust the view settings .	open the page that you want to connect bookmark , and adjust the settings view .	4,00	3,00
44	if you don ' t select a bookmark , the new bookmark is automatically added at the end of the list	if you do not select a bookmark , the new bookmark automatically will be added to the end of the list .	4,00	3,33
45	type or edit the name of the new bookmark .	type or edit a new bookmarks .	3,00	3,00
46	in Reader , you can make bookmarks easier to read by changing their text appearance .	in Reader , bookmarks , you can do čitljivijima if you change the appearance of text .	2,67	2,67
47	wrap text in a long bookmark	Wrapping text in dugoj to a bookmark	1,67	2,00
48	click the Bookmarks button , and choose Wrap Long Bookmarks from the options menu .	click the Bookmarks button , and from the Options menu , choose wrap long bookmarks .	4,00	4,00
49	you can change the appearance of a bookmark to draw attention to it .	you can change how the bookmarks to highlight .	2,67	2,00
50	in the Bookmarks panel , select one or more bookmarks .	in the Bookmarks panel , select one or more bookmarks .	4,00	4,00
51	in the document pane , move to the location you want to specify as the new destination .	in the document pane , navigate to the destination you want to specify as a new target .	4,00	4,00
52	in the Bookmark properties dialog box , click Actions .	in the Bookmark Properties dialog box , click the Actions .	4,00	4,00
53	Deleting a bookmark deletes any bookmarks that are subordinate to it .	delete bookmarks deletes the and all its bookmarks .	1,67	2,00
54	you can nest a list of bookmarks to show a relationship between topics .	the list of bookmarks you can ugnijezditi to ilustrirali relationships between skin .	1,67	1,33
55	Nesting creates a parent / child relationship .	Gniježđenje creates relationship nadređeno / Child .	1,67	1,67
56	Nesting a bookmark (left) , and the result (right)	Gniježđenje bookmarks (left) , and the result (right)	2,00	1,67
57	move bookmarks out of a nested position	move bookmarks from a nested position	4,00	3,67
58	from the options menu , choose expand Top-Level Bookmarks or Collapse Top-Level Bookmarks .	from the options menu , select Expand top-level Bookmarks or Collapse top-level Bookmarks .	4,00	4,00
59	tagged bookmarks give you greater control over page content than do regular bookmarks .	tagged bookmarks you have greater control the page content names bookmarks .	2,67	2,67
60	converted web pages typically include tagged bookmarks .	converted web pages typically include the bookmarks .	3,00	4,00
61	select the structure elements you want specified as tagged bookmarks	select the elements structure that you want to specify as a tagged	3,33	3,00

	.	bookmarks .		
62	Edit tags with the tags tab	Edit tags with the Tags tab	4,00	4,00
63	add multimedia to PDFs	adding multimedia to PDFs	3,67	4,00
64	drag a rectangle where you want to create a link .	drag a rectangle to where you want to create a link .	4,00	4,00
65	select the destination file and click select .	select the target file , and click Select .	3,67	4,00
66	select the options you want in the create link dialog box .	select the options you want in the Create Link dialog box .	4,00	4,00
67	changing the properties of an existing link affects only the currently selected link .	changing the properties existing links affects only trenutačno the selected link .	2,67	3,00
68	select the link tool and double-click the link rectangle .	select the Link tool , and then double-click the link rectangle .	4,00	4,00
69	select the Locked option if you want to prevent users from accidentally changing your settings .	choose Locked if you want to prevent inadvertently change the settings .	4,00	3,00
70	you can attach PDFs and other types of files to a PDF .	the PDF , you can prilagati PDFs and other types of files .	2,67	2,33
71	in the Add files dialog box , select the file you want to attach , and click Open .	in the dialog box , Add Files , select the file you want to attach , and click Open .	4,00	4,00
72	in the attachments panel , select an attachment , and then choose Delete attachment from the options menu .	in the Attachments from the options menu , select an attachment , and choose Delete an attachment .	3,33	3,00
73	click Use advanced Search options at the bottom of the window , and then select include attachments .	in the bottom of the window click Use Advanced Search Options , and then select Include attachments .	4,00	3,33
74	actions are set in the properties dialog box .	action is postavljaju in the Properties dialog box .	2,67	2,67
75	add actions with page thumbnails	add actions with page thumbnails	4,00	4,00
76	to enhance the interactive quality of a document , you can specify actions , such as changing the zoom value , to occur when a page is opened or closed .	to improve interactivity the document , you can define actions , such as changing its the zoom , which is start when the page is open it .	3,00	2,00
77	Executes a specified menu command as the action .	run the command menu as action .	2,67	3,00
78	plays the specified sound file .	plays the specified zvučnu file .	2,67	2,67
79	plays a specified movie that was created as acrobat 6-compatible .	plays film compatible with Acrobat 6 .	4,00	3,00
80	before you add this action , specify the appropriate layer settings .	before you add this action specify the appropriate settings layer .	3,33	2,00
81	Toggles between showing and hiding a field in a PDF document .	Prebacuje between pokazivanja and hiding fields in the PDF .	2,67	2,33
82	Triggers determine how actions are activated in Media clips , pages , and form fields .	Okidači determine how run actions in media clips , and on the pages in the form fields .	2,67	2,00
83	when the page containing the Media clip becomes the current page .	when pages with media isječkom becomes the current page .	2,67	2,67

84	you can also use JavaScript with PDF forms and batch sequences .	with PDF forms and batch sequences you can also use JavaScript .	4,00	3,33
85	tagged web bookmarks are initially all at the same level , but you can rearrange them and nest them in family groups to help keep track of the hierarchy of material on the web pages .	tagged bookmarks for are initially in the same levels , but you can move and ugnijezditi in assemblies to help you pratili hierarchy material on the web pages .	2,67	2,00
86	you can display a dialog box with the current page ' s URL , title , date and time downloaded , and other information .	you can include view URL-om dialog box to the current page , title , Date and vremenom downloads , and other information .	2,67	2,00
87	the browser opens in a new application window to the page you specify .	Viewer opens in the new navedenoj aplikacijskom window on the page .	2,67	2,00
88	drag a rectangle to define the first article box .	drag a rectangle to define the first the article .	2,33	3,00
89	Use the Article tool to create, display, and make changes to an article box in the PDF document.	to create it , viewing and modify the article in the PDF , use the for articles .	2,67	2,67
90	when editing a batch sequence , click output options .	when editing in the batch sequence click Output Options .	4,00	4,00
91	optimizing : fast web View option	Optimizing : option your Fast Web View	3,00	2,67
92	Downsample	downsampling	3,67	3,67
93	Reduces file size by eliminating unnecessary pixel data .	the file size removing nepotrebnih pixels .	2,67	2,33
94	Disables all actions related to submitting or importing form data , and resets form fields .	disables any actions associated with slanjem or importing the form data and again sets the form fields .	2,67	3,33
95	form data is merged with the page to become page content .	form data stapaju with page and part njezina content .	1,67	1,67
96	removes all versions of an image except the one destined for on-screen viewing .	removes all versions images except those to be such as no on the screen .	2,67	2,00
97	embedded thumbnails , deleting	embedded thumbnails , deleting	4,00	4,00
98	fragmented images , merging	fragmentirane images , merge	2,67	2,67
99	removes embedded search indexes , which reduces file size .	removes embedded index search and reduces the file .	3,00	3,67
100	removes all bookmarks from the document .	removes all bookmarks from the document .	4,00	4,00
		sustav 3		
Rbr.	**referentan prijevod**	**strojni prijevod**	**A**	**F**
1	Command + Q	Command + Q	4,00	4,00
2	Keyboard shortcuts	tipkovni prečaci	1,00	1,00
3	this section lists common shortcuts for moving around a document .	in this are section najčešći popisani prečaci for kretanje by dokumentu .	1,67	1,67
4	marquee Zoom tool	zumiranje Označivač for	1,00	1,00
5	select Object tool	objekata Alat to select	1,33	1,00
6	crop tool	obrezivanje Alat for	1,00	1,00

7	cycle through drawing markup tools : arrow , Line , Rectangle , Oval , Polygon Line , Polygon , pencil tool , Eraser tool	tools for označavanje Kruženje crteža : Strelica , linija , pravokutnik , Elipsa , linija poligona , Poligon , Olovka , Gumica	1,00	1,33
8	cycle through attach tools : attach file , Record Audio Comment	Kružno kretanje through alate for prilaganje : Priloži datoteku , Alat for snimanje zvučnih komentara	1,00	1,00
9	keys for navigating a PDF	PDF-a navigaciju within tipke for	1,67	1,00
10	move focus to menus (Windows , UNIX) ; expand first menu item (UNIX)	premještanje žarišta on izbornike (Windows , UNIX) ; proširenje first stavke izbornika (UNIX)	1,33	1,00
11	move focus to toolbar in browser	premještanje žarišta on pregledniku alatnu traku in	1,00	1,00
12	move focus to next tab in a tabbed dialog box	premještanje žarišta on kartičnom dijaloškom okviru sljedeću karticu in	1,00	1,00
13	move to next search result and highlight it in the document	premještanje on the next result pretraživanja and it dokumentu isticanje in	1,33	1,00
14	keys for working with navigation panels	tipke for working with navigaciju pločama for	1,33	1,00
15	move among the elements of the active navigation panel	kretanje among elementima aktivne navigaciju ploče for	1,00	1,00
16	up arrow or Down arrow	Strelica Strelica up or down	1,67	1,00
17	move focus to previous item in a navigation panel	premještanje žarišta on navigacijskoj ploči prethodnu stavku on	1,00	1,00
18	move to previous pane	prethodno okno premještanje on	1,00	1,00
19	Reflow a tagged PDF , and return to unreflowed view	change prijeloma PDF-a with strukturnim oznakama and vraćanje on prijeloma prikaz without changes	1,00	1,00
20	activate and deactivate Read Out Loud	Reading aloud deaktiviranje aktiviranje and	1,67	1,00
21	after you create page thumbnails , you can embed them in the PDF .	after you can stvaranja , sličice pages ugraditi in PDF .	1,67	1,00
22	Embedding prevents the page thumbnails from redrawing each time you click the Pages button , often a time-consuming process .	ugradnja iscrtavanje sprječava again sličica pages each time kliknete Stranice button , because that can zahtijevati many time .	1,67	1,33
23	in the Pages panel , choose Embed All Page Thumbnails or Remove Embedded Page Thumbnails from the options menu .	on ploči Stranice with izbornika Opcije odaberite Ugradi all sličice pages or Ukloni ugrađene sličice pages .	1,00	1,00
24	Embed or unembed page thumbnails in a PDF Portfolio	ugradnja or removing portfelja sličica pages from PDF	1,67	1,00
25	to embed page thumbnails , click Embed Page Thumbnails , and then click Run Sequence .	to ugradili sličice pages , kliknite Ugradi sličice pages , and then Pokreni slijed .	1,00	1,00
26	follow the instructions provided .	follow dobivene instructions .	2,67	2,67
27	choose one of the following file formats : BMP , GIF , JPEG , PCX , PNG , EPS / PS , or TIFF .	odaberite one of the following formata datoteka : BMP , GIF , JPEG , PCX , PNG , EPS / PS or TIFF .	2,67	2,00
28	Page thumbnails are miniature	sličice pages are minijaturni prikazi	1,00	1,00

	previews of the pages in a document	pages dokumenta .		
29	you can use page thumbnails to jump quickly to a selected page or to adjust the view of the page .	sličice pages in you can use for quickly jump to odabranu page or prilagodbu prikaza pages .	1,67	1,67
30	in adobe Reader ® , when you move a page thumbnail , you move the corresponding page .	when in program Adobe Reader ® premještate sličicu pages , zapravo premještate odgovarajuću page .	1,00	1,00
31	in acrobat , when you move , copy , or delete a page thumbnail , you move , copy , or delete the corresponding page .	when in program Acrobat premještate , kopirate sličice pages , or brišete zapravo premještate , kopirate or brišete odgovarajuću page .	1,67	1,00
32	Page thumbnails appear in the navigation pane .	in navigacijskom will be oknu pojaviti sličice pages .	1,00	1,00
33	define the tabbing order	definiranje redoslijeda kretanja	1,00	1,00
34	in the Pages panel , you can set the order in which a user tabs through form fields , links , and comments for each page .	on ploči Stranice you can for each page zadati redoslijed which will be korisnik pritiskom Tab to move between tipku veza obrasca fields , and .	1,00	1,00
35	select a page thumbnail , and choose Page properties from the options menu .	odaberite sličicu pages , and then with izbornika Opcije odaberite Svojstva pages .	1,00	1,00
36	moves in the order specified by the authoring application .	moves redoslijedom definiranim in autorskoj aplikaciji .	1,00	1,00
37	if the document was created in an earlier version of acrobat , the tab order is Unspecified by default .	if the document made in starijoj verziji Acrobata , zadana value is redoslijeda kretanja Nije određen .	1,00	1,00
38	about bookmarks	about knjižnim oznakama	1,00	1,67
39	each bookmark goes to a different view or page in the document .	each knjižna oznaka pridružuje se različitom prikazu or page dokumenta .	1,00	1,00
40	in acrobat , you can set bookmark destinations as you create each bookmark .	stvaranja each time knjižne marks in program Acrobat you can set a certain odredišta for knjižne marks .	1,67	1,00
41	Bookmarks can also perform actions , such as executing a menu item or submitting a form .	knjižne marks can and izvršavati action , like stavki izvršavanja izbornika or slanja obrasca .	1,00	1,00
42	Bookmarks act as a table of contents for some PDFs .	knjižne se marks for some PDF-ove ponašaju as pregledi sadržaja .	1,00	1,00
43	open the page where you want the bookmark to link to , and adjust the view settings .	open the page that you want to connect knjižnu tag and prilagodite postavke prikaza .	1,67	1,67
44	if you don ’ t select a bookmark , the new bookmark is automatically added at the end of the list	if not odaberete knjižnu tag , new knjižna oznaka automatically will be add to the end of the popisa .	1,67	2,00
45	type or edit the name of the new bookmark .	upišite or uredite the name new knjižne marks .	1,00	1,00
46	in Reader , you can make bookmarks easier to read by changing their text appearance .	in the program Reader knjižne marks you can do čitljivijima if change appearance teksta .	1,67	1,00
47	wrap text in a long bookmark	prelamanje dugoj knjižnoj oznaci	1,00	1,00

		teksta in		
48	click the Bookmarks button , and choose Wrap Long Bookmarks from the options menu .	kliknite button Opcije Knjižne marks and with izbornika Prelomi long odaberite knjižne marks .	1,00	1,00
49	you can change the appearance of a bookmark to draw attention to it .	you can change appearance knjižne marks to is istaknuli .	1,67	1,67
50	in the Bookmarks panel , select one or more bookmarks .	on ploči Knjižne marks odaberite one or more knjižnih oznaka .	1,00	1,00
51	in the document pane , move to the location you want to specify as the new destination .	in oknu dokumenta odredište that you want to move to navesti as new odredište .	1,00	1,00
52	in the Bookmark properties dialog box , click Actions .	in dijaloškom okviru Svojstva knjižne marks kliknite Akcije .	1,00	1,00
53	Deleting a bookmark deletes any bookmarks that are subordinate to it .	brisanje izbrisat knjižne marks will and all podređene knjižne marks .	1,00	1,00
54	you can nest a list of bookmarks to show a relationship between topics .	you can knjižnih list oznaka ugnijezditi to ilustrirali odnose between tema .	1,00	1,00
55	Nesting creates a parent / child relationship .	Gniježđenje stvara relationship nadređeno / podređeno .	1,00	1,00
56	Nesting a bookmark (left) , and the result (right)	Gniježđenje knjižne marks (left) and result (right)	1,67	1,00
57	move bookmarks out of a nested position	premještanje knjižnih ugniježđenog position oznaka from	1,00	1,00
58	from the options menu , choose expand Top-Level Bookmarks or Collapse Top-Level Bookmarks .	with izbornika opcija odaberite Proširi knjižne marks most knjižne marks most razine or Sažmi razine .	1,00	1,00
59	tagged bookmarks give you greater control over page content than do regular bookmarks .	knjižne marks you strukturirane pružaju veću control uobičajenih sadržajem pages of knjižnih oznaka .	1,00	1,00
60	converted web pages typically include tagged bookmarks .	pretvorene web-stranice usually sadrže označene knjižne marks .	1,00	1,00
61	select the structure elements you want specified as tagged bookmarks .	odaberite elemente strukture that you want to navesti as strukturirane knjižne marks .	1,00	1,00
62	Edit tags with the tags tab	Uređivanje strukturnih oznaka with the cards Strukturne marks	1,00	1,00
63	add multimedia to PDFs	dodavanje PDF-ove multimedije in	1,33	1,00
64	drag a rectangle where you want to create a link .	opišite pravokutnik on place on which you want to create a connection .	1,00	1,00
65	select the destination file and click select .	odaberite odredišnu datoteku and kliknite Odaberi .	1,00	1,00
66	select the options you want in the create link dialog box .	odaberite željene opcije in dijaloškom okviru Stvaranje connection .	1,00	1,00
67	changing the properties of an existing link affects only the currently selected link .	connection mijenjanje svojstava postojeće trenutačno utječe only on odabranu connection .	1,33	1,00
68	select the link tool and double-click the link rectangle .	odaberite Alat for connection and dvokliknite pravokutnik connection .	1,33	1,00

69	select the Locked option if you want to prevent users from accidentally changing your settings .	odaberite opciju Zaključano if you want to prevent slučajno mijenjanje postavki .	1,33	1,00
70	you can attach PDFs and other types of files to a PDF .	you can prilagati PDF-u PDF-ove and other kinds of datoteka .	1,00	1,00
71	in the Add files dialog box , select the file you want to attach , and click Open .	in dijaloškom okviru Dodaj files odaberite datoteku which you want to priložiti and kliknite Open .	1,00	1,00
72	in the attachments panel , select an attachment , and then choose Delete attachment from the options menu .	on ploči Privici odaberite privitak and with izbornika opcija odaberite Izbriši privitak .	1,00	1,00
73	click Use advanced Search options at the bottom of the window , and then select include attachments .	at the bottom of the window Koristi kliknite napredne opcije pretraživanja , and then odaberite Uključi privitke .	1,33	1,67
74	actions are set in the properties dialog box .	action se postavljaju in dijaloškom okviru Svojstva .	1,00	1,00
75	add actions with page thumbnails	dodavanje action with sličicama pages	1,00	1,00
76	to enhance the interactive quality of a document , you can specify actions , such as changing the zoom value , to occur when a page is opened or closed .	to poboljšali interaktivnost dokumenta , you can definirati action , like stupnja mijenjanja zumiranja , that pokreću when pages open or it .	1,00	1,00
77	Executes a specified menu command as the action .	navedenu command izvršava izbornika as action .	1,00	1,00
78	plays the specified sound file .	reproducira navedenu zvučnu datoteku .	1,00	1,00
79	plays a specified movie that was created as acrobat 6-compatible .	reproducira film kompatibilan with Acrobatom 6 .	1,00	1,00
80	before you add this action , specify the appropriate layer settings .	before dodavanja this action navedite appropriate postavke sloja .	1,00	1,00
81	Toggles between showing and hiding a field in a PDF document .	Prebacuje se between pokazivanja and sakrivanja fields in PDF dokumentu .	1,00	1,00
82	Triggers determine how actions are activated in Media clips , pages , and form fields .	Okidači određuju way pokretanja action in medijskim isječcima te on obrazaca stranicama and in fields .	1,00	1,00
83	when the page containing the Media clip becomes the current page .	when pages with medijskim isječkom becomes trenutna pages .	1,67	1,00
84	you can also use JavaScript with PDF forms and batch sequences .	with PDF obrascima and skupnim sljedovima you can use and JavaScript .	1,67	1,67
85	tagged web bookmarks are initially all at the same level , but you can rearrange them and nest them in family groups to help keep track of the hierarchy of material on the web pages .	strukturirane knjižne marks for web are in početku on the same razinama , but you can premještati and ugnijezditi in skupine to lakše pratili hijerarhiju material on web-stranicama .	1,00	1,00
86	you can display a dialog box with the current page ' s URL , title , date and time downloaded , and other information .	you can uključiti prikaz dijaloškog okvira with URL-om trenutne pages , and vremenom naslovom , datumom preuzimanja te other informacijama .	1,00	1,00
87	the browser opens in a new	preglednik will be open in a new	1,67	1,67

	application window to the page you specify .	navedenoj aplikacijskom window on the page .		
88	drag a rectangle to define the first article box .	opišite pravokutnik to define the first frame članka .	1,67	2,00
89	Use the Article tool to create, display, and make changes to an article box in the PDF document.	for creating , and mijenjanje pregledavanje okvira articles in PDF dokumentu use Alat for articles .	1,67	2,00
90	when editing a batch sequence , click output options .	time skupnog uređivanja slijeda kliknite on Opcije izlaza .	1,00	1,00
91	optimizing : fast web View option	Optimizacija : opcija brzog prikaza for web	1,00	1,00
92	Downsample	piksela smanjivanje number	1,00	1,00
93	Reduces file size by eliminating unnecessary pixel data .	smanjuje datoteku uklanjanjem nepotrebnih piksela .	1,00	1,00
94	Disables all actions related to submitting or importing form data , and resets form fields .	onemogućuje all action povezane with slanjem or uvozom podataka from obrazaca postavlja fields obrazaca and again .	1,00	1,00
95	form data is merged with the page to become page content .	the data obrazaca stapaju with stranicom and postaju part njezina sadržaja .	1,00	1,00
96	removes all versions of an image except the one destined for on-screen viewing .	uklanja all verzije slike except those namijenjene prikazivanju on zaslonu .	1,00	1,00
97	embedded thumbnails , deleting	ugrađene brisanje sličice ,	1,00	1,00
98	fragmented images , merging	fragmentirane spajanje slike ,	1,00	1,00
99	removes embedded search indexes , which reduces file size .	uklanja ugrađena kazala for pretraživanje te smanjuje datoteku .	1,00	1,00
100	removes all bookmarks from the document .	uklanja all knjižne marks from dokumenta .	1,67	1,00
		sustav 4		
Rbr.	**referentan prijevod**	**strojni prijevod**	**A**	**F**
1	Command + Q	Command + Q	4,00	4,00
2	Keyboard shortcuts	tipkovni prečaci	1,00	1,00
3	this section lists common shortcuts for moving around a document .	in this are section popisani najčešći prečaci for kretanje by dokumentu .	1,67	1,67
4	marquee Zoom tool	Označivač for zumiranje	1,00	1,00
5	select Object tool	Alat to select objekata	1,33	1,00
6	crop tool	Alat for obrezivanje	1,00	1,00
7	cycle through drawing markup tools : arrow , Line , Rectangle , Oval , Polygon Line , Polygon , pencil tool , Eraser tool	Kruženje tools for označavanje crteža : Strelica , linija , pravokutnik , Elipsa , linija poligona , Poligon , Olovka , Gumica	1,33	1,33
8	cycle through attach tools : attach file , Record Audio Comment	Kružno kretanje through alate for prilaganje : Priloži datoteku , Alat for snimanje zvučnih komentara	1,00	1,00
9	keys for navigating a PDF	tipke for navigaciju within PDF-a	1,67	2,00
10	move focus to menus (Windows , UNIX) ; expand first menu item (UNIX)	premještanje žarišta on izbornike (Windows , UNIX) ; proširenje first stavke izbornika (UNIX)	1,00	1,00

11	move focus to toolbar in browser	premještanje žarišta on alatnu traku in pregledniku	1,00	1,00
12	move focus to next tab in a tabbed dialog box	premještanje žarišta on sljedeću karticu in kartičnom dijaloškom okviru	1,00	1,00
13	move to next search result and highlight it in the document	premještanje on next result pretraživanja and it isticanje in dokumentu	1,00	1,00
14	keys for working with navigation panels	tipke for working with pločama for navigaciju	1,00	1,00
15	move among the elements of the active navigation panel	kretanje among elementima aktivne ploče for navigaciju	1,33	1,00
16	up arrow or Down arrow	Strelica up or Strelica down	1,67	2,00
17	move focus to previous item in a navigation panel	premještanje žarišta on prethodnu stavku on navigacijskoj ploči	1,00	1,00
18	move to previous pane	premještanje on prethodno okno	1,00	1,00
19	Reflow a tagged PDF , and return to unreflowed view	change prijeloma PDF-a with strukturnim oznakama and vraćanje on prikaz without changes prijeloma	1,00	1,00
20	activate and deactivate Read Out Loud	aktiviranje and deaktiviranje Reading aloud	1,67	2,00
21	after you create page thumbnails , you can embed them in the PDF .	after stvaranja , sličice pages you can ugraditi in PDF .	1,00	1,00
22	Embedding prevents the page thumbnails from redrawing each time you click the Pages button , often a time-consuming process .	ugradnja sprječava again iscrtavanje sličica pages every time kliknete button Stranice because that can zahtijevati much time .	1,00	1,00
23	in the Pages panel , choose Embed All Page Thumbnails or Remove Embedded Page Thumbnails from the options menu .	on ploči Stranice with izbornika Opcije odaberite Ugradi all sličice pages or Ukloni ugrađene sličice pages .	1,00	1,00
24	Embed or unembed page thumbnails in a PDF Portfolio	ugradnja or removing sličica pages from PDF portfelja	1,00	1,00
25	to embed page thumbnails , click Embed Page Thumbnails , and then click Run Sequence .	to ugradili sličice pages , kliknite Ugradi sličice pages , and then Pokreni slijed .	1,00	1,00
26	follow the instructions provided .	follow dobivene instructions .	2,33	2,33
27	choose one of the following file formats : BMP , GIF , JPEG , PCX , PNG , EPS / PS , or TIFF .	odaberite one of the following formata datoteka : BMP , GIF , JPEG , PCX , PNG , EPS / PS or TIFF .	2,67	2,00
28	Page thumbnails are miniature previews of the pages in a document	sličice pages are minijaturni prikazi pages dokumenta .	1,00	1,00
29	you can use page thumbnails to jump quickly to a selected page or to adjust the view of the page .	sličice pages in you can use for quickly jump on odabranu page or prilagodbu prikaza pages .	1,67	1,67
30	in adobe Reader ® , when you move a page thumbnail , you move the corresponding page .	when in program Adobe Reader ® premještate sličicu pages , zapravo premještate odgovarajuću page .	1,67	1,00
31	in acrobat , when you move , copy , or delete a page thumbnail , you move , copy , or delete the	when in program Acrobat premještate , kopirate or brišete sličice pages , zapravo premještate , kopirate or	1,33	1,00

	corresponding page .	brišete odgovarajuću page .		
32	Page thumbnails appear in the navigation pane .	in navigacijskom will oknu pojaviti sličice pages .	1,00	1,00
33	define the tabbing order	definiranje redoslijeda kretanja	1,00	1,00
34	in the Pages panel , you can set the order in which a user tabs through form fields , links , and comments for each page .	on ploči Stranice you can for every page zadati redoslijed which will korisnik pritiskom on tipku Tab move between fields obrasca , veza and .	1,33	1,00
35	select a page thumbnail , and choose Page properties from the options menu .	odaberite sličicu pages , and then with izbornika Opcije odaberite Svojstva pages .	1,00	1,00
36	moves in the order specified by the authoring application .	moves redoslijedom definiranim in autorskoj aplikaciji .	1,00	1,00
37	if the document was created in an earlier version of acrobat , the tab order is Unspecified by default .	if is document made in starijoj verziji Acrobata , zadana value redoslijeda kretanja is Nije određen .	1,00	1,00
38	about bookmarks	about knjižnim oznakama	1,00	1,33
39	each bookmark goes to a different view or page in the document .	each knjižna oznaka pridružuje be različitom prikazu or page dokumenta .	1,00	1,00
40	in acrobat , you can set bookmark destinations as you create each bookmark .	time stvaranja every knjižne marks in program Acrobat you can set a certain odredišta for knjižne marks .	1,00	1,00
41	Bookmarks can also perform actions , such as executing a menu item or submitting a form .	knjižne marks can and izvršavati action , like izvršavanja stavki izbornika or slanja obrasca .	1,00	1,00
42	Bookmarks act as a table of contents for some PDFs .	knjižne be marks for some PDF-ove ponašaju as pregledi sadržaja .	1,00	1,00
43	open the page where you want the bookmark to link to , and adjust the view settings .	open page with want to connect knjižnu tag and prilagodite postavke prikaza .	1,67	1,00
44	if you don ' t select a bookmark , the new bookmark is automatically added at the end of the list	if not odaberete knjižnu tag , new knjižna oznaka automatically will add on end of popisa .	1,00	1,00
45	type or edit the name of the new bookmark .	upišite or uredite the name new knjižne marks .	1,00	1,00
46	in Reader , you can make bookmarks easier to read by changing their text appearance .	in program Reader knjižne marks you can do čitljivijima if change appearance teksta .	1,67	1,67
47	wrap text in a long bookmark	prelamanje teksta in dugoj knjižnoj oznaci	1,00	1,00
48	click the Bookmarks button , and choose Wrap Long Bookmarks from the options menu .	kliknite button Knjižne marks and with izbornika Opcije odaberite Prelomi long knjižne marks .	1,00	1,00
49	you can change the appearance of a bookmark to draw attention to it .	you can change appearance knjižne marks to is istaknuli .	1,67	1,67
50	in the Bookmarks panel , select one or more bookmarks .	on ploči Knjižne marks odaberite one or more knjižnih oznaka .	1,00	1,00
51	in the document pane , move to the location you want to specify as the new destination .	in oknu dokumenta move on odredište that want navesti as new odredište .	1,00	1,00

52	in the Bookmark properties dialog box , click Actions .	in dijaloškom okviru Svojstva knjižne marks kliknite Akcije .	1,00	1,00
53	Deleting a bookmark deletes any bookmarks that are subordinate to it .	brisanje knjižne marks izbrisat will and all podređene knjižne marks .	1,00	1,00
54	you can nest a list of bookmarks to show a relationship between topics .	list knjižnih oznaka you can ugnijezditi to ilustrirali odnose between tema .	1,00	1,00
55	Nesting creates a parent / child relationship .	Gniježđenje stvara relationship nadređeno / podređeno .	1,00	1,00
56	Nesting a bookmark (left) , and the result (right)	Gniježđenje knjižne marks (left) and result (right)	1,33	1,33
57	move bookmarks out of a nested position	premještanje knjižnih oznaka from ugniježđenog position	1,00	1,00
58	from the options menu , choose expand Top-Level Bookmarks or Collapse Top-Level Bookmarks .	with izbornika opcija odaberite Proširi knjižne marks most razine or Sažmi knjižne marks most razine .	1,00	1,00
59	tagged bookmarks give you greater control over page content than do regular bookmarks .	strukturirane knjižne marks pružaju you veću control sadržajem pages of uobičajenih knjižnih oznaka .	1,00	1,00
60	converted web pages typically include tagged bookmarks .	pretvorene web-stranice usually sadrže označene knjižne marks .	1,00	1,00
61	select the structure elements you want specified as tagged bookmarks .	odaberite elemente strukture that want navesti as strukturirane knjižne marks .	1,00	1,00
62	Edit tags with the tags tab	Uređivanje strukturnih oznaka by the cards Strukturne marks	1,00	1,00
63	add multimedia to PDFs	dodavanje multimedije in PDF-ove	1,00	1,00
64	drag a rectangle where you want to create a link .	opišite pravokutnik on place on which want to create a connection .	1,67	1,67
65	select the destination file and click select .	odaberite odredišnu datoteku and kliknite Odaberi .	1,00	1,00
66	select the options you want in the create link dialog box .	odaberite željene opcije in dijaloškom okviru Stvaranje connection .	1,00	1,00
67	changing the properties of an existing link affects only the currently selected link .	mijenjanje svojstava postojeće connection utječe only on trenutačno odabranu connection .	1,00	1,00
68	select the link tool and double-click the link rectangle .	odaberite Alat for connection and dvokliknite pravokutnik connection .	1,33	1,00
69	select the Locked option if you want to prevent users from accidentally changing your settings .	odaberite opciju Zaključano if you want to prevent slučajno mijenjanje postavki .	1,33	1,33
70	you can attach PDFs and other types of files to a PDF .	PDF-u you can prilagati PDF-ove and other kinds of datoteka .	1,67	1,67
71	in the Add files dialog box , select the file you want to attach , and click Open .	in dijaloškom okviru Dodaj files odaberite datoteku which want priložiti and kliknite Open .	1,00	1,00
72	in the attachments panel , select an attachment , and then choose Delete attachment from the options menu .	on ploči Privici odaberite privitak and with izbornika opcija odaberite Izbriši privitak .	1,00	1,00
73	click Use advanced Search options	on bottom window kliknite Koristi	1,33	1,00

	at the bottom of the window , and then select include attachments .	napredne opcije pretraživanja , and then odaberite Uključi privitke .		
74	actions are set in the properties dialog box .	action be postavljaju in dijaloškom okviru Svojstva .	1,00	1,00
75	add actions with page thumbnails	dodavanje action with sličicama pages	1,00	1,00
76	to enhance the interactive quality of a document , you can specify actions , such as changing the zoom value , to occur when a page is opened or closed .	to poboljšali interaktivnost dokumenta , you can definirati action , like mijenjanja stupnja zumiranja , that pokreću when pages opens or it .	1,33	1,33
77	Executes a specified menu command as the action .	izvršava navedenu command izbornika as action .	1,00	1,00
78	plays the specified sound file .	reproducira navedenu zvučnu datoteku .	1,00	1,00
79	plays a specified movie that was created as acrobat 6-compatible .	reproducira film kompatibilan with Acrobatom 6 .	1,67	1,67
80	before you add this action , specify the appropriate layer settings .	before dodavanja these action navedite appropriate postavke sloja .	1,00	1,00
81	Toggles between showing and hiding a field in a PDF document .	Prebacuje be between pokazivanja and sakrivanja fields in PDF dokumentu .	1,00	1,00
82	Triggers determine how actions are activated in Media clips , pages , and form fields .	Okidači određuju way pokretanja action in medijskim isječcima those on stranicama and in fields obrazaca .	1,00	1,00
83	when the page containing the Media clip becomes the current page .	when pages with medijskim isječkom becomes trenutna pages .	1,33	1,33
84	you can also use JavaScript with PDF forms and batch sequences .	with PDF obrascima and skupnim sljedovima you can use and JavaScript .	1,67	1,67
85	tagged web bookmarks are initially all at the same level , but you can rearrange them and nest them in family groups to help keep track of the hierarchy of material on the web pages .	strukturirane knjižne marks for web are in početku on the same razinama , but them you can premještati and ugnijezditi in skupine to lakše pratili hijerarhiju material on web-stranicama .	1,00	1,00
86	you can display a dialog box with the current page ' s URL , title , date and time downloaded , and other information .	you can uključiti prikaz dijaloškog okvira with URL-om trenutne pages , naslovom , datumom and vremenom preuzimanja those other informacijama .	1,00	1,00
87	the browser opens in a new application window to the page you specify .	preglednik will open in a new aplikacijskom window on navedenoj page .	1,33	1,33
88	drag a rectangle to define the first article box .	opišite pravokutnik to define first frame članka .	1,67	1,67
89	Use the Article tool to create, display, and make changes to an article box in the PDF document.	for creating , pregledavanje and mijenjanje okvira articles in PDF dokumentu use Alat for articles .	1,67	1,33
90	when editing a batch sequence , click output options .	time uređivanja skupnog slijeda kliknite on Opcije izlaza .	1,00	1,00
91	optimizing : fast web View option	Optimizacija : opcija brzog prikaza	1,00	1,00

		for web		
92	Downsample	smanjivanje number piksela	1,00	1,00
93	Reduces file size by eliminating unnecessary pixel data .	smanjuje datoteku uklanjanjem nepotrebnih piksela .	1,00	1,00
94	Disables all actions related to submitting or importing form data , and resets form fields .	onemogućuje all action povezane with slanjem or uvozom podataka from obrazaca and again postavlja fields obrazaca .	1,00	1,00
95	form data is merged with the page to become page content .	the data obrazaca stapaju with stranicom and postaju part njezina sadržaja .	1,00	1,00
96	removes all versions of an image except the one destined for on-screen viewing .	uklanja all verzije slike except those namijenjene prikazivanju on zaslonu .	1,33	1,00
97	embedded thumbnails , deleting	ugrađene sličice , brisanje	1,00	1,00
98	fragmented images , merging	fragmentirane slike , spajanje	1,00	1,00
99	removes embedded search indexes , which reduces file size .	uklanja ugrađena kazala for pretraživanje those smanjuje datoteku .	1,00	1,00
100	removes all bookmarks from the document .	uklanja all knjižne marks from dokumenta .	1,33	1,00
		sustav 5		
Rbr.	**referentan prijevod**	**strojni prijevod**	**A**	**F**
1	Command + Q	zapovijedam + dr kao	1,00	1,00
2	tipkovni prečaci	Keyboard prečice	2,33	1,33
3	u ovom su odjeljku popisani najčešći prečaci za kretanje po dokumentu .	ovo odjeljak poginulih zdravog prečice za kreće je dokument .	1,00	1,00
4	Označivač za zumiranje	marquee Zoom oruđe	1,67	1,33
5	Alat za odabir objekata	odabir Prigovor oruđe	1,00	1,00
6	Alat za obrezivanje	pripremimo oruđe	1,00	1,67
7	Kruženje alatima za označavanje crteža : Strelica , linija , pravokutnik , Elipsa , linija poligona , Poligon , Olovka , Gumica	motor kroz primaće markup sredstvo : strijelom , Maginot , Rectangle , Oval , Polygon Maginot , Polygon , olovku oruđe , Eraser oruđe	1,67	1,00
8	Kružno kretanje kroz alate za prilaganje : Priloži datoteku , Alat za snimanje zvučnih komentara	motor kroz attach sredstvo : attach podnio , Record Audio vi plesačica	1,00	1,00
9	tipke za navigaciju unutar PDF-a	ključeve za navigating malo PDF	1,67	1,00
10	premještanje žarišta na izbornike (Windows , UNIX) ; proširenje prve stavke izbornika (UNIX)	pomakni focus da menus (Windows , UNIX) ; expand prvo jelovnik pojedinost Mana Ganda UNIX	1,00	1,00
11	premještanje žarišta na alatnu traku u pregledniku	pomakni focus da toolbar u browser	1,67	1,00
12	premještanje žarišta na sljedeću karticu u kartičnom dijaloškom okviru	pomakni focus da sljedeći tab u tabbed dialog kutiju	1,67	1,00
13	premještanje na sljedeći rezultat pretraživanja i njegovo isticanje u dokumentu	se sljedeći pretražiti rezultat i highlight to u dokument	1,67	1,00
14	tipke za rad s pločama za navigaciju	ključeve za raditi s navigaciji panels	1,67	1,33

15	kretanje među elementima aktivne ploče za navigaciju	pomakni među prirodni od aktivnu navigaciji panel	1,00	1,00
16	Strelica gore ili Strelica dolje	gore strijelom ili onog strijelom	1,00	1,00
17	premještanje žarišta na prethodnu stavku na navigacijskoj ploči	pomakni focus da prijašnje pojedinost u navigaciji panel	1,67	1,33
18	premještanje na prethodno okno	se prijašnje pane	1,00	1,00
19	promjena prijeloma PDF-a sa strukturnim oznakama i vraćanje na prikaz bez promjene prijeloma	Reflow malo oznaиiti PDF , i vratiti u unreflowed pogled	1,00	1,00
20	aktiviranje i deaktiviranje čitanja naglas	activate i deactivate Read Out Loud	1,00	1,00
21	nakon stvaranja , sličice stranica možete ugraditi u PDF .	poslije vas slikati stranici thumbnails , možeš embed ih u PDF .	1,00	1,00
22	ugradnja sprječava ponovno iscrtavanje sličica stranica svaki put kad kliknete gumb Stranice jer to može zahtijevati mnogo vremena .	Embedding da stranicu thumbnails iz redrawing svaki put click na Pages dugme , često je time-consuming proces .	1,00	1,00
23	na ploči Stranice s izbornika Opcije odaberite Ugradi sve sličice stranica ili Ukloni ugrađene sličice stranica .	u Pages panel , biram Embed U Page Thumbnails ili Remove Embedded Page Thumbnails iz obveza jelovnik .	1,00	1,00
24	ugradnja ili uklanjanje sličica stranica iz PDF portfelja	Embed ili unembed stranici thumbnails u PDF Portfolio	1,00	1,00
25	da biste ugradili sličice stranica , kliknite Ugradi sličice stranica , a zatim Pokreni slijed .	da embed stranici thumbnails , click Embed Page Thumbnails , i onda click Runa Sequence .	1,00	1,00
26	slijedite dobivene upute .	slijedi upute pod .	2,00	1,67
27	odaberite jedan od sljedećih formata datoteka : BMP , GIF , JPEG , PCX , PNG , EPS / PS ili TIFF .	Izaberite jedan od slijedeće dosijeima formats : BMP , GIF , JPEG , PCX , PNG , EPS Gelula PS , ili TIFF .	1,67	1,33
28	sličice stranica su minijaturni prikazi stranica dokumenta .	Page thumbnails su miniature previews od stranica u dokument .	1,67	1,00
29	sličice stranica u možete koristiti za brzo skakanje na odabranu stranicu ili prilagodbu prikaza stranice .	moћeљ koristiti stranici thumbnails da skočim brzo do izabrani ste stranici ili da adjust pogled na stranici .	1,67	1,00
30	kada u programu Adobe Reader ® premještate sličicu stranice , zapravo premještate odgovarajuću stranicu .	u adobe Reader ® , kad ti se stranicu thumbnail , ti se corresponding stranici .	1,00	1,00
31	kada u programu Acrobat premještate , kopirate ili brišete sličice stranica , zapravo premještate , kopirate ili brišete odgovarajuću stranicu .	u akrobata , kad ti se , jednu , ili delete stranicu thumbnail , ti se , jednu , ili delete na corresponding stranici .	1,00	1,00
32	u navigacijskom će se oknu pojaviti sličice stranica .	Page thumbnails pojavio u navigaciji pane .	1,00	1,00
33	definiranje redoslijeda kretanja	definirali na tabbing red	1,00	1,00
34	na ploči Stranice možete za svaku stranicu zadati redoslijed kojim će se korisnik pritiskom na tipku Tab kretati između polja obrasca , veza i komentara .	u Pages panel , možeš set zapovijed u kojoj je user tabs kroz oblik polja , links , i što za dodati za jedno izdanje .	1,00	1,00

35	odaberite sličicu stranice , a zatim s izbornika Opcije odaberite Svojstva stranice .	odabir stranicu thumbnail , i biraš Page posjedima iz obveza jelovnik .	1,00	1,00
36	kreće se redoslijedom definiranim u autorskoj aplikaciji .	seli se u red specified po authoring spis .	1,00	1,00
37	ako je dokument stvoren u starijoj verziji Acrobata , zadana vrijednost redoslijeda kretanja jest Nije određen .	ako je dokument je stvorio u ranije verzija od akrobata , tab red je Unspecified po default .	1,00	1,00
38	o knjižnim oznakama	o bookmarks	1,00	1,67
39	svaka knjižna oznaka pridružuje se različitom prikazu ili stranici dokumenta .	svaki bookmark ide u drukčiji pogled ili stranici u dokument .	1,00	1,00
40	prilikom stvaranja svake knjižne oznake u programu Acrobat možete postaviti određene odredišta za knjižne oznake .	u akrobata , možeš set bookmark destinations kao da stvarate jedno bookmark .	1,00	1,00
41	knjižne oznake mogu i izvršavati akcije , poput izvršavanja stavki izbornika ili slanja obrasca .	Bookmarks i izvesti djela , tako kao executing malo jelovnik pojedinost ili submitting je oblik .	1,00	1,00
42	knjižne se oznake za neke PDF-ove ponašaju kao pregledi sadržaja .	Bookmarks kao stol od a za neke PDFs .	1,00	1,00
43	otvorite stranicu s kojom želite povezati knjižnu oznaku i prilagodite postavke prikaza .	otvori stranicu gdje želite bookmark da kariku da , i adjust pogled settings .	1,67	1,33
44	ako ne odaberete knjižnu oznaku , nova knjižna oznaka automatski će se dodati na kraj popisa .	ako ti ne odabir malo bookmark , novi bookmark je automatski dodala na kraju popis .	1,00	1,00
45	upišite ili uredite naziv nove knjižne oznake .	tip ili edit ime novi bookmark .	1,00	1,00
46	u programu Reader knjižne oznake možete učiniti čitljivijima ako promijenite izgled teksta .	u Reader , možeš napraviti bookmarks lakše da čitati zamijenio njihov tekst za izgled .	1,67	1,33
47	prelamanje teksta u dugoj knjižnoj oznaci	zamotaj tekst za u dugo bookmark	1,00	1,00
48	kliknite gumb Knjižne oznake i s izbornika Opcije odaberite Prelomi duge knjižne oznake .	click na Bookmarks dugme , i biraš Wrap Long Bookmarks iz obveza jelovnik .	1,00	1,00
49	možete promijeniti izgled knjižne oznake da biste je istaknuli .	možeš promijeniti izgled od bookmark da potegne pozornost .	1,67	1,33
50	na ploči Knjižne oznake odaberite jednu ili više knjižnih oznaka .	u Bookmarks panel , odabir jedan ili više bookmarks .	1,00	1,00
51	u oknu dokumenta pomaknite se na odredište koje želite navesti kao novo odredište .	u dokument pane , pomakni na terenu želiš specify kao nova odredi ste .	1,00	1,00
52	u dijaloškom okviru Svojstva knjižne oznake kliknite Akcije .	u Bookmark posjedima dialog kutiju , click Actions .	1,00	1,00
53	brisanje knjižne oznake izbrisat će i sve podređene knjižne oznake .	Deleting malo bookmark deletes bilo bookmarks to su subordinate da to .	1,00	1,00
54	popis knjižnih oznaka možete ugnijezditi da biste ilustrirali odnose	možeš gnijezdu popis od bookmarks pokazati malo odnosa između topics .	1,00	1,00

	između tema .			
55	Gniježđenje stvara odnos nadređeno / podređeno .	Nesting creates roditelj Gelula dijete odnos .	1,00	1,00
56	Gniježđenje knjižne oznake (lijevo) i rezultat (desno)	Nesting malo bookmark (ostavio) , i rezultat (redu !	1,00	1,00
57	premještanje knjižnih oznaka iz ugniježđenog položaja	pomakni bookmarks iz nested položaj	1,33	1,33
58	s izbornika opcija odaberite Proširi knjižne oznake najviše razine ili Sažmi knjižne oznake najviše razine .	iz obveza jelovnik , biram expand Top-Level Bookmarks ili kolaps Top-Level Bookmarks .	1,00	1,00
59	strukturirane knjižne oznake pružaju vam veću kontrolu nad sadržajem stranice od uobičajenih knjižnih oznaka .	oznaиiti bookmarks dati više kontrolirate stranici zadovoljna nego raditi uobičajeni bookmarks .	1,00	1,00
60	pretvorene web-stranice obično sadrže označene knjižne oznake .	Ęucaje web stranica typically Sadrži oznaиiti bookmarks .	1,00	1,00
61	odaberite elemente strukture koje želite navesti kao strukturirane knjižne oznake .	odabir građevina koja elements želiš specified kao oznaиiti bookmarks .	1,00	1,00
62	Uređivanje strukturnih oznaka pomoću kartice Strukturne oznake	Edit privjeske s privjesci tab	1,00	1,00
63	dodavanje multimedije u PDF-ove	dodati multimedia da PDFs	1,33	1,00
64	opišite pravokutnik na mjestu na kojem želite stvoriti vezu .	odvući malo rectangle gdje želiš stvoriti kariku .	1,67	1,00
65	odaberite odredišnu datoteku i kliknite Odaberi .	odabir na odrediљte dosijeima i click odabir .	1,67	1,00
66	odaberite željene opcije u dijaloškom okviru Stvaranje veze .	odabir na obveza želiš u stvaranje kariku dialog kutiju .	1,67	1,00
67	mijenjanje svojstava postojeće veze utječe samo na trenutačno odabranu vezu .	zamijenio je posjedima na postojanje kariku affects samo currently izabrani ste kariku .	1,00	1,00
68	odaberite Alat za veze i dvokliknite pravokutnik veze .	odabir na kariku oruđe i double-click na kariku rectangle .	1,00	1,00
69	odaberite opciju Zaključano ako želite spriječiti slučajno mijenjanje postavki .	odabir na Zaključèano option ako želiš sprečavaju users iz ubijete mijenja vašeg settings .	1,00	1,00
70	PDF-u možete prilagati PDF-ove i druge vrste datoteka .	možeš attach PDFs i drugim vrstama datoteke u PDF .	1,33	1,00
71	u dijaloškom okviru Dodaj datoteke odaberite datoteku koju želite priložiti i kliknite Otvori .	u Add datoteke dialog kutiju , odabir arhive želiš attach , i click Otvori .	1,00	1,00
72	na ploči Privici odaberite privitak i s izbornika opcija odaberite Izbriši privitak .	u attachments panel , odabir sat vezanost , i onda izabrati Delete vezanost iz obveza jelovnik .	1,00	1,00
73	na dnu prozora kliknite Koristi napredne opcije pretraživanja , a zatim odaberite Uključi privitke .	click Upotrijebite vidjeh Search obveza na dnu prozor , i onda odabir Sadrži attachments .	1,00	1,00
74	akcije se postavljaju u dijaloškom okviru Svojstva .	djela su set u posjedima dialog kutiju .	1,00	1,00
75	dodavanje akcija sa sličicama stranica	dodati djela s stranici thumbnails	1,00	1,00

76	da biste poboljšali interaktivnost dokumenta , možete definirati akcije , poput mijenjanja stupnja zumiranja , koje se pokreću kada se stranica otvori ili zatvori .	da enhance na interactive kvaliteta od dokument , možeš specify djela , tako kao mijenja na zoom vrijednost , da pomislili kad stranicu je otvorio ili zatvorena .	1,00	1,00
77	izvršava navedenu naredbu izbornika kao akciju .	Executes malo specified jelovnik zapovijedaš kao akciju .	1,00	1,00
78	reproducira navedenu zvučnu datoteku .	svira na specified zvuk dosijeima .	1,00	1,00
79	reproducira film kompatibilan s Acrobatom 6 .	uvijek svira specified filmske to je stvorio kao akrobata 6-compatible .	1,00	1,00
80	prije dodavanja ove akcije navedite odgovarajuće postavke sloja .	prije si dodati ovu akciju , specify je prikladno sloj settings .	1,00	1,00
81	Prebacuje se između pokazivanja i sakrivanja polja u PDF dokumentu .	Toggles između pokažem i skriva polju u PDF dokument .	1,00	1,00
82	Okidači određuju način pokretanja akcija u medijskim isječcima te na stranicama i u poljima obrazaca .	Triggers ustanovi kako djela su potpuna u Media clips , stranica , i oblik poljima .	1,00	1,00
83	kada stranica s medijskim isječkom postaje trenutna stranica .	kad stranicu containing na Media režem postaje struju stranici .	1,00	1,00
84	s PDF obrascima i skupnim sljedovima možete koristiti i JavaScript .	možeš i koristi JavaScript s PDF niste i pošiljci sequences .	1,00	1,00
85	strukturirane knjižne oznake za web nalaze se u početku na istim razinama , ali ih možete premještati i ugnijezditi u skupine da biste lakše pratili hijerarhiju materijala na web-stranicama .	oznaиiti web bookmarks su initially sve u isto , ali moћeљ rearrange ih i gnijezdu ih u obitelji grupe pomoći pratiti na hierarchy od materijal na netu stranica .	1,00	1,00
86	možete uključiti prikaz dijaloškog okvira s URL-om trenutne stranice , naslovom , datumom i vremenom preuzimanja te drugim informacijama .	možeš ispadne malo dialog kutiju s struju stranici je URL , naslov , sastanak i vrijeme downloaded , i druge informacije .	1,67	1,00
87	preglednik će se otvoriti u novom aplikacijskom prozoru na navedenoj stranici .	na browser uđe u novi spis prozor da stranicu ti specify .	1,00	1,00
88	opišite pravokutnik da biste definirali prvi okvir članka .	odvući malo rectangle da definirali prvi član kutiju .	1,00	1,00
89	Za stvaranje, pregledavanje i mijenjanje okvira članaka u PDF dokumentu koristite Alat za članke.	upotrebi Article oruđe stvoriti , ispadne , i mijenja da članak kutiju u PDF dokument .	1,00	1,00
90	prilikom uređivanja skupnog slijeda kliknite na Opcije izlaza .	kad editing malo pošiljci sequence , click output obveza .	1,00	1,00
91	Optimizacija : opcija brzog prikaza za web	optimizing : brzo web View option	1,00	1,00
92	smanjivanje broja piksela	Downsample	1,00	1,00
93	smanjuje datoteku uklanjanjem nepotrebnih piksela .	Reduces dosje broj do eliminating nepotrebno pixel data .	1,00	1,00
94	onemogućuje sve akcije povezane sa slanjem ili uvozom podataka iz	Disables sve djela u da submitting ili importing oblik data , i resets oblik	1,00	1,00

	obrazaca i ponovo postavlja polja obrazaca .	poljima .		
95	podaci obrazaca stapaju se sa stranicom i postaju dio njezina sadržaja .	oblik data je novina je preuzeto sa stranicu postati stranici zadovoljna .	1,00	1,00
96	uklanja sve verzije slike osim one namijenjene prikazivanju na zaslonu .	ukÉď ni sve versions na Predstavljaću osim je netko tko je unaprijed izabran za on-screen viewing .	1,00	1,00
97	ugrađene sličice , brisanje	embedded thumbnails , deleting	1,00	1,00
98	fragmentirane slike , spajanje	fragmented images , merging	1,00	1,00
99	uklanja ugrađena kazala za pretraživanje te smanjuje datoteku .	ukÉď ni embedded pretražiti indexes , koji reduces dosje broj .	1,00	1,00
100	uklanja sve knjižne oznake iz dokumenta .	ukÉď ni sve bookmarks iz dokument .	1,00	1,00
		sustav 6		
Rbr.	**referentan prijevod**	**strojni prijevod**	**A**	**F**
1	Command + Q	Command + Q	4,00	4,00
2	tipkovni prečaci	tipkovni prečaci	4,00	4,00
3	u ovom su odjeljku popisani najčešći prečaci za kretanje po dokumentu .	ovaj odjeljak prikazuje najčešće prečace za kretanje po dokumenta .	4,00	4,00
4	Označivač za zumiranje	Označivač za zumiranje	4,00	4,00
5	Alat za odabir objekata	odabir objekata	3,00	4,00
6	Alat za obrezivanje	za obrezivanje	2,67	4,00
7	Kruženje alatima za označavanje crteža : Strelica , linija , pravokutnik , Elipsa , linija poligona , Poligon , Olovka , Gumica	listanje : označavanje crtanje Cycle strelicu , crte , pravokutnike , ovale , segmentirane crte za segmentirane , Gumica za Eraser	3,00	2,67
8	Kružno kretanje kroz alate za prilaganje : Priloži datoteku , Alat za snimanje zvučnih komentara	Cycle kroz priložiti Alati : Priloži datoteku snimanje zvučnih komentara	1,67	1,67
9	tipke za navigaciju unutar PDF-a	ključeva za Navigacija PDF	1,67	1,67
10	premještanje žarišta na izbornike (Windows , UNIX) ; proširenje prve stavke izbornika (UNIX)	premještanje žarišta na izbornicima (Windows , UNIX) ; proširite prvi izbornik stavka (UNIX)	3,00	2,67
11	premještanje žarišta na alatnu traku u pregledniku	premještanje žarišta na alatna traka u pregledniku	4,00	3,00
12	premještanje žarišta na sljedeću karticu u kartičnom dijaloškom okviru	premještanje žarišta na sljedeći karticu u dijaloškom okviru tabbed	2,67	3,00
13	premještanje na sljedeći rezultat pretraživanja i njegovo isticanje u dokumentu	sljedeći prešli na rezultat pretraživanja i ga istaknuli unutar dokumenta	2,67	2,00
14	tipke za rad s pločama za navigaciju	strelicama za rad s navigacijskih ploča	2,67	1,67
15	kretanje među elementima aktivne ploče za navigaciju	pomicanje između elemente aktivnog Navigacijska ploča	3,00	2,00
16	Strelica gore ili Strelica dolje	gore lijevo ili lijevo	1,67	3,00
17	premještanje žarišta na prethodnu stavku na navigacijskoj ploči	premještanje žarišta na prijašnji stavka u navigacijskoj ploči	3,00	3,00
18	premještanje na prethodno okno	premještanje bivšem okno	2,67	2,00
19	promjena prijeloma PDF-a sa	promjena prijeloma strukturiranog	4,00	4,00

	strukturnim oznakama i vraćanje na prikaz bez promjene prijeloma	PDF-a i povratak na prikaz bez promjene prijeloma		
20	aktiviranje i deaktiviranje čitanja naglas	aktiviranje ili deaktiviranje čitanja naglas	4,00	4,00
21	nakon stvaranja , sličice stranica možete ugraditi u PDF .	nakon što stvorite sličice stranica možete ugrađivati ih u PDF .	4,00	3,00
22	ugradnja sprječava ponovno iscrtavanje sličica stranica svaki put kad kliknete gumb Stranice jer to može zahtijevati mnogo vremena .	ugrađivanje sprječava sličice stranica iz ponovno iscrtavanje svaki put kada kliknete gumb Stranice , često dugotrajni procesa .	3,00	2,00
23	na ploči Stranice s izbornika Opcije odaberite Ugradi sve sličice stranica ili Ukloni ugrađene sličice stranica .	na ploči Stranice odaberite Ugradi sve sličice stranica ili Ukloni ugrađenim sličice stranica u izborniku opcija .	2,67	2,33
24	ugradnja ili uklanjanje sličica stranica iz PDF portfelja	ugrađivanje ili unembed sličice stranica u PDF portfelju	1,67	1,67
25	da biste ugradili sličice stranica , kliknite Ugradi sličice stranica , a zatim Pokreni slijed .	da biste ugradili sličice stranica , kliknite Ugradi sličice stranica , a zatim kliknite Pokreni slijed .	4,00	4,00
26	slijedite dobivene upute .	slijedite upute koje .	2,67	3,33
27	odaberite jedan od sljedećih formata datoteka : BMP , GIF , JPEG , PCX , PNG , EPS / PS ili TIFF .	odaberite jednu od sljedećih formata datoteka : BMP , GIF , JPEG , PCX , PNG , EPS / PS , ili TIFF .	4,00	3,33
28	sličice stranica su minijaturni prikazi stranica dokumenta .	sličice stranica su minijaturne previews stranice u dokumentu .	2,33	3,00
29	sličice stranica u možete koristiti za brzo skakanje na odabranu stranicu ili prilagodbu prikaza stranice .	možete koristiti sličice stranica za skok brzo na odabrane stranice ili za prilagodbu prikaz stranice .	2,67	2,33
30	kada u programu Adobe Reader ® premještate sličicu stranice , zapravo premještate odgovarajuću stranicu .	u programu Adobe Reader ® kada pomaknete sličicu stranice , pomičete odgovarajuće stranice .	4,00	3,67
31	kada u programu Acrobat premještate , kopirate ili brišete sličice stranica , zapravo premještate , kopirate ili brišete odgovarajuću stranicu .	u Acrobatu , kada pomaknete , kopiranje ili izbrisala sličicu stranice , možete premještati , kopirati ili izbrisala odgovarajuće stranice .	3,00	2,00
32	u navigacijskom će se oknu pojaviti sličice stranica .	sličice stranica pojavile u navigacijskom oknu .	3,00	2,67
33	definiranje redoslijeda kretanja	određivanje redoslijeda kretanja	4,00	4,00
34	na ploči Stranice možete za svaku stranicu zadati redoslijed kojim će se korisnik pritiskom na tipku Tab kretati između polja obrasca , veza i komentara .	na ploči Stranice možete postaviti redoslijedom kojim se korisnik tabulatorom kroz polja obrasca , veze i komentari za svaku stranicu .	3,00	2,00
35	odaberite sličicu stranice , a zatim s izbornika Opcije odaberite Svojstva stranice .	odaberite sličicu stranice i odaberite Svojstva stranice u izborniku opcija .	3,00	3,67
36	kreće se redoslijedom definiranim u autorskoj aplikaciji .	premješta redoslijedom odredio autorske aplikacije .	1,67	2,00
37	ako je dokument stvoren u starijoj verziji Acrobata , zadana vrijednost redoslijeda kretanja jest Nije	ako dokument stvoren u nekoj prethodnoj verzije Acrobata , kretanja tabulatorom je Unspecified prema	1,67	1,67

	određen .	zadanom .		
38	o knjižnim oznakama	o knjižne oznake	3,00	3,00
39	svaka knjižna oznaka pridružuje se različitom prikazu ili stranici dokumenta .	svaku knjižnu oznaku goes na drugi prikaz ili stranicu u dokumentu .	2,67	2,00
40	prilikom stvaranja svake knjižne oznake u programu Acrobat možete postaviti određene odredišta za knjižne oznake .	u Acrobatu , možete postaviti oznakom odredišta što stvorite svaku knjižnu oznaku .	2,67	2,00
41	knjižne oznake mogu i izvršavati akcije , poput izvršavanja stavki izbornika ili slanja obrasca .	knjižne oznake možete izvršiti akcije , poput executing stavku izbornika ili predavanja obrasca .	3,00	2,00
42	knjižne se oznake za neke PDF-ove ponašaju kao pregledi sadržaja .	knjižne oznake u pravilu ponašaju kao sadržaj za neke PDF-ova .	2,67	2,00
43	otvorite stranicu s kojom želite povezati knjižnu oznaku i prilagodite postavke prikaza .	otvorite stranici gdje želite da biste povezali knjižnu oznaku te prilagodite prikaz postavki .	2,67	2,33
44	ako ne odaberete knjižnu oznaku , nova knjižna oznaka automatski će se dodati na kraj popisa .	ako ne odaberete knjižnu oznaku , novu knjižnu oznaku se automatski dodaju na kraju popisa .	3,67	3,00
45	upišite ili uredite naziv nove knjižne oznake .	upišite ili uredite naziv novu knjižnu oznaku .	4,00	3,00
46	u programu Reader knjižne oznake možete učiniti čitljivijima ako promijenite izgled teksta .	u Readeru možete učiniti knjižne oznake lakše čitljivim mijenjajući svoje izmjene izgled .	3,67	3,00
47	prelamanje teksta u dugoj knjižnoj oznaci	Prelomi teksta u dugo oznakom	1,67	1,67
48	kliknite gumb Knjižne oznake i s izbornika Opcije odaberite Prelomi duge knjižne oznake .	kliknite gumb Knjižne oznake i odaberite Prelomi dugi knjižne oznake u izborniku opcija .	2,67	3,00
49	možete promijeniti izgled knjižne oznake da biste je istaknuli .	možete promijeniti izgled knjižnu oznaku da biste nacrtali attention .	1,67	1,67
50	na ploči Knjižne oznake odaberite jednu ili više knjižnih oznaka .	na ploči Knjižne oznake odaberite jedan ili više knjižnih oznaka .	4,00	3,33
51	u oknu dokumenta pomaknite se na odredište koje želite navesti kao novo odredište .	u oknu dokumenta , pomaknite kamo želite da biste naveli kao nove .	3,00	2,00
52	u dijaloškom okviru Svojstva knjižne oznake kliknite Akcije .	u dijaloškom okviru Svojstva knjižne oznake kliknite na Akcije .	4,00	4,00
53	brisanje knjižne oznake izbrisat će i sve podređene knjižne oznake .	brisanje knjižnu oznaku briše sve knjižne oznake koje su subordinate .	3,00	2,33
54	popis knjižnih oznaka možete ugnijezditi da biste ilustrirali odnose između tema .	možete nest popis knjižnih oznaka da biste prikazali odnos između teme .	3,00	2,00
55	Gniježđenje stvara odnos nadređeno / podređeno .	stvara gniježđenja nadređenim / podređene odnos .	1,67	1,67
56	Gniježđenje knjižne oznake (lijevo) i rezultat (desno)	knjižne oznake gniježđenja (lijevo) i rezultat (desno)	1,67	2,00
57	premještanje knjižnih oznaka iz ugniježđenog položaja	premještanje knjižne oznake izvan ugniježđenog položaj	3,67	2,67
58	s izbornika opcija odaberite Proširi	na izborniku opcija odaberite Proširi	4,00	4,00

	knjižne oznake najviše razine ili Sažmi knjižne oznake najviše razine .	knjižne oznake najviše razine ili Sažmi knjižne oznake najviše razine .		
59	strukturirane knjižne oznake pružaju vam veću kontrolu nad sadržajem stranice od uobičajenih knjižnih oznaka .	strukturirane knjižne oznake pružaju veću kontrolu nad sadržajem stranice od učinite redovna knjižne oznake .	4,00	3,00
60	pretvorene web-stranice obično sadrže označene knjižne oznake .	pretvorene web-stranice obično obuhvaćaju strukturiranih knjižnih oznaka .	3,67	3,00
61	odaberite elemente strukture koje želite navesti kao strukturirane knjižne oznake .	odaberite strukturu elemente želite naveli kao strukturiranih knjižnih oznaka .	2,67	2,00
62	Uređivanje strukturnih oznaka pomoću kartice Strukturne oznake	Uređivanje strukturnih oznaka pomoću kartice Strukturne oznake	4,00	4,00
63	dodavanje multimedije u PDF-ove	dodavanje multimedije u PDF-ove	4,00	4,00
64	opišite pravokutnik na mjestu na kojem želite stvoriti vezu .	povucite pravokutnik gdje želite stvoriti vezu .	4,00	3,33
65	odaberite odredišnu datoteku i kliknite Odaberi .	odaberite odredište datoteku i kliknite Odaberi .	4,00	3,00
66	odaberite željene opcije u dijaloškom okviru Stvaranje veze .	odaberite željene opcije u dijaloškom okviru Stvaranje veze .	4,00	4,00
67	mijenjanje svojstava postojeće veze utječe samo na trenutačno odabranu vezu .	promjena svojstva postojeći vezu utječe samo trenutno odabrane vezu .	3,00	2,33
68	odaberite Alat za veze i dvokliknite pravokutnik veze .	odaberite Alat za veze i dvokliknuti pravokutnik veze .	3,67	3,33
69	odaberite opciju Zaključano ako želite spriječiti slučajno mijenjanje postavki .	odaberite tu opciju ako želite spriječiti korisnika accidentally promjeni postavki .	2,67	2,33
70	PDF-u možete prilagati PDF-ove i druge vrste datoteka .	možete priložiti PDF-ove i druge vrste datoteka u PDF .	3,67	2,67
71	u dijaloškom okviru Dodaj datoteke odaberite datoteku koju želite priložiti i kliknite Otvori .	u dijaloškom okviru Dodavanje datoteka odaberite datoteku koju želite priložiti i kliknite Otvori .	4,00	4,00
72	na ploči Privici odaberite privitak i s izbornika opcija odaberite Izbriši privitak .	u Privici ploče odaberite privitak , a zatim odaberite Izbriši priloge u izborniku opcija .	2,67	2,67
73	na dnu prozora kliknite Koristi napredne opcije pretraživanja , a zatim odaberite Uključi privitke .	kliknite na Koristi napredne opcije pretraživanja pri dnu prozora i zatim odaberite Obuhvati Privici .	3,67	3,33
74	akcije se postavljaju u dijaloškom okviru Svojstva .	akcije su postavljene u dijaloškom okviru Svojstva .	4,00	4,00
75	dodavanje akcija sa sličicama stranica	dodavanje akcija sa sličicama stranica	4,00	4,00
76	da biste poboljšali interaktivnost dokumenta , možete definirati akcije , poput mijenjanja stupnja zumiranja , koje se pokreću kada se stranica otvori ili zatvori .	interaktivni da biste poboljšali kvalitetu dokumenta možete odrediti akcije , poput promjena uvećanja vrijednost za doći kada se stranica ili zatvoren .	3,00	2,00
77	izvršava navedenu naredbu izbornika kao akciju .	naredba izvršava navedenu izbornika kao akciju .	3,00	2,33

78	reproducira navedenu zvučnu datoteku .	reproducira naveli zvučne datoteke .	2,67	1,67
79	reproducira film kompatibilan s Acrobatom 6 .	reproducira navedenu filmski koji je stvoren kao Acrobat 6-compatible .	2,67	2,00
80	prije dodavanja ove akcije navedite odgovarajuće postavke sloja .	prije dodavanja ta akcija , odredite odgovarajuće postavke sloja .	4,00	3,00
81	Prebacuje se između pokazivanja i sakrivanja polja u PDF dokumentu .	služi za prijelaz između prikazivanje i sakrivanje polja u PDF dokumentu .	3,67	2,67
82	Okidači određuju način pokretanja akcija u medijskim isječcima te na stranicama i u poljima obrazaca .	pokreće akcija određuju kako se aktivira u medijskim isječcima , stranice , i polja obrazaca .	3,00	2,00
83	kada stranica s medijskim isječkom postaje trenutna stranica .	kada stranice koja sadrži medijskog isječka postaje trenutnoj stranici .	1,67	1,67
84	s PDF obrascima i skupnim sljedovima možete koristiti i JavaScript .	također možete koristiti JavaScript s PDF obrazaca i skupni sljedovi .	2,67	2,00
85	strukturirane knjižne oznake za web nalaze se u početku na istim razinama , ali ih možete premještati i ugnijezditi u skupine da biste lakše pratili hijerarhiju materijala na web-stranicama .	web strukturirane knjižne oznake prikazuju se svi na razini , ali možete im promijeniti raspored i nest skupa ih u grupe za Ostavi praćenje hijerarhije materijala na web-stranicama .	3,00	2,00
86	možete uključiti prikaz dijaloškog okvira s URL-om trenutne stranice , naslovom , datumom i vremenom preuzimanja te drugim informacijama .	možete prikazati dijaloški okvir s njihovoj trenutnoj stranici URL naslov , datum i vrijeme preuzimati i druge informacije .	3,00	3,00
87	preglednik će se otvoriti u novom aplikacijskom prozoru na navedenoj stranici .	pretraživač otvara u novom prozor aplikacije na stranicu .	2,67	1,67
88	opišite pravokutnik da biste definirali prvi okvir članka .	povucite pravokutnik da biste definirali prvi članka .	2,00	2,67
89	Za stvaranje, pregledavanje i mijenjanje okvira članaka u PDF dokumentu koristite Alat za članke.	koristite Alat za Article , prikaz i izvršite promjene članka okvir u PDF dokument .	1,67	1,67
90	prilikom uređivanja skupnog slijeda kliknite na Opcije izlaza .	kada uređujete skupni slijed kliknite Opcije izlaza .	3,33	3,67
91	Optimizacija : opcija brzog prikaza za web	Optimizacija : Brzi prikaz za web opcija	4,00	3,33
92	smanjivanje broja piksela	smanjivanje broja piksela	4,00	4,00
93	smanjuje datoteku uklanjanjem nepotrebnih piksela .	smanjuje veličinu datoteke , uklanjanje unnecessary piksel podataka .	3,00	2,33
94	onemogućuje sve akcije povezane sa slanjem ili uvozom podataka iz obrazaca i ponovo postavlja polja obrazaca .	onemogućuje sve akcije povezane s predavanje ili uvoz podatke obrasca i resets polja obrasca .	2,00	2,00
95	podaci obrazaca stapaju se sa stranicom i postaju dio njezina sadržaja .	podaci se obrasca stranici da biste budu objedinjeni sa sadržajem stranice .	2,33	2,00
96	uklanja sve verzije slike osim one	uklanja sve verzije sliku osim jedan	2,67	2,00

	namijenjene prikazivanju na zaslonu .	destined za pregledavanje na zaslonu .		
97	ugrađene sličice , brisanje	ugrađeni sličice brisanje	3,67	3,67
98	fragmentirane slike , spajanje	Objedinjavanje slike fragmented	2,33	2,00
99	uklanja ugrađena kazala za pretraživanje te smanjuje datoteku .	uklanja ugrađene pretraživanje kazala koja smanjuje veličinu datoteke .	2,67	2,00
100	uklanja sve knjižne oznake iz dokumenta .	uklanja sve knjižne oznake iz dokumenta .	4,00	4,00
		sustav 7		
Rbr.	**referentan prijevod**	**strojni prijevod**	**A**	**F**
1	Command + Q	naredbu + Q	1,67	1,67
2	tipkovni prečaci	keyboard shortcuts	1,00	1,00
3	u ovom su odjeljku popisani najčešći prečaci za kretanje po dokumentu .	ovaj odjeljak lists uobičajene shortcuts za moving oko a dokument .	1,00	1,00
4	Označivač za zumiranje	Marquee Zoom tool	1,00	1,00
5	Alat za odabir objekata	odabir Object tool	1,00	1,00
6	Alat za obrezivanje	crop tool	1,00	1,00
7	Kruženje alatima za označavanje crteža : Strelica , linija , pravokutnik , Elipsa , linija poligona , Poligon , Olovka , Gumica	Cycle kroz drawing markup alati : arrow , Line , Rectangle , Oval , Polygon Line , Polygon , Pencil Tool , Eraser Tool	1,33	1,00
8	Kružno kretanje kroz alate za prilaganje : Priloži datoteku , Alat za snimanje zvučnih komentara	Cycle kroz attach alati : Attach File , Record Audio Comment	1,00	1,00
9	tipke za navigaciju unutar PDF-a	za ključeve navigating a PDF	1,67	1,00
10	premještanje žarišta na izbornike (Windows , UNIX) ; proširenje prve stavke izbornika (UNIX)	move focus da menus (Windows , UNIX) ; expand prvi menu stavka (UNIX)	1,00	1,00
11	premještanje žarišta na alatnu traku u pregledniku	move focus da toolbar u browser	1,33	1,00
12	premještanje žarišta na sljedeću karticu u kartičnom dijaloškom okviru	move focus da next tab u tabbed dialog box	1,00	1,00
13	premještanje na sljedeći rezultat pretraživanja i njegovo isticanje u dokumentu	move da next rezultat pretražiti i highlight ga u dokument	1,00	1,00
14	tipke za rad s pločama za navigaciju	za rad s navigation panels ključeve	1,00	1,00
15	kretanje među elementima aktivne ploče za navigaciju	među move na elements od aktivan navigation panel	1,00	1,00
16	Strelica gore ili Strelica dolje	gore ili dolje Arrow Arrow	1,67	1,00
17	premještanje žarišta na prethodnu stavku na navigacijskoj ploči	move focus da previous stavka u navigation panel	1,33	1,00
18	premještanje na prethodno okno	move da previous pane	1,00	1,00
19	promjena prijeloma PDF-a sa strukturnim oznakama i vraćanje na prikaz bez promjene prijeloma	reflow a tagged PDF i povratak na unreflowed vidi	1,00	1,00
20	aktiviranje i deaktiviranje čitanja naglas	activate i deactivate Read Out Loud	1,33	1,33
21	nakon stvaranja , sličice stranica možete ugraditi u PDF .	nakon što ste stvaranje stranicu thumbnails , možete ih embed u PDF	1,67	1,67

		.		
22	ugradnja sprječava ponovno iscrtavanje sličica stranica svaki put kad kliknete gumb Stranice jer to može zahtijevati mnogo vremena .	embedding prevents stranicu thumbnails iz redrawing svaki put ste click Pages na gumb , često a time-consuming proces .	1,00	1,00
23	na ploči Stranice s izbornika Opcije odaberite Ugradi sve sličice stranica ili Ukloni ugrađene sličice stranica .	u Pages panel , izabrati Sve Embed Page Thumbnails ili Remove Embedded Page Thumbnails iz options menu .	1,00	1,00
24	ugradnja ili uklanjanje sličica stranica iz PDF portfelja	ili embed stranici unembed thumbnails Portfolio u PDF	1,00	1,00
25	da biste ugradili sličice stranica , kliknite Ugradi sličice stranica , a zatim Pokreni slijed .	na stranici , thumbnails embed click Embed Page Thumbnails , a zatim click Run Sequence .	1,00	1,00
26	slijedite dobivene upute .	slijedite upute na ako .	1,67	1,00
27	odaberite jedan od sljedećih formata datoteka : BMP , GIF , JPEG , PCX , PNG , EPS / PS ili TIFF .	izabrati jednu od sljedećih file formats : BMP , GIF , JPEG , PCX , PNG , EPS / PS ili TIFF .	2,33	1,67
28	sličice stranica su minijaturni prikazi stranica dokumenta .	stranicu thumbnails su miniature previews od stranica u dokument .	1,67	1,33
29	sličice stranica u možete koristiti za brzo skakanje na odabranu stranicu ili prilagodbu prikaza stranice .	možete upotrijebiti stranicu thumbnails za skok brzo u selected stranicu ili adjust od vidi na stranici .	1,00	1,00
30	kada u programu Adobe Reader ® premještate sličicu stranice , zapravo premještate odgovarajuću stranicu .	u Adobe Reader ® , kada ste move stranicu thumbnail , corresponding move na stranici .	1,00	1,00
31	kada u programu Acrobat premještate , kopirate ili brišete sličice stranica , zapravo premještate , kopirate ili brišete odgovarajuću stranicu .	u Acrobat , kada ste move , kopirati ili delete stranicu , thumbnail move , kopirati ili delete corresponding na stranici .	1,00	1,00
32	u navigacijskom će se oknu pojaviti sličice stranica .	stranicu thumbnails se pojaviti u navigation pane .	1,00	1,00
33	definiranje redoslijeda kretanja	definirali na tabbing order	1,00	1,00
34	na ploči Stranice možete za svaku stranicu zadati redoslijed kojim će se korisnik pritiskom na tipku Tab kretati između polja obrasca , veza i komentara .	u Pages panel , možete postavite na order u kojem a user tabs kroz oblik polja , links i comments za svaku stranicu .	1,00	1,00
35	odaberite sličicu stranice , a zatim s izbornika Opcije odaberite Svojstva stranice .	odabir stranicu thumbnail i izabrati Page Properties iz options menu .	1,00	1,00
36	kreće se redoslijedom definiranim u autorskoj aplikaciji .	kreće se u order specified by na authoring application .	1,00	1,00
37	ako je dokument stvoren u starijoj verziji Acrobata , zadana vrijednost redoslijeda kretanja jest Nije određen .	ako dokument je stvorio u earlier verzija od Acrobat , tab je order Unspecified by default .	1,00	1,00
38	o knjižnim oznakama	o bookmarks	1,00	1,67
39	svaka knjižna oznaka pridružuje se različitom prikazu ili stranici	svaki bookmark goes da drukčiji vidi ili stranicu u dokument .	1,67	1,00

	dokumenta .			
40	prilikom stvaranja svake knjižne oznake u programu Acrobat možete postaviti određene odredišta za knjižne oznake .	u Acrobat , možete postavite bookmark destinations kao stvaranje svaki bookmark .	1,67	1,00
41	knjižne oznake mogu i izvršavati akcije , poput izvršavanja stavki izbornika ili slanja obrasca .	također može bookmarks perform actions primjer , a executing menu stavka ili submitting a oblik .	1,67	1,00
42	knjižne se oznake za neke PDF-ove ponašaju kao pregledi sadržaja .	bookmarks act kao table od contents za neke PDFs .	1,00	1,00
43	otvorite stranicu s kojom želite povezati knjižnu oznaku i prilagodite postavke prikaza .	otvorite stranicu gdje želite da bookmark link da i adjust na vidi settings .	1,67	1,33
44	ako ne odaberete knjižnu oznaku , nova knjižna oznaka automatski će se dodati na kraj popisa .	ako ne odabir a bookmark , novi bookmark je automatski added na kraju popis .	1,67	1,00
45	upišite ili uredite naziv nove knjižne oznake .	tip ili edit naziv novog bookmark .	1,00	1,00
46	u programu Reader knjižne oznake možete učiniti čitljivijima ako promijenite izgled teksta .	u Reader , možete make bookmarks easier da read by changing njihove text izgled .	1,00	1,00
47	prelamanje teksta u dugoj knjižnoj oznaci	wrap text u dugo bookmark	1,00	1,00
48	kliknite gumb Knjižne oznake i s izbornika Opcije odaberite Prelomi duge knjižne oznake .	click na gumb , Bookmarks i izabrati wrap Long Bookmarks iz options menu .	1,00	1,00
49	možete promijeniti izgled knjižne oznake da biste je istaknuli .	možete promijeniti izgled od bookmark da draw attention .	1,67	1,00
50	na ploči Knjižne oznake odaberite jednu ili više knjižnih oznaka .	u Bookmarks panel , odabir jedan ili više bookmarks .	1,00	1,00
51	u oknu dokumenta pomaknite se na odredište koje želite navesti kao novo odredište .	u dokument , pane move na lokaciju želite specify kao novi destination .	1,00	1,00
52	u dijaloškom okviru Svojstva knjižne oznake kliknite Akcije .	u Bookmark Properties dialog box , click Actions .	1,00	1,00
53	brisanje knjižne oznake izbrisat će i sve podređene knjižne oznake .	deleting a bookmark deletes any bookmarks koji su subordinate .	1,00	1,00
54	popis knjižnih oznaka možete ugnijezditi da biste ilustrirali odnose između tema .	možete nest popis od bookmarks a da odnos između topics .	1,00	1,00
55	Gniježđenje stvara odnos nadređeno / podređeno .	nesting creates a parent / child odnos .	1,00	1,00
56	Gniježđenje knjižne oznake (lijevo) i rezultat (desno)	nesting a bookmark (lijevo) i rezultat (desno)	1,00	1,00
57	premještanje knjižnih oznaka iz ugniježđenog položaja	move bookmarks out od nested položaj	1,00	1,00
58	s izbornika opcija odaberite Proširi knjižne oznake najviše razine ili Sažmi knjižne oznake najviše razine .	iz options menu , izabrati Expand top-level Bookmarks ili Collapse top-level Bookmarks .	1,00	1,00
59	strukturirane knjižne oznake pružaju vam veću kontrolu nad sadržajem	tagged bookmarks veći ste give preko kontrolirati stranici content nego	1,00	1,00

	stranice od uobičajenih knjižnih oznaka .	učiniti regular bookmarks .		
60	pretvorene web-stranice obično sadrže označene knjižne oznake .	converted web stranice typically include tagged bookmarks .	1,00	1,00
61	odaberite elemente strukture koje želite navesti kao strukturirane knjižne oznake .	odabir na struktura elements želite kao specified tagged bookmarks .	1,00	1,00
62	Uređivanje strukturnih oznaka pomoću kartice Strukturne oznake	Edit tags s Tags tab	1,00	1,00
63	dodavanje multimedije u PDF-ove	dodati multimedia da PDFs	1,67	1,67
64	opišite pravokutnik na mjestu na kojem želite stvoriti vezu .	povucite a rectangle gdje želite stvoriti link .	1,67	1,00
65	odaberite odredišnu datoteku i kliknite Odaberi .	odabir na destination file i click Select .	1,33	1,00
66	odaberite željene opcije u dijaloškom okviru Stvaranje veze .	odabir na options želite u Create Link dialog .	1,00	1,00
67	mijenjanje svojstava postojeće veze utječe samo na trenutačno odabranu vezu .	na changing properties od an existing link affects samo currently selected link .	1,00	1,00
68	odaberite Alat za veze i dvokliknite pravokutnik veze .	odabir na Link tool i double-click na link rectangle .	1,00	1,00
69	odaberite opciju Zaključano ako želite spriječiti slučajno mijenjanje postavki .	odabir na Locked option ako želite spriječiti users iz accidentally changing your settings .	1,33	1,33
70	PDF-u možete prilagati PDF-ove i druge vrste datoteka .	možete attach PDFs i druge vrstama datoteke u PDF .	1,67	1,67
71	u dijaloškom okviru Dodaj datoteke odaberite datoteku koju želite priložiti i kliknite Otvori .	u Add Files dialog box , odabir na file želite attach i click Otvori .	1,00	1,00
72	na ploči Privici odaberite privitak i s izbornika opcija odaberite Izbriši privitak .	u Attachments panel , odabir an attachment , a zatim izabrati Delete Attachment iz options menu .	1,00	1,00
73	na dnu prozora kliknite Koristi napredne opcije pretraživanja , a zatim odaberite Uključi privitke .	click Use Advanced Search Options na dnu prozora , a zatim odabir Include Attachments .	1,00	1,00
74	akcije se postavljaju u dijaloškom okviru Svojstva .	postavite su actions u Properties dialog .	1,00	1,00
75	dodavanje akcija sa sličicama stranica	dodati actions s stranicu thumbnails	1,00	1,00
76	da biste poboljšali interaktivnost dokumenta , možete definirati akcije , poput mijenjanja stupnja zumiranja , koje se pokreću kada se stranica otvori ili zatvori .	kvaliteta interactive na da enhance od dokument , možete specify changing na primjer , actions zoom vrijednost , za occur kada stranicu ili otvaranja je zatvoren .	1,67	1,00
77	izvršava navedenu naredbu izbornika kao akciju .	Executes a specified naredbu menu kao akciju .	1,00	1,00
78	reproducira navedenu zvučnu datoteku .	na plays specified zvuk file .	1,00	1,00
79	reproducira film kompatibilan s Acrobatom 6 .	plays a specified koji je stvorio kao movie Acrobat 6-compatible .	1,00	1,00
80	prije dodavanja ove akcije navedite	prije dodati ovo akciju , specify na	1,33	1,00

	odgovarajuće postavke sloja .	odgovarajuće sloj settings .		
81	Prebacuje se između pokazivanja i sakrivanja polja u PDF dokumentu .	Toggles između showing i skrivanje a polje u PDF dokument .	1,67	1,67
82	Okidači određuju način pokretanja akcija u medijskim isječcima te na stranicama i u poljima obrazaca .	triggers determine kako su actions activated u media clips , stranica i oblik polja .	1,00	1,00
83	kada stranica s medijskim isječkom postaje trenutna stranica .	kada stranicu containing media clip postaje current na stranici .	1,00	1,00
84	s PDF obrascima i skupnim sljedovima možete koristiti i JavaScript .	također možete koristiti JavaScript s PDF forms i batch sequences .	1,67	1,67
85	strukturirane knjižne oznake za web nalaze se u početku na istim razinama , ali ih možete premještati i ugnijezditi u skupine da biste lakše pratili hijerarhiju materijala na web-stranicama .	web tagged su bookmarks initially sve u isto razina , ali možete ih rearrange i nest pomoći da ih u grupe family keep pratiti od hierarchy od materijal na web stranice .	1,67	1,67
86	možete uključiti prikaz dijaloškog okvira s URL-om trenutne stranice , naslovom , datumom i vremenom preuzimanja te drugim informacijama .	možete display a dialog box s current stranicu je URL , titulu , datum i vrijeme downloaded i druge informacije .	1,67	1,67
87	preglednik će se otvoriti u novom aplikacijskom prozoru na navedenoj stranici .	na browser otvara u novom prozoru application da stranicu ste specify .	1,00	1,00
88	opišite pravokutnik da biste definirali prvi okvir članka .	povucite a rectangle da definirali prvi članak .	1,67	1,67
89	Za stvaranje, pregledavanje i mijenjanje okvira članaka u PDF dokumentu koristite Alat za članke.	koristite za stvaranje tool Article , display , i napravite promjene u članak an box u PDF dokument .	1,67	1,67
90	prilikom uređivanja skupnog slijeda kliknite na Opcije izlaza .	kada editing a batch sequence , click Output Options .	1,33	1,33
91	Optimizacija : opcija brzog prikaza za web	Fast Web View option Optimizing :	1,00	1,00
92	smanjivanje broja piksela	downsample	1,00	1,00
93	smanjuje datoteku uklanjanjem nepotrebnih piksela .	veličina file reduces by eliminating unnecessary pixel data .	1,00	1,00
94	onemogućuje sve akcije povezane sa slanjem ili uvozom podataka iz obrazaca i ponovo postavlja polja obrazaca .	sve disables actions related da submitting ili importing oblik data i resets oblik polja .	1,33	1,00
95	podaci obrazaca stapaju se sa stranicom i postaju dio njezina sadržaja .	oblik data je merged s stranicu da postat stranicu content .	1,33	1,00
96	uklanja sve verzije slike osim one namijenjene prikazivanju na zaslonu .	sve removes versions od an image osim na jedan destined za on-screen viewing .	1,00	1,00
97	ugrađene sličice , brisanje	embedded thumbnails , deleting	1,00	1,00
98	fragmentirane slike , spajanje	fragmented images , merging	1,00	1,00
99	uklanja ugrađena kazala za pretraživanje te smanjuje datoteku .	removes embedded pretražite indexes , koji reduces file veličine .	1,33	1,00

100	uklanja sve knjižne oznake iz dokumenta .	sve removes bookmarks iz dokument .	1,67	1,00
		sustav 8		
Rbr.	**referentan prijevod**	**strojni prijevod**	**A**	**F**
1	Command + Q	naredbu + Q	1,67	1,67
2	tipkovni prečaci	keyboard shortcuts	1,00	1,00
3	u ovom su odjeljku popisani najčešći prečaci za kretanje po dokumentu .	ovaj odjeljak lists uobičajene shortcuts za moving oko a dokument .	1,67	1,00
4	Označivač za zumiranje	Marquee Zoom tool	1,00	1,00
5	Alat za odabir objekata	odabir Object tool	1,00	1,00
6	Alat za obrezivanje	crop tool	1,00	1,00
7	Kruženje alatima za označavanje crteža : Strelica , linija , pravokutnik , Elipsa , linija poligona , Poligon , Olovka , Gumica	Cycle kroz drawing markup alati : arrow , Line , Rectangle , Oval , Polygon Line , Polygon , Pencil Tool , Eraser Tool	1,00	1,00
8	Kružno kretanje kroz alate za prilaganje : Priloži datoteku , Alat za snimanje zvučnih komentara	Cycle kroz attach alati : Attach File , Record Audio Comment	1,00	1,00
9	tipke za navigaciju unutar PDF-a	za ključeve navigating a PDF	1,33	1,00
10	premještanje žarišta na izbornike (Windows , UNIX) ; proširenje prve stavke izbornika (UNIX)	move focus da menus (Windows , UNIX) ; expand prvi menu stavka (UNIX)	1,33	1,00
11	premještanje žarišta na alatnu traku u pregledniku	move focus da toolbar u browser	1,00	1,00
12	premještanje žarišta na sljedeću karticu u kartičnom dijaloškom okviru	move focus da next tab u tabbed dialog box	1,00	1,00
13	premještanje na sljedeći rezultat pretraživanja i njegovo isticanje u dokumentu	move da next rezultat pretražiti i highlight ga u dokument	1,00	1,00
14	tipke za rad s pločama za navigaciju	za rad s navigation panels ključeve	1,67	1,67
15	kretanje među elementima aktivne ploče za navigaciju	među move na elements od aktivan navigation panel	1,00	1,00
16	Strelica gore ili Strelica dolje	gore ili dolje Arrow Arrow	2,00	1,00
17	premještanje žarišta na prethodnu stavku na navigacijskoj ploči	move focus da previous stavka u navigation panel	1,00	1,00
18	premještanje na prethodno okno	move da previous pane	1,00	1,00
19	promjena prijeloma PDF-a sa strukturnim oznakama i vraćanje na prikaz bez promjene prijeloma	reflow a tagged PDF i povratak na unreflowed vidi	1,00	1,00
20	aktiviranje i deaktiviranje čitanja naglas	activate i deactivate Read Out Loud	1,00	1,00
21	nakon stvaranja , sličice stranica možete ugraditi u PDF .	nakon što ste stvaranje stranicu thumbnails , možete ih embed u PDF .	1,67	1,67
22	ugradnja sprječava ponovno iscrtavanje sličica stranica svaki put kad kliknete gumb Stranice jer to može zahtijevati mnogo vremena .	embedding prevents stranicu thumbnails iz redrawing svaki put ste click Pages na gumb , često a time-consuming proces .	1,00	1,00
23	na ploči Stranice s izbornika Opcije	u Pages panel , izabrati Sve Embed	1,00	1,00

	odaberite Ugradi sve sličice stranica ili Ukloni ugrađene sličice stranica .	Page Thumbnails ili Remove Embedded Page Thumbnails iz options menu .		
24	ugradnja ili uklanjanje sličica stranica iz PDF portfelja	ili embed stranici unembed thumbnails Portfolio u PDF	1,00	1,00
25	da biste ugradili sličice stranica , kliknite Ugradi sličice stranica , a zatim Pokreni slijed .	na stranici , thumbnails embed click Embed Page Thumbnails , a zatim click Run Sequence .	1,00	1,00
26	slijedite dobivene upute .	slijedite upute na ako .	1,67	1,67
27	odaberite jedan od sljedećih formata datoteka : BMP , GIF , JPEG , PCX , PNG , EPS / PS ili TIFF .	izabrati jednu od sljedećih file formats : BMP , GIF , JPEG , PCX , PNG , EPS / PS ili TIFF .	2,00	1,67
28	sličice stranica su minijaturni prikazi stranica dokumenta .	stranicu thumbnails su miniature previews od stranica u dokument .	1,67	1,00
29	sličice stranica u možete koristiti za brzo skakanje na odabranu stranicu ili prilagodbu prikaza stranice .	možete upotrijebiti stranicu thumbnails za skok brzo u selected stranicu ili adjust od vidi na stranici .	1,67	1,67
30	kada u programu Adobe Reader ® premještate sličicu stranice , zapravo premještate odgovarajuću stranicu .	u Adobe Reader ® , kada ste move stranicu thumbnail , corresponding move na stranici .	1,00	1,00
31	kada u programu Acrobat premještate , kopirate ili brišete sličice stranica , zapravo premještate , kopirate ili brišete odgovarajuću stranicu .	u Acrobat , kada ste move , kopirati ili delete stranicu , thumbnail move , kopirati ili delete corresponding na stranici .	1,00	1,00
32	u navigacijskom će se oknu pojaviti sličice stranica .	stranicu thumbnails se pojaviti u navigation pane .	1,00	1,00
33	definiranje redoslijeda kretanja	definirali na tabbing order	1,00	1,00
34	na ploči Stranice možete za svaku stranicu zadati redoslijed kojim će se korisnik pritiskom na tipku Tab kretati između polja obrasca , veza i komentara .	u Pages panel , možete postavite na order u kojem a user tabs kroz oblik polja , links i comments za svaku stranicu .	1,00	1,00
35	odaberite sličicu stranice , a zatim s izbornika Opcije odaberite Svojstva stranice .	odabir stranicu thumbnail i izabrati Page Properties iz options menu .	1,00	1,00
36	kreće se redoslijedom definiranim u autorskoj aplikaciji .	kreće se u order specified by na authoring application .	1,00	1,00
37	ako je dokument stvoren u starijoj verziji Acrobata , zadana vrijednost redoslijeda kretanja jest Nije određen .	ako dokument je stvorio u earlier verzija od Acrobat , tab je order Unspecified by default .	1,00	1,00
38	o knjižnim oznakama	o bookmarks	1,00	2,00
39	svaka knjižna oznaka pridružuje se različitom prikazu ili stranici dokumenta .	svaki bookmark goes da drukčiji vidi ili stranicu u dokument .	1,00	1,00
40	prilikom stvaranja svake knjižne oznake u programu Acrobat možete postaviti određene odredišta za knjižne oznake .	u Acrobat , možete postavite bookmark destinations kao stvaranje svaki bookmark .	1,67	1,67
41	knjižne oznake mogu i izvršavati	također može bookmarks perform	1,00	1,00

	akcije , poput izvršavanja stavki izbornika ili slanja obrasca .	actions primjer , a executing menu stavka ili submitting a oblik .		
42	knjižne se oznake za neke PDF-ove ponašaju kao pregledi sadržaja .	bookmarks act kao table od contents za neke PDFs .	1,00	1,00
43	otvorite stranicu s kojom želite povezati knjižnu oznaku i prilagodite postavke prikaza .	otvorite stranicu gdje želite da bookmark link da i adjust na vidi settings .	1,00	1,00
44	ako ne odaberete knjižnu oznaku , nova knjižna oznaka automatski će se dodati na kraj popisa .	ako ne odabir a bookmark , novi bookmark je automatski added na kraju popis .	1,00	1,00
45	upišite ili uredite naziv nove knjižne oznake .	tip ili edit naziv novog bookmark .	1,00	1,00
46	u programu Reader knjižne oznake možete učiniti čitljivijima ako promijenite izgled teksta .	u Reader , možete make bookmarks easier da read by changing njihove text izgled .	1,00	1,00
47	prelamanje teksta u dugoj knjižnoj oznaci	wrap text u dugo bookmark	1,00	1,00
48	kliknite gumb Knjižne oznake i s izbornika Opcije odaberite Prelomi duge knjižne oznake .	click na gumb , Bookmarks i izabrati wrap Long Bookmarks iz options menu .	1,00	1,00
49	možete promijeniti izgled knjižne oznake da biste je istaknuli .	možete promijeniti izgled od bookmark da draw attention .	1,33	1,00
50	na ploči Knjižne oznake odaberite jednu ili više knjižnih oznaka .	u Bookmarks panel , odabir jedan ili više bookmarks .	1,00	1,00
51	u oknu dokumenta pomaknite se na odredište koje želite navesti kao novo odredište .	u dokument , pane move na lokaciju želite specify kao novi destination .	1,00	1,00
52	u dijaloškom okviru Svojstva knjižne oznake kliknite Akcije .	u Bookmark Properties dialog box , click Actions .	1,00	1,00
53	brisanje knjižne oznake izbrisat će i sve podređene knjižne oznake .	deleting a bookmark deletes any bookmarks koji su subordinate .	1,00	1,00
54	popis knjižnih oznaka možete ugnijezditi da biste ilustrirali odnose između tema .	možete nest popis od bookmarks a da odnos između topics .	1,67	1,00
55	Gniježđenje stvara odnos nadređeno / podređeno .	nesting creates a parent / child odnos .	1,33	1,00
56	Gniježđenje knjižne oznake (lijevo) i rezultat (desno)	nesting a bookmark (lijevo) i rezultat (desno)	1,33	1,00
57	premještanje knjižnih oznaka iz ugniježđenog položaja	move bookmarks out od nested položaj	1,00	1,00
58	s izbornika opcija odaberite Proširi knjižne oznake najviše razine ili Sažmi knjižne oznake najviše razine .	iz options menu , izabrati Expand top-level Bookmarks ili Collapse top-level Bookmarks .	1,00	1,00
59	strukturirane knjižne oznake pružaju vam veću kontrolu nad sadržajem stranice od uobičajenih knjižnih oznaka .	tagged bookmarks veći ste give preko kontrolirati stranici content nego učiniti regular bookmarks .	1,00	1,00
60	pretvorene web-stranice obično sadrže označene knjižne oznake .	converted web stranice typically include tagged bookmarks .	1,00	1,00
61	odaberite elemente strukture koje želite navesti kao strukturirane	odabir na struktura elements želite kao specified tagged bookmarks .	1,00	1,00

	knjižne oznake .			
62	Uređivanje strukturnih oznaka pomoću kartice Strukturne oznake	Edit tags s Tags tab	1,00	1,00
63	dodavanje multimedije u PDF-ove	dodati multimedia da PDFs	1,67	1,67
64	opišite pravokutnik na mjestu na kojem želite stvoriti vezu .	povucite a rectangle gdje želite stvoriti link .	1,67	2,00
65	odaberite odredišnu datoteku i kliknite Odaberi .	odabir na destination file i click Select .	1,33	1,00
66	odaberite željene opcije u dijaloškom okviru Stvaranje veze .	odabir na options želite u Create Link dialog .	1,67	1,00
67	mijenjanje svojstava postojeće veze utječe samo na trenutačno odabranu vezu .	na changing properties od an existing link affects samo currently selected link .	1,33	1,00
68	odaberite Alat za veze i dvokliknite pravokutnik veze .	odabir na Link tool i double-click na link rectangle .	1,33	1,00
69	odaberite opciju Zaključano ako želite spriječiti slučajno mijenjanje postavki .	odabir na Locked option ako želite spriječiti users iz accidentally changing your settings .	1,00	1,00
70	PDF-u možete prilagati PDF-ove i druge vrste datoteka .	možete attach PDFs i druge vrstama datoteke u PDF .	1,67	1,67
71	u dijaloškom okviru Dodaj datoteke odaberite datoteku koju želite priložiti i kliknite Otvori .	u Add Files dialog box , odabir na file želite attach i click Otvori .	1,00	1,00
72	na ploči Privici odaberite privitak i s izbornika opcija odaberite Izbriši privitak .	u Attachments panel , odabir an attachment , a zatim izabrati Delete Attachment iz options menu .	1,00	1,00
73	na dnu prozora kliknite Koristi napredne opcije pretraživanja , a zatim odaberite Uključi privitke .	click Use Advanced Search Options na dnu prozora , a zatim odabir Include Attachments .	1,00	1,00
74	akcije se postavljaju u dijaloškom okviru Svojstva .	postavite su actions u Properties dialog .	1,00	1,33
75	dodavanje akcija sa sličicama stranica	dodati actions s stranicu thumbnails	1,67	1,67
76	da biste poboljšali interaktivnost dokumenta , možete definirati akcije , poput mijenjanja stupnja zumiranja , koje se pokreću kada se stranica otvori ili zatvori .	kvaliteta interactive na da enhance od dokument , možete specify changing na primjer , actions zoom vrijednost , za occur kada stranicu ili otvaranja je zatvoren .	1,00	1,00
77	izvršava navedenu naredbu izbornika kao akciju .	Executes a specified naredbu menu kao akciju .	1,33	1,00
78	reproducira navedenu zvučnu datoteku .	na plays specified zvuk file .	1,00	1,00
79	reproducira film kompatibilan s Acrobatom 6 .	plays a specified koji je stvorio kao movie Acrobat 6-compatible .	1,67	1,67
80	prije dodavanja ove akcije navedite odgovarajuće postavke sloja .	prije dodati ovo akciju , specify na odgovarajuće sloj settings .	1,33	1,00
81	Prebacuje se između pokazivanja i sakrivanja polja u PDF dokumentu .	Toggles između showing i skrivanje a polje u PDF dokument .	1,00	1,00
82	Okidači određuju način pokretanja akcija u medijskim isječcima te na stranicama i u poljima obrazaca .	triggers determine kako su actions activated u media clips , stranica i oblik polja .	1,00	1,00

83	kada stranica s medijskim isječkom postaje trenutna stranica .	kada stranicu containing media clip postaje current na stranici .	1,33	1,00
84	s PDF obrascima i skupnim sljedovima možete koristiti i JavaScript .	također možete koristiti JavaScript s PDF forms i batch sequences .	1,00	1,00
85	strukturirane knjižne oznake za web nalaze se u početku na istim razinama , ali ih možete premještati i ugnijezditi u skupine da biste lakše pratili hijerarhiju materijala na web-stranicama .	web tagged su bookmarks initially sve u isto razina , ali možete ih rearrange i nest pomoći da ih u grupe family keep pratiti od hierarchy od materijal na web stranice .	1,67	1,67
86	možete uključiti prikaz dijaloškog okvira s URL-om trenutne stranice , naslovom , datumom i vremenom preuzimanja te drugim informacijama .	možete display a dialog box s current stranicu je URL , titulu , datum i vrijeme downloaded i druge informacije .	1,67	1,67
87	preglednik će se otvoriti u novom aplikacijskom prozoru na navedenoj stranici .	na browser otvara u novom prozoru application da stranicu ste specify .	1,00	1,00
88	opišite pravokutnik da biste definirali prvi okvir članka .	povucite a rectangle da definirali prvi članak .	1,67	1,67
89	Za stvaranje, pregledavanje i mijenjanje okvira članaka u PDF dokumentu koristite Alat za članke.	koristite Article tool da stvarate , display , i napraviti mijenja da na članak box u PDF dokument .	1,67	1,67
90	prilikom uređivanja skupnog slijeda kliknite na Opcije izlaza .	kada editing a batch sequence , click Output Options .	1,00	1,00
91	Optimizacija : opcija brzog prikaza za web	Fast Web View option Optimizing :	1,00	1,00
92	smanjivanje broja piksela	downsample	1,00	1,00
93	smanjuje datoteku uklanjanjem nepotrebnih piksela .	veličina file reduces by eliminating unnecessary pixel data .	1,33	1,00
94	onemogućuje sve akcije povezane sa slanjem ili uvozom podataka iz obrazaca i ponovo postavlja polja obrazaca .	sve disables actions related da submitting ili importing oblik data i resets oblik polja .	1,00	1,00
95	podaci obrazaca stapaju se sa stranicom i postaju dio njezina sadržaja .	oblik data je merged s stranicu da postat stranicu content .	1,00	1,00
96	uklanja sve verzije slike osim one namijenjene prikazivanju na zaslonu .	sve removes versions od an image osim na jedan destined za on-screen viewing .	1,00	1,00
97	ugrađene sličice , brisanje	embedded thumbnails , deleting	1,00	1,00
98	fragmentirane slike , spajanje	fragmented images , merging	1,00	1,00
99	uklanja ugrađena kazala za pretraživanje te smanjuje datoteku .	removes embedded pretražite indexes , koji reduces file veličine .	1,33	1,00
100	uklanja sve knjižne oznake iz dokumenta .	sve removes bookmarks iz dokument .	1,67	1,00

DODATAK C Deskriptivna statistika rezultata ljudske evaluacije

sustav 1						
	evaluator 1		evaluator 2		evaluator 3	
	adekvatnost	fluentnost	adekvatnost	fluentnost	adekvatnost	fluentnost
aritmetička sredina	1.28	1.26	1.12	1.14	1.28	1.2
standardna pogreška aritmetičke sredine	0.0533	0.0543	0.0456	0.0450	0.0533	0.0492
medijan	1	1	1	1	1	1
mod	1	1	1	1	1	1
standardna devijacija	0.5333	0.5434	0.4557	0.4499	0.5333	0.4924
varijanca	0.2844	0.2954	0.2077	0.2024	0.2844	0.2424
mjera spljoštenosti	6.2094	6.8415	20.9938	18.8639	6.2094	11.1947
mjera asimetrije	2.1843	2.4038	4.4026	3.9960	2.1843	3.0051
raspon	3	3	3	3	3	3
minimum	1	1	1	1	1	1
maksimum	4	4	4	4	4	4
sustav 2						
	evaluator 1		evaluator 2		evaluator 3	
	adekvatnost	fluentnost	adekvatnost	fluentnost	adekvatnost	fluentnost
aritmetička sredina	3.32	3.21	2.91	2.93	3.32	3.13
standardna pogreška aritmetičke sredine	0.0633	0.0701	0.1035	0.0934	0.0617	0.0812
medijan	3	3	3	3	3	3
mod	3	3	4	4	3	4
standardna devijacija	0.6337	0.7005	1.0356	0.9347	0.6175	0.8122
varijanca	0.4016	0.4908	1.0726	0.8738	0.3814	0.6596
mjera spljoštenosti	-0.6568	-0.9189	-1.2508	-1.1368	-0.6302	-1.0693
mjera	-0.3825	-0.3158	-0.3178	-0.2372	-0.3293	-0.3595

asimetrije						
raspon	2	2	3	3	2	3
minimum	2	2	1	1	2	1
maksimum	4	4	4	4	4	4
sustav 3						
	evaluator 1		**evaluator 2**		**evaluator 3**	
	adekvatnost	**fluentnost**	**adekvatnost**	**fluentnost**	**adekvatnost**	**fluentnost**
aritmetička sredina	1.36	1.19	1.05	1.11	1.2929	1.1414
standardna pogreška aritmetičke sredine	0.0577	0.0486	0.0329	0.0423	0.0560	0.0431
medijan	1	1	1	1	1	1
mod	1	1	1	1	1	1
standardna devijacija	0.5777	0.4860	0.3295	0.4239	0.5579	0.4288
varijanca	0.3337	0.2362	0.1085	0.1796	0.3112	0.1838
mjera spljoštenosti	3.5603	12.1375	67.3236	25.5874	5.4969	20.2592
mjera asimetrije	1.6920	3.1397	7.8855	4.7353	2.1411	3.9723
raspon	3	3	3	3	3	3
minimum	1	1	1	1	1	1
maksimum	4	4	4	4	4	4
sustav 4						
	evaluator 1		**evaluator 2**		**evaluator 3**	
	adekvatnost	**fluentnost**	**adekvatnost**	**fluentnost**	**adekvatnost**	**fluentnost**
aritmetička sredina	1.32	1.23	1.04	1.08	1.2626	1.2020
standardna pogreška aritmetičke sredine	0.0564	0.0509	0.0315	0.0368	0.0548	0.0475
medijan	1	1	1	1	1	1
mod	1	1	1	1	1	1
standardna devijacija	0.5664	0.5096	0.3153	0.3674	0.5455	0.4733
varijanca	0.3208	0.2596	0.0994	0.1349	0.2976	0.2240
mjera spljoštenosti	4.5838	8.8799	81.2239	41.8628	6.7437	11.6662
mjera asimetrije	1.9419	2.6513	8.8194	5.9759	2.3869	2.9332
raspon	3	3	3	3	3	3
minimum	1	1	1	1	1	1
maksimum	4	4	4	4	4	4
sustav 5						
	evaluator 1		**evaluator 2**		**evaluator 3**	
	adekvatnost	**fluentnost**	**adekvatnost**	**fluentnost**	**adekvatnost**	**fluentnost**
aritmetička	1.21	1.03	1	1.01	1.25	1.11

sredina						
standardna pogreška aritmetičke sredine	0.0433	0.0172	0	0.01	0.0479	0.0314
medijan	1	1	1	1	1	1
mod	1	1	1	1	1	1
standardna devijacija	0.4333	0.1714	0	0.1	0.4794	0.3145
varijanca	0.1878	0.0294	0	0.01	0.2298	0.0989
mjera spljoštenosti	2.3315	29.8977	-	100	2.1306	4.4956
mjera asimetrije	1.8193	5.5946	-	10	1.7193	2.5310
raspon	2	1	0	1	2	1
minimum	1	1	1	1	1	1
maksimum	3	2	1	2	3	2
sustav 6						
	evaluator 1		**evaluator 2**		**evaluator 3**	
	adekvatnost	**fluentnost**	**adekvatnost**	**fluentnost**	**adekvatnost**	**fluentnost**
aritmetička sredina	3.24	2.83	2.76	2.57	3.24	2.88
standardna pogreška aritmetičke sredine	0.0683	0.0817	0.1045	0.0967	0.0683	0.0819
medijan	3	3	3	2	3	3
mod	3	2	3	2	3	2
standardna devijacija	0.6834	0.8171	1.0456	0.9667	0.6834	0.8199
varijanca	0.4671	0.6677	1.0933	0.9344	0.4671	0.6723
mjera spljoštenosti	-0.8333	-1.4286	-1.0913	-1.0025	-0.8333	-1.4773
mjera asimetrije	-0.3447	0.3252	-0.3127	0.1402	-0.3447	0.2268
raspon	2	2	3	3	2	2
minimum	2	2	1	1	2	2
maksimum	4	4	4	4	4	4
sustav 7						
	evaluator 1		**evaluator 2**		**evaluator 3**	
	adekvatnost	**fluentnost**	**adekvatnost**	**fluentnost**	**adekvatnost**	**fluentnost**
aritmetička sredina	1.33	1.15	1.01	1	1.27	1.14
standardna pogreška aritmetičke sredine	0.0472	0.0358	0.01	0	0.0468	0.0348
medijan	1	1	1	1	1	1
mod	1	1	1	1	1	1
standardna	0.4725	0.3589	0.1	0	0.4682	0.3487

devijacija						
varijanca	0.2233	0.1287	0.01	0	0.2192	0.1216
mjera spljoštenosti	-1.4912	2.0012	100	-	0.5781	2.4877
mjera asimetrije	0.7341	1.9903	10	-	1.3496	2.1067
raspon	1	1	1	0	2	1
minimum	1	1	1	1	1	1
maksimum	2	2	2	1	3	2
sustav 8						
	evaluator 1		**evaluator 2**		**evaluator 3**	
	adekvatnost	**fluentnost**	**adekvatnost**	**fluentnost**	**adekvatnost**	**fluentnost**
aritmetička sredina	1.3	1.18	1.02	1.02	1.27	1.17
standardna pogreška aritmetičke sredine	0.0460	0.0386	0.0140	0.0140	0.0446	0.0377
medijan	1	1	1	1	1	1
mod	1	1	1	1	1	1
standardna devijacija	0.4605	0.3861	0.1407	0.1407	0.4461	0.3775
varijanca	0.2121	0.1490	0.0197	0.0197	0.1990	0.1425
mjera spljoštenosti	-1.2398	0.8777	47.4177	47.4177	-0.9119	1.2060
mjera asimetrije	0.8862	1.6913	6.9620	6.9620	1.0519	1.7839
raspon	1	1	1	1	1	1
minimum	1	1	1	1	1	1
maksimum	2	2	2	2	2	2

ŽIVOTOPIS I POPIS RADOVA

Ivan Dunđer, mag. inf., inženjer je informacijske tehnologije i zaposlen kao asistent na Odsjeku za informacijske i komunikacijske znanosti pri Filozofskom fakultetu Sveučilišta u Zagrebu. Rođen je 29.12.1985. u Zagrebu. Znanstveno-istraživački je angažiran na području računalne obrade prirodnog jezika, strojnog prevođenja i evaluacije, jezičnih i govornih tehnologija, upravljanja znanjem te modeliranja i razvoja baza podataka i informacijskih sustava. Izlagao je na brojnim međunarodnim konferencijama s međunarodnom recenzijom te se znanstveno i stručno usavršavao na seminarima, certificiranim programima edukacije, javnim tribinama, na konferencijama i radionicama u Hrvatskoj i inozemstvu. Aktivno sudjeluje u radu znanstvenih i stručnih udruženja te se služi njemačkim, engleskim i francuskim jezikom. Oženjen.

Matični broj znanstvenika:

345536

E-pošta:

ivandunder@gmail.com

Znanstveno-istraživački interesi:

strojno prevođenje i evaluacija

računalna obrada prirodnog jezika (NLP)

strojno učenje

jezične tehnologije i resursi

govorne tehnologije

obrada i upravljanje znanjem

modeliranje i razvoj baza podataka i informacijskih sustava

Popis radova u Hrvatskoj znanstvenoj bibliografiji (CROSBI) dostupan je na: http://bib.irb.hr/lista-radova?autor=345536

Odabranih 5 radova:

1. Sanja Seljan, Marko Tucaković, **Ivan Dunđer**. *Human Evaluation of Online Machine Translation Services for English/Russian-Croatian*. WorldCIST'15 - 3rd World Conference on Information Systems and Technologies. New Contributions in Information Systems and Technologies (vol 1): Advances in Intelligent Systems and Computing (volume 353). Álvaro Rocha, Ana Maria Correia, Sandra Costanzo, Luís Paulo Reis (ur.). Springer International Publishing. 1.-3.4.2015. Azori, Portugal. ISBN 978-3-319-16485-4; ISSN 2194-5357, pp. 1089-1098 (predavanje, međunarodna konferencija, međunarodna recenzija, rad objavljen in extenso, znanstveni rad)
2. Sanja Seljan, **Ivan Dunđer**. *Automatic Quality Evaluation of Machine-Translated Output in Sociological-Philosophical-Spiritual Domain*. CISTI'2015 - 10th Iberian Conference on Information Systems and Technologies: Sistemas e Tecnologias de Informação - Vol. II Artigos Curtos, Artigos Poster, Simpósio Doutoral. Álvaro Rocha, Arnaldo Martins, Gonçalo Paiva Dias, Luís P. Reis, Manuel Pérez Cota (ur.). Águeda : AISTI (Associação Ibérica de Sistemas e Tecnologias de Informação). 17.-20.06.2015. Aveiro, Portugal. ISBN: 978-989-98434-5-5, pp. 128-131 (predavanje, međunarodna konferencija, međunarodna recenzija, rad objavljen in extenso, znanstveni rad)
3. **Ivan Dunđer**. *CroSS: Croatian Speech Synthesizer - design and implementation*. 16th International Multiconference INFORMATION SOCIETY - IS 2013 / Collaboration, Software and Services in Information Society (CSS'2013). Jožef Stefan Institute (Ljubljana). Matjaž Gams, Rok Piltaver, Dunja Mladenić, Marko Grobelnik, Franc Novak, Bojan Blažica, Ciril Bohak, Luka Čehovin, Marjan Heričko, Urban Kordeš, Zala Kurinčič, Katarina Marjanovič, Toma Strle, Vladimir A. Fomichov, Olga S. Fomichova, Vladislav Rajkovič, Tanja Urbančič, Mojca Bernik, Andrej Brodnik (ur.). 7.-11.10.2013. Ljubljana, Slovenija. Vol. A, pp. 257-260 (predavanje, međunarodna konferencija, međunarodna recenzija, rad objavljen in extenso, znanstveni rad)
4. **Ivan Dunđer**, Sanja Seljan, Marko Arambašić. *Domain-Specific Evaluation of Croatian Speech Synthesis in CALL*. 7th European Computing Conference (ECC '13) - Recent Advances in

Information Science (Recent Advances in Computer Engineering Series 13) / Language and Text Processing. World Scientific and Engineering Academy and Society - WSEAS (Atena). Damir Boras, Nives Mikelić Preradović, Francisco Moya, Mohamed Roushdy, Abdel-Badeeh M. Salem (ur.). 25.-27.6.2013. Dubrovnik, Hrvatska. ISBN: 978-960-474-304-9, ISSN: 1790-5109, pp. 142-147 (predavanje, međunarodna konferencija, međunarodna recenzija, rad objavljen in extenso, znanstveni rad)

5. Sanja Seljan, **Ivan Dunđer**, Angelina Gašpar. *From Digitisation Process to Terminological Digital Resources*. 36th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2013) / Intelligent Systems (CIS). Croatian Society for Information and Communication Technology, Electronics and Microelectronics - MIPRO (Rijeka). Petar Biljanović (ur.). 20.-24.5.2013. Opatija, Hrvatska. ISBN: 978-953-233-074-8, ISSN: 1847-3938, pp. 1329-1334 (predavanje, međunarodna konferencija, međunarodna recenzija, rad objavljen in extenso, znanstveni rad)