Zdravo,

Sem se prebil skozi, čeprav je bilo težko. Predvsem je grozno koliko je tipkarskih in drugih napak v tekstu, kar je še kar žaljivo do bralca, saj ti nisem lektor. No, ko sem jih opazil sem jih tudi dodal v komentarje, čeprav sem se nekje na sredi naveličal. Naslednja verzija prosim, da je čistopis, v nasprotnem primeru jo vrnem, ne da bi jo sploh bral.

Sicer je dr. boljši kot prejšnjič, mu pa še vedno manjka, da bi bil ok. Struktura se mi zdi nekam zmedena, iste stvari se večkrat vpeljuje. Zdi se mi tudi, da so obrobne stvari zelo na dolgo pisane, ključne pa zelo na kratko, ravno obratno kot bi bilo fino. Npr. MULTEXT-East in 1984 bi moral nekje na začetku predstavit, skupaj z referencami, saj ti je osnovna podatkovna množica, nato pa se enostavno na to sklicevat, npr. korpus »1984« ne pa vsakič korpus (Erjavec 2004). Terminologijo bi tudi moral uvesti, in se jo potem konsistentno držat. Sedaj je enkrat tako, drugič drugače. Svojih člankov niti ne omenjaš vseh pred zaključkom, če pa že jih pa dostikrat ni jasno kako se navezujejo na besedilo. Večkrat tui ni jasno kaj povzemaš po drugih, kje si pa ti naredil novo zadevo.

Podrobne opombe spodaj, bi pa prosil, da za naslednji verziji priložiš te komentarje skupaj z razlago kako si jih upošteval:

1. angleške termine ven iz naslovov poglavij

Oznaˇcevanje POS - Part Of Speech -> Oznaˇcevanje POS

Konˇcno urejanje – post-editing —> Konˇcno urejanje

Podobnosti slovanskih jezikov pri prevajanju (free rides) —> Podobnosti slovanskih jezikov pri prevajanju (v besedilu pa uvedem angleški opis)

1. MSD ni »oznaka morfoloških in sintaktičnih lastnosti besede« ampak morfosintaktičnih lastnosti. Torej presek, ne unija, kar je dosti huda razlika.

popravil

1. parsing je (skladenjsko) razčlenjevanje, ne (kar) analiza.

Popravil na vseh mestih

pregledal še vse primere, ko se pojavi analiz\* in tudi tukaj ustrezno popravil (dva primera nesmiselnih stavkov)

1. »To delo predstavlja pregled metod,« verjetno še kaj več od tega..

Pričujoče delo predstavlja sisteme za strojno prevajanje sorodnih jezikov. Prikazan je pregled že izdelanih prevajalnih sistemov za izbrane jezikovne pare. Opisane so osnovne komponenete ter jezikovna gradiva, ki jih uporablja večina teh sistemov. Poseben poudarek je na opisu komponent ter gradiv zbirke orodij in gradiv za postavitev novih sistemov za strojno prevajanje sorodnih jezikov Apertium \cite{corbi1}. V tem delu so med drugim predstavljene metode, ki omogočajo samodejno ustvarjanje vseh gradiv, ki so potrebna za postavitev sistema za prevajanje naravnih sorodnih jezikov.

1. »Sistem je uporabljal globoko sintaktično analizo (deep syntactical analisys) in .«   
   Ma joj, da moraš tako šlampat! Narobe angleški termin (2x) in ni konca stavka. Pa že V2..

popravil na 1 mestu

1. Na začetku poglavij (npr. 2.1) je tudi lepo dati pregled o čem bo tekla beseda.

Naredil, tudi za 2.0

1. 2.1 po ne paše sem, ampak prej pred 3.4. Če drugega ne, tu operiraš s pojmi, ki sploh še niso bili uvedeni. Pa tudi sicer vun štrli.

sem pisal Tomažu

popravil po napotkih s telefona

1. »Plitvi so predlagali ter ga implementirali na poskusni razliˇcici sistema« ???

Pri izdelavi stvarnega kazala sem zamenjal vse besede transfer z \index{transfer}, namesto transfer\index{transfer}, to sem popravil

1. »Predstavljeno je ogrodje Apertium, ki predstavlja osnovno platformo za

skupaj z naslednjo točko

1. veˇcino implementacij metod, ki so predstavljene v tem delu. Poglavje 4 predstavlja …«   
   5x »predstavlja/eno« in naprej še ničkolikokrat. Daj malo variiraj, se bo lepše bralo.

Res je, 9x predstavlja:(

Zdaj samo 2x

1. »Upogibna morfologija« to je kalk iz angleščine. V slovenščini je drug izraz, katerega odkritje prepustim avtorju.

Toporišič v knjigi sploh nima tega, ker je šele pozneje začel pisati o „oblikoskladnji“. Zaenkrat mi je še najlepši izraz „oblikoskladenjska slovnica“

1. »po (Janda, 2007)« če se neposredno sklicuje na delo, se navaja a la »Janda (2007)«

popravil

1. »Oznake POS predstavljajo oznake besednih vrst, z oznaˇcevanjem POS pa opisujemo proces oznaˇcevanja morfoloških ter v nekaterih primerih tudi morfosintakti ˇcnih znaˇcilk kot so oznake MSD definirane v (Erjavec, 2004).« To je zelo zmedeno povedano, pa še precej stara referenca. Poglej članke o JOSu. Odloči se tudi, ali boš res govoril o PoS označevanju, ali o označevanju MSD in ustrezno uvedi ti dve kratici.

Uredil, predelal obe poglavji ter dodal najpomembnejša dela za slovenščino

1. »velik del besed predstavlja razliˇcne oznake POS« predstavlja?

popravil

1. 2.1.1.1 bi lahko bil daljši, da bi vsaj omenil par metod in programov za PoS tagging.

to delam v kompletu z 13

1. Paradigme so .. »potencialno rekurzivne matrike« Prvič slišim, da so lahko rekurzivne. Imaš za to kakšno referenco?

Dodal referenco

1. »je kontroverzno razpravljan« pa kakšna slovenščina je to? Kot da bi kakšnega američana slovenskega rodu 3 generacije poslušal..

lepše zapisal isto vsebino, nadaljevanje pa povsem spremenil

1. »so paradigme so le skupine«

popravil

1. »zadošča če opis razred elementov s podobnostmi, v našem primeru bodo elementi besede oziroma besedne zveze (words, phrases and phrase chunks).« kako naj paradigma opisuje besedne zveze? Odstavek ki sledi temu, je tudi zelo zmeden. Na novo napisat, bolj koherentno razložit kaj (tebi) paradigma pomeni kaj čemu jo potrebuješ.

Na novo napisal, upošteval pripombe, zbrisal del o besednih zvezah v paradigmah.

1. »2.1.1.2.1 Samodejno lušˇcenje paradigem« ali te podrobnosti res spadajo sem? No, mogoče, ampak v vsakem primeru ne moreš samo navesti reference brez kakršnekoli diskusije.
2. (Erjavec in Džeroski, 2004) se ne ukvarja z luščenjem paradigem temveč z lematizacijo.

Spremenil v manandhar98

1. »kakršna je tudi slovenšˇcine.«

popravil

1. »Sestavljeno je iz korena, vrhnje vozlišˇce, vej drevesa, notranja vozlišˇca ter listov« ujemanje po sklonu

popravil

1. »Rdeˇci avto je vozil« to se meni sploh ne sliši pravilen stavek. Če je to edini primer, se pa ja spomni nekega normalnega.

Popravil v Avto rdeče barve je vozil po cesti}

1. »vendar pa ne navaja notranjo gradnikov«, »Plitki je naravno nadaljevanje plitke analize«   
   teh napak je enostavno preveč. Daj prosim komu v lekturo, ker takle res ne moreš (ne bi smel) iz rok dat. Naslednjič zavrnem branje če not najdem več kot tri brutalne jezikovne napake.

Popravil, tudi ti dve napaki izvirata iz \index označevanja

tudi v lekturo bom dal, ko odpravim napake

1. »Statistiˇcni model jezika dodeljuje verjetnost zaporedju besed« Stat. modeli lahko delajo še marsikaj drugega, npr. določitev ključnih besed.

Napisal sem malo bolj splošno ter tudi predstavil izbrane LMje

1. »V ˇclanku je uporabljen« to ni članek ampak disertacija. Tu tudi ni mesto govorit o tem, ker drugod tudi nisi.

Popravil.

1. »tako bi lahko, z doloˇcenimi omejitvami, ta jezik uporabili kot univerzalni slovanski jezik za strojno prevajanje« Ta je pa lovska.

Zbrisal, ideja je padla na pijači s Tadićem

1. »2.3 … Free rides« V naslovih res ni treba pisat angleških terminov

popravil

1. »“My brother read a book.”« zakaj prevod v angleščino?

Ker sem primer skopiral iz članka:( popravil, zdaj je samo sl primer

1. »Sutra ´cu kupiti kolo« Srbi ne rečejo »kolo« razen če mislijo na ples.. Ali nisem tega že komentiral v prejšnji verziji?

popravil

1. »Slika 2.2: pravilo..« Po moje tu ni mesto tej sliki, ker nisi stvari ilustriral za druga podpoglavja, pa tudi ne uvedel formata ali vsega ostalega, in je zato primer nerazumljiv.

Popravil ter dodal referenco na priloga1.

Dodaj še natančnejši opis formata nekje v nadaljevanju, verjetno pri pravilih transferja – to sem dodal in sicer v razdelku label{section:apertiumovformatpravil}, tukaj na primeru razložim format pravil.

1. »Tabela 2.1« Lahko bi našel še malo več primerov, posebej s SH.

Dodal še dva primera prav za sr/hr

1. Kaj pa podobnice / cognates? S tem se da uloviti dosti prevodov besed (mleko – mlijeko itd) Spodobilo bi se vsaj omenit.

Dodal razdelek ter tabelo s primeri

1. »full-fledged translation of natural languages with no user intervention« Temu se je tradicionalno reklo FAMT

popravil, zdaj piše FAMT

1. »(Apertium, 2010).« Čudno navajanje sistemov

Popravil v footnote

1. »SMT« zakaj ni podana tudi razvezana oblika tule, ko prvič uvedeš kratico?

popravil

1. »3.1.1 Statistiˇcno strojno prevajanje – SMT« Tudi kratic ni treba v naslove poglavij

popravil

1. »Statistiˇcno strojno prevajanje, Statistical Machine Translation – SMT« argh, prvič ne poveš kaj to je, sedaj pa že tretjič ponavljaš. Res si v nulo šlampast.

popravil

1. »globlje razumevanje med modelom in predmetom« A modeli potem razumejo predmet?

Tudi sam ne vem kaj sem želel povedat, brisal ven, poved ima še vedno smisel, oziroma šele sedaj ga ima.

1. »Sistemi te paradigme izvorno besedilo« ker paradigma uporabljaš v tehničnem smislu, se raje izogni drugačni uporabi.

Paradigme -> skupine

1. Str. 27 »strukturni« kaj?

Strukturni transfer, \index..., avtomatsko označevanje

1. »finite state transductors« verjetno transducers

popravil sem v transducers

1. »Sledi ponovna uvedba oznaˇcevanja« kaj to pomeni? Se tudi ne sklada s sliko.

Sledi ponovno dodajanje delov besedila, ki ga je prvi modul posebej označil in se zato niso prevajali. Popravil sem vrsti red opisa modulov, post-generator je pred morfološko sintezo.

1. »izbere najverjetnejšo izbiro izmed ponujenih izbir« joooj..

Sistem uporablja stohastični označevalnik oblikoskladenjskih oznak, ki izbere iz seznama možnih oznak, ki jih je našel modul za oblikoskladenjsko analizo, najverjetnejšo oznako.

1. »Rdeˇce drevo« zakaj vedno taki bizarni primeri?

Mogoče še iz najbolj znanega primera iz članka Brown et. al.: Una bruja verde (zelena čarovnica)

nov primer: odprto okno

zamenjal sem tudi oznake (zdaj so po JOSu)

1. »Osnova za ugotavljanje kakovosti prevoda je statistiˇcni jezikovni model ciljnega jezika.« To je ključni nov modul v spremenjeni arhitekturi, in moral bi ga bolj podrobno opisa. Statistični jezikovni model lahko pomeni karkoli..

popravil, opisal sem razliične jezikovne modele in predstavil zakaj sem izbral enostavne trigram modele

1. »Izdelava oznaˇcevalnika POS, še posebej dobre kakovosti, ni enostavna naloga.« To je še kako res, ampak za slovenščino je bil že narejen, c.f. JOS in tako ta problem odpade, in s tem tudi tvoj argument. Glede na relevantnost zate, bi te prosil, da omeniš publikacije in rezultate tega projekta in kakšno zvezo imajo s tvojim delom, oz. zakaj nimajo nobene. Poleg tega bi nekje želel še diskusijo o tem, da če bi uporabil PoS označevalnik, niti ne rabiš morfološkega analizatorja, saj že tagger vsebuje leksikon besednih oblik, in ti torej ena stopnja analize odpade.

JERNEJNOVO Mail Erjavec

1. »Stopnja kakovosti oznaˇcevanja današnjih najboljših, state-of-the-art, oznaˇcevalnikov POS za visoko pregibne jezike, kot sta (Hajiˇc, 2000) in (Erjavec, 2006) je relativno nizka, v primerjavi s kakovostjo oznaˇcevalnikov POS za analitiˇcne jezike, kot je angleški jezik, in tudi v primerjavi s splošno kakovostjo prevajalnih sistemov za sorodne jezike.«   
   Erjavec 2006 je stara publikacija, in modeli JOS dajo bistveno boljše rezultate. Poleg tega je kvaliteta MSD označevanje res nizka, ampak PoS pa čisto primerljiva z angleščino. (sploh bi bilo dobro da razdelaš ta pojma na začetku – kaj je MSD in kaj PoS označevanje in katerega uporabljaš ti.)

Na novo napisal

1. »Veˇc kandidatov za prevode omogoˇca izbor najboljših kandidatov v zakljuˇcni fazi, ko so zbrani vsi razpoložljivi podatki za prevod.« To je verjetno res, ampak kaj pa kombinatorična eksplozija? Kaj besed o tem.

napisal

1. »Pregled uporabljenih podatkovnih tipov« to so po moje prej jezikovni viri (ali modeli), ne pa podatkovni tipi, kar prej spominja na spremenljivke v programskih jezikih (integer, string ipd)

Pregled uporabljenih podatkovnih tipov -> jezikovnih virov, povsod v besedilu

1. »Statistiˇcni jezikovni model ciljnega jezika 1« fusnota piše isto kot besedilo; če ne prej, bi pa tu moral podrobno razlužiti kako ta modul dela.

Popravil, nova razlaga je nekoliko više v besedilu

1. »Jezikovni model morfoloških oznak (POS) izvornega jezika 2« Velja isto ko pri prejšnji alineji.

Popravljeno,

1. »je bila izdelana nova metoda oziroma je bila uporabljena že znana metoda za samodejno izdelavo podatkov tega tipa« ???

popravil: glej 55

1. »Veˇc popolnoma delujoˇcih sistemov smo izdelali s pomoˇcjo opisanih metod.« Čuden vrsten red; bolj pomembno; kje so ti sistemi predstavljeni? Upam, da v nadaljevanju disertacije!

Popravil: Gornji spisek vsebuje vsa jezikovna gradiva, ki so potrebna za postavitev prevajalnih sistemov spremenjene arhitekture, kot je opisana v Razdelku \ref{section:predlaganaspremenjenaarhitektura}. Za vsako jezikovno gradivo smo poiskali metodo, ki omogoča samodejno izdelavo tega gradiva. Če metode nismo našli, oziroma še ni bila predstavljena ali pa so bila gradiva neprimerna, smo izdelali novo metodo. Metode so natančneje predstavljene v Razdelku \ref{section:metode}.\\

Več popolnoma delujočih sistemov smo izdelali s pomočjo opisanih metod. Sistemi so predstavljeni v Razdelku \label{section:description}, kakovost prevodov\index{prevod} je ovrednotena v Poglavju \ref{chapter:evalvacija}.

1. »Sistemi za strojno prevajanje na osnovi pravil plitkega transferja enostavne metode

Verjetno je bila napaka samo v tisti verziji pdf, zdaj je vse OK

1. analize, transferja ter sinteze pri izdelavi prevodov iz izvornega v ciljni jezik.« Glagoli so ti očitno odveč.

Verjetno je bila napaka samo v tisti verziji pdf, zdaj je vse OK

1. »morfološko oznaˇcenih enojeziˇcnih slovarjev« zakaj ne kar »morfološki leksikon«?
2. »Z drugimi

nič

1. besedami, vsebujejo vse leme, katerih vse besedne oblike se spreminjajo na enak način za vse morfološke oznake (Erjavec, 2004).« Tega jaz sigurno nisem tako napisal.

Popravil: za vse morfološke oznake - MSD, morfološke oznake so obširneje predstavljene v Razdelku \ref{section:msd}.

1. »Morfemi vkljuˇcujejo krne« To ne drži. Krn nima prave veze z morfologijo.

Popravil (res grda napaka): (concatenated) morfemov. Morfemi so natančneje razliženi v Razdelku \ref{section:morfem}

1. »N,NT« kaj je NT? In, zakaj ne uporabljaš oznak MULTEXT V4 / JOS?   
   V Sloveniji so postale skorajda standard (npr. Fida, SSJ); zakaj je dobro, da si ti izmisliš nove?

vprašaj Erjavca: uporabil sem standardne oznake Apertiuma, verjetno pa bi bilo za to predstavitev bolje ali: vse pretvoriti v JOS ali pa izdelati poseben razdelek, ki bi opisal primerjavo/razlike

Sem vprašal in pretvoril vse v JOS, dodal sem tabele z opisi MSD oznak, ki jih uporabljam

1. »smernicah TEI-P4 (Consortium, 2007).« V resnici P5. Kar nek »Consortium« tudi ne more biti avtor..

popravil po njihovih navodilih: TEI-Consortium

1. »Uporabili smo korpus (Erjavec, 2010)« Zakaj ne kar »1984«, tako samo zmedeš bralca.

Popravil: Uporabili smo korpus "1984", natančneje je predstavljen v Razdelku

...

1. »zapisan \*me morfo-sintaktiˇcna oznaka« odloči se tudi kako jih boš poimenoval in se potem tega drži

morfosintaktična oznaka

1. »iz oznaˇcenega korpusa, kot je (Erjavec, 2010).« ditto 65., Erjavec ni označen korpus…

popravil

1. Slika 4.6: »for vsaka do« vsaka kaj? Tudi sicer je ta algoritem zelo lakonično zapisan; prosim razširi.

popravil

1. »ˇce se noben par vsi zapisov ne izkljuˇcuje«

popravil

1. »Torej dve paradigmi družimo, ˇce ena predstavlja popolno podmnožico druge paradigme.« Torej ni treba, da sta enaki? Ampak če je ena za množinske samostalnike, a ne bodo potem ti postali tudi edninski? Npr. škarje, vrata, Jesenice,… Prosim pojasni.

JERNEJNOVO pojasni, najprej sortiram

1. »Enojeziˇcni morfološko oznaˇceni slovarji izvornega in ciljnega jezika so bili zgrajeni« Pojasni še, zakaj nisi enostavno vzel MULTEXT-East lexicona za slovenščino (in srbščino).

Metoda omogoča gradnjo takšnih leksikonov, preizkušena je bila na jezikovnem paru, ki takšen leksikon sicer vsebuje, vendar jo lahko uporabljamo tudi za ostale jezikovne pare, ki takšnih leksikonov nimajo (recimo še vedno hrvaščina:). Ko je bila izdelana metoda tudi srbščina še ni imela leksikona.

1. Prvi odstavek 4.4.2: ali te trditve ne obstaja kakšna referenca?

Dodal sem referenco katz1987

1. »Izdelali smo metodo poravnave lem.« To je nekam abruptno - malo bolj upelji – zakaj, kje?

Popravil, najprej predstavim sparse data problem, povem, da smo ga reševali na dva načina ter ju predstavim

1. »morfološko-skladenjska oznaka« ampak čisto vsakič uporabiš drug termin.. Takole je še najgrše, da združiš latinski in slovenski termin..

popravil v slovensko oblikoskladenjska in ne morfosintaktična – to sem izbral

1. »Metoda poravnave lem omogoˇca boljše rezultate v primerjavi s poravnavo besed v korpusu, zaradi zmanjšanja prostora iskanja.« Zakaj to dvakrat poveš?

popravil

1. »Oglejmo si zmanjšanje iskalnega prostora z opisano metodo:…« Nekaj, kar je zelo enotavno (da je lem manj kot besednih oblik) si tu silno zakompliciral, v bistvu mi ni jasno zakaj, saj je stvarže inuitivno čisto jasna.

Popravil, skrajšal, vendar še vedno razložim bistvo

1. »S predpostavko da je roman Georga Orwella "1984"(Orwell, 1949), ki sestavlja veˇcjeziˇcni povedno poravnani del korpusa (Erjavec, 2010),« Pa saj 10 strani skoz pišeš o tem korpusu, od kod sedaj potreba, da ga tako na široko ponovno predstavljaš? Dosti bolje bi bilo, da ga na začetku spodobno opišeš (kar ga ne), in poveš točno kako je označen, v katerih jezikih je in koliko je velik. Tako pa nekaj po kapljicah..

dodal sem poseben razdelek, kjer predstavim MULTEXT-EAST ter wiki korpuse

1. »odan modul za rangiranje prevodov na osnovi stohastiˇcnih metod« To je tvoje delo ali od čehov? Če drugo, se pohvali, če drugo, daj referenco.

Dodal referenco vicic2009b

1. »predhodni moduli prevajalnega sistema pa zagotavljajo, da se pri prevodu ohrani pomen izvorne povedi« To je pa močna trditev, da kar zagotavlja! Še ljudje se kdaj zmotijo, pa se ne bi program… Če drugega ne, kaj pa dvoumne besede?

modul poišče le slovnično najbolj pravilno poved v ciljnem jeziku. Arhitektura se zanaša na predhodne module pri ohranjanju pomena v končnem prevodu.

1. »uporabljenih jezikov iz Wikipedije« Če se ne motim, sploh še ni bilo predstavljeno, kateri so »uporabljeni« jeziki.

Dodal poseben razdelek med uporabljena učna gradiva

1. »Velikost korpusov je bilo«

Ta del sem napisal popolnoma na novo, ga skrajšal ter opis korpusov prestavil v poseben razdelek (više)

1. »ujemanje morfoloških kategorij lahko modeliramo s pravili, ki temeljijo na regularnih izrazih.«

popravil: find/replace problem

1. »saj je bilo najbolj enostavno uporabiti enako tehnologijo.« nepotrebno.

zbrisal

1. Slika 4.10: ta del je zanimiv, pa ga predstaviš v dveh stavkih, nobenega primera, nerazumljiv algoritem. Prosim razširi.

Razširil in opisal zakaj to potrebujem in tudi kako metoda deluje.

1. Začetek Poglavja 5: ali je to res mesto, da spet razlagaš osnove MT? To paše v uvod, ne sem; če kaj, bi sem pasalo konkretno kako dela Apertium.
2. »na Primeru 5.2« kaj je ta primer 5.2?

popravil: na Sliki 5.2

1. »Primer pravila je predstavljen na Sliki 5.2.« to pravilo je popolnoma nerazumljivo. Razloži lepo kaj naredi in kaj vsak posamezni del pomeni, in pojasni kriptične labele: vbserfti, temps, persona, …

popravil, format je opisan v Apertiumov format pravil, kriptične labele so zamenjane z opisi, primer: pomožni glagol v prihodnjiku

1. »Veˇc pravil je predstavljenih v Prilogi A.« Več ali vsa? Jaz pi pričakoval vsa. Btw, doktoratu bi se spodobilo kot priloga dodati CD, kjer je cel sistem.

Vsa pravila, ki so bila ročno pregledana in smiselna, izdelal sem tudi cel kup pravil, ki so narejena s samodejno metodo, teh ne bi izpisoval.

CD bo priložen, to sem imel v načrtu (JERNEJ NE POZABI)

1. »Samodejne metode popolne slovniˇcne analize povedi ne dosegajo« že mogoče res, samo kaj ima to opraviti s samodejno izdelavo pravil za plitvi prenos?

Popravil: zbrisal, del pa prestavil v splošni opis

1. Če še nisem omenil; ko navajaš eksplicitno avtorje, jih daš ven iz oklepaja. Namesto »To dejstvo predstavljajo avtorji (Homola in Kuboˇn, 2008a)« naj bo »To dejstvo predstavljajo Homola in Kuboˇn (2008a)« To je treba povsod popravit.

popravil citep v citet\*

1. »z uporabo besednih razredov.« kaj so b.r.?

Poljubni ekvivalenčni razredi, na novo sem jih opisal v pregledu področja

1. »povezave med fraza,«

popravil sklon

1. »lingvistiˇcnim znanjem« prosim odloči se ali uporabljaš termine iz latinščine ali iz slovenščine (jezikoslovni). Eno ali drugo je ok, samo bodi konsistenten.

Popravil: slovensko

1. »5.3.2 Opis metode« je eden bolj zanimivih delov, pa je spet tako razložen, da ni jasno zakaj gre. Daj prosim kakšen primer, za začetek ATja, potem pa omejitev; po možnosti poenostavljen v didaktične namene.

Dodal slikice in na novo napisal

1. »Generirana koda izvede najbolj pogosto AT, ki« najbolj pogosto po kakšen kriteriju?

Popravil: Generirana koda izvede AT, ki zadošča omejitvam ciljnega jezika za izbran primer in ki je bila sestavljena iz največjega števila poravnav v učnem korpusu, narejena je bila iz največjega števila učnih primerov.

1. Prvi odstavek na strani 52 je spet precej pomemben, pa meni precej nerazumljivo napisan; privošči si malo več prostora in ilustriraj s primeri.

Popolnoma na novo napisal, s primeri

1. »semantiˇcno enostavno veže na besede v ciljnem« kaj to pomeni, da se »veže«? A propos, kaj pa če je beseda dvoumna, kako potem prevedeš v pravi pomen? Je to kje pojasnjeno?

Leksikalna podobnost\index{podobnost} sorodnih jezikov, obširneje predstavljena v Razdelku \ref{section:leksikalnapodobnost}, zagotavlja, da se večina besed pomensko enostavno prevaja na besede v ciljnem jeziku po pravilu "ena-na-ena", torej ena beseda izvornega jezika se prevede v eno besedo v ciljnem jeziku. Tako je izdelava slovarjev enostavna in napake so redke. Kljub temu pa obstajajo izjeme, ki jih poskušamo opisati s pomočjo razširitve dvojezičnih slovarjev z morfološkimi podatki ter s pomočjo pravil. \\

Dvoumnost: del dvoumnosti odpravljam že z nemogočimi kombinacijami v izvornem jeziku (pravila lokalnega ujemanja), del spremenim s pravili, tiste, ki ostanejo pa naj bi „pravilo“ ovrednostil ranker.

1. 5.3.5 Primeri uporabe pravil: zakaj šele tu končno vidimo uporabo pravila? To bi moralo priti najprej, skupaj z razlago, da si človek sploh lahko kaj predstavlja, torej tam okoli 5.2.

nov subsection: Uporaba pravil za strukturni transfer

1. Tudi Slika 5.4 mi je nerazumljiva, in izgleda nepotrebno komplicirana in nedosledna. Zakaj enostavno ne uporabljaš multext-east imena lastnosti, torej Adjective namesto a\_adjec, gender namesto gen? Kar je »nbr«? Kaj je »lemh«? Poleg tega so vse v angleščini, naenkrat pa »sklon« v slovenščini. Prosim poenostavi, in (povsod!) dosledno uporabljaj MULTEXT-East specifikacije za poimenovanje atributov in lastnosti, saj bo tako vse bolj enostavno in pregledno.

spremenil sem jih kar v polne opise v slovenščini, btw nbr je number

1. Uh, že stota pripomba..

:(

1. »Pridevnik in samostalnik na Primeru 5.1« pa saj primer ima sam. in glagol!?

Popravil: ni na primeru 5.1 ampak na pravilu na Sliki fig:praviloleksikalna1

1. »long-distance relations« dependencies

popravil

1. 6. 6.1 6.1.1. 6.1.1.1. – Kaj če bi tu dal kaj uvodnih pojasnil in motivacij. Za MT imaš cele strani nekih uvodov, celo po večkrat isto, tu pa čisto nič. Saj se da kaj povedat; zakaj je pomembna, kako delimo metode, splošne prednosti, slabosti ročnih/avto, pregled pomembne literature.

Popravil, zdaj je poglavje še enkrat daljše, po opisu: naredi splošen pregled evalvacije in kaj je evalvacija, poglej po starih člankih

dodal še pregled ročnih metod, čeprav jih ne uporabljam

1. »preciznosti (precision)« valjda obstaja boljša slovenska beseda!

AI na FRI uporabljajo preciznost, vsaj Janez Demšar in tudi Branko Kavšek, pri nas

1. »ki za razred predstavlja število« kakšen razred?

popravil

1. »Sprememba z osnovno preciznostjo naj bi poskrbela za lastnost sistemov strojnega prevajanja, ki težijo k daljšim prevodom.« Tega sploh ne razumem.

Popravil, upam, da je zdaj boljše.

1. Metrika METEOR: glede na to, da jo potem tudi uporabljaš jo bolj natančno predstavi.

uredil

1. Slika 6.1: a se res ne moreš kakšnega bolj realističnega primera spomnit?

Spremenil, čeprav je v osnovnem članku ravno tak primer

1. Enačba 6.2 izgleda, kot da računa samo za eno besedo. Kaj pa ostali dve? To moraš bolje razložit.

m je število besed kandidata za prevode, ki so v referenčnih povedih

1. »predstavljene v tem prispevku v Razdelku 4.4 se« briši »v tem prispevku«

popravil

1. »angleški jezik je bila izbran kot linearni« kaj je to linerani?

najdi boljši izraz, že od začetka mi ni bil všeč, uporabil ga je Homola v članku

izolativni jezik, jezik z malo pregibanji

1. »Sistem je izdelala ena oseba v dveh dneh« koliko je pa trajalo za –et, -en?

Tudi, sem zapisal:

Za vsak predstavljen sistem smo porabili enako časa in sicer je sisteme izdelala ena oseba z uporabo običajnega osebnega računalnika \footnote{prenosni računalnik z 2 GB RAM in procesorjem Intel Core2 duo.)} in za vsak sistem porabila po dva delovna dneva.

1. »Metrika BLEU (Papineni et al., 2001) je najbolj razširjena metrika za evalvacijo sistemov strojnega prevajanja, vendar je nismo uporabili, saj mnogi« to si že povedal z istimi besedami par strani nazaj.

Opis sem prestavil v razdelek z opisom BLEU, tukaj sem le zapisal, da je nismo uporabili.

1. »Porter-stem« stemmer. Ali si implementiral stemmer tudi za slovenščino!? To bolje razloži.

Stemmer sem opisal v posebnem razdelku in povedaj, kje ga uporabljam

1. »petkratno preˇcno preverjanje … saj gradnja sistema za prevajanje ni popolnoma avtomatizirana« kaj ima to vezo? Razloži.

V našem primeru je bilo uporabljeno petkratno prečno preverjanje namesto pogosteje uporabljenega desetkratnega prečnega preverjanja, saj gradnja sistema za prevajanje ni popolnoma avtomatizirana, desetkratno prečno preverjanje pa v našem primeru zahteva postavitev desetih prevajalnih sistemov.

1. Polni testni set skupaj z ocenami priložil na CD-ju. (za pregled komisiji pa pač .zip)

JERNEJ ne pozabi!

1. »Cohenov kapa koeficient« Razloži kaj to je in kako dela. Tudi 6.1 bi si zaslužila kakšno razlago.

Popravi in opisal ter dodal razdelek, ki opisuje Cohenov kapa koeficient

1. Pogl. 7: tu na začetku nekje povej, kaj boš naredil, kaj poglavje vsebuje, in kako se navezuje na to kar je bilo narejeno prej.

Dodal opis zakaj je ta metoda potrebna

1. 7.1 Osnove: naslov poglavja bi bil lahko bolj informativen, not pa da bolj eksplicitno poveš v čem si spremenil metodo, zakaj, kako.

popravil

1. »POS oznake« slovensko »oznake POS«, pa v resnici PoS, pa še to ne vem, ali niso MSD? Kot že večkrat rečeno, poenoti terminologijo, nekje na začetku dr. jo pa uvedi.

Poravnave se vršijo med deli nizov oznak besednih vrst, kar v bistvu je PoS, vendar se ga zdaj načrtno izogibam, v pregledupodročja pa to tudi izpostavim

1. »z brisanjem besed.« to mi je malo štorasto rečeno. Pač ohraniš samo MSDje, raje kot da brišeš besede (in leme, pa mogoče če kaj, odvisno kako je korpus označen).

Popravil

1. »(z abstrahiranjem besed).« ? saj jih ne abstrahiraš, pač pa zanemariš. Bolje to kar brisat.

popravil

1. »niz POS oznak, izdelan iz izvorne oznaˇcene povedi je poravnan z drevesom izpeljave v ciljnem jeziku. Predstavljena metoda uporablja poravnave med POS oznake izvorne povedi ter drevesom izpeljave v ciljnem jeziku.« Ni to 2x povedano isto?

Ja, popravil, brisal sem poved

1. »dreves izpeljave (treebank)« za treebank se je že lepo uveljavila drevesnica; pa tudi, skladenjsko označen korpus ne vsebuje dreves izpeljave (izpeljave česa?) pač pa skladenjska drevesa (če res so drevesa, kar pri odvisnostni slovenici niso..)

za uporabo slovenske odvisnostne slovnice, bi morali najprej predelati zapis v skladenjska drevesa npr. PENN treebank

1. »V prvo skupino sodi le pešˇcica najveˇcjih svetovnih jezikov« to pa že dolgo ni več res. Npr. češčina ni največji svetovni jezik, pa ima super drevesnice. Konec koncev je tudi jos100k drevesnica. In bi bilo v resnici zelo zanimivo, da bi ga uporabil.

Prvi del sem popravil, drugi del pa se bom potrudil in naredil en sistem (enkart), zdaj pa bi to izpustil

1. »drevo izpeljave z izraˇcunano stopnjo zaupanja (confidence score). Primer drevesa izpeljave je predstavljen na Sliki 7.1.« Kje so pa confidence scores v 7.1.?

nova slika

1. »Ciljna poved je analizirana z analizatorjem (Collins, 2003), ki je bil nauˇcen na skladenjsko oznaˇcenem korpusu (Marcus et al., 1993);« Če bi bil malo manj kriptičen, da npr. poveš da gre za angleščino in imenuješ analizatorje in korpuse z imenon, referenco pa navedeš samo prvič, bi ti bil bralec zelo hvaležen.

Popravil Penn treebank, Charniakov razčlenjevalnik

1. O (Collins, 2003) bi bilo tudi dobro še kaj povedat, če ga že uporabljaš. Npr. ali res vedno vrne prava drevesa – kaj naredi s praznimi kategorijami (traces)?
2. »Opisano zaporedje pare oblike razvidne iz primera« ???

Gledav v tla, ta del sem popolnoma na novo napisal, nisem sploh končal poglavja

1. »v, po možnosti pravilno, poved« je to res treba izpostavljat?

brisal

1. »Algoritem poišˇce niz oznak POS v uˇcnih podatkih« glede na kaj? Verjetno izvorni niz, samo dobro bi bilo povedat. Kakšne dolžine so ti nizi? Za celo poved, ali kaj manj?

V tej fazi se odvija prevod\index{prevod} vhodne izvorne povedi v poved v ciljnem jeziku. Vhodna poved, poved, ki jo prevajamo, je oblikoskladenjsko označena. Iz teh oznak izberemo le besedne vrste in sestavimo iskalni niz. Algoritem poišče niz oznak besednih vrst v izvornem delu učnih podatkov (za vsako poved). Rezultati so ocenjeni v skladu z uporabljeno iskalno metodo:

1. »Rezultati so ocenjeni v skladu z uporabljeno iskalno metodo (metode so toˇckovane)« Kakšne so te metode?

\begin{itemize}

\item \emph{točno ujemanje;}; popolno ujemanje izvornega niza s ciljnim nizom. Ta metoda je ocenjena brez penalizacije.

\item \emph{ujemanje podobnih nizov}; izvorni in ciljni niz se lahko razlikujeta le za vnaprej določeno uteženo Levenshteinovo razdaljo \citep{levenshtein1965}. Penalizacija je odvisna od utežene Levenshteinove razdalje.

\item \emph{ujemanje podnizov}; ujemanje izvornega nizi z večimi nizi ciljnega korpusa oziroma ujemanje izvornega dela s podnizom ciljnega niza. Penalizacija je odvisna od števila podnizov in je bila določena le na osnovi majhnega empiričnega testiranja.

\end{itemize}

1. Ostanek tega poglavja ne razumem preveč – predvsem pa, zadeve se kar naenkrat abruptno konča, sem mislil, da boš naredil eksperiment in evalviral rezultate, tako pa vse obvisu v zraku, in sploh ni jasno kaj si naredil. To bo očitno treba še pošteno dopolnit!

ja, evalvacijo sem dodal v poglavje evalvacija

1. »8.1 Prispevki k znanosti« To mi je v resnici čuden naslov poglavja, pričakoval bi da najprej povzameš disertacijo, in kje vmes poveš kaj so prispevki k znanosti.

popravil

1. »v naslednjih glavnih publikacijah (Viˇciˇc in Brodnik, 2008; Homola et al., 2009; Mikoliˇc et al., 2009; Viˇciˇc, 2009)« Zakaj pa te publikacije niso bile (vsaj ne vse) omenjene tudi v doktoratu?

poglej ce katera ni bila

Zdaj upam, da so vse

1. Pa, precej važna stvar: ali ni na FRI pogoj za dr. da imaš članek v IF reviji? Jaz nič takega ne vidim v referencah.

Kononenko je napisal, da za moj doktorat ni nujna, ker sem se sprehodil prek magisterija. Če bo slučajno nujna, bom čakal na objavo.

1. »Metoda, za prevajanje s pomoˇcjo dreves izpeljave« samo v doktoratu o tem nimaš dovolj napisanega; zakaj enostavno ne povzameš po že objavljenih člankih?

Enostavno sem pozabil na ta razdelek, zdaj je dodan, evalvacija pa je premaklnjena v poglavje Evalvacija

1. »Ocenjevanje pravil za strukturni« ta goli »strukturni« se v dr. ne vem kolikokrat pojavi. A si delal kak neumen poišči zamenjaj?

Ja, transfer sem označil z \index{transfer} namesto transfer\index{transfer}

sem popravil

1. »Izdelani so napotki za hitro izdelavo podobnih sistemov z drugimi jezikovnimi pari« Kje so ti napotki – jaz jih nisem našel.

JERNEJNOVO povzami metode v napotke

1. »razdelitev paradigme strojnega« ne uporabljat paradigma, ker zmedeš.

popravil

1. »osnovo ˇclanka« kakšnega članka?

V ospredju so metode, ki slonijo na pravilih\index{pravilo}, to delo se največ posveča prevajalnim sistemom te skupine.

1. Te umestitve ne razumem, bo treba bolje razložit kaj s tem sploh hočeš povedat. Tudi se mi nekako zdi, da bi taka slika bolj pasala na začetek, ne pa na konec disertacije.

Prestavil v pregled področja, na konec, saj šele tu uvedem vse pojme

1. 8.1.2 itd. Zakaj sedaj spet svoja podpoglavja za vse to? Posebej če imajo samo po en odstavek. Kot rečeno, »obširneje predstavljena v Poglavju 7« ravno ni. Tam še precej manjka.

popravil

1. »Število sklanjatvenih vzorcev v naravnih jezikih je relativno majhno in pravila lahko roˇcno vnesemo« To pa ne drži – če vzameš vse premene kot svoje vzorce jih je za slovenščino sto!

Popolnoma spremenil, tukaj enostavno zbrisal

1. »ki jih ta metoda samo oznaˇci, prepozna jih pravilno kot izjeme« slovenščina..

popolnoma spremenil

1. »Oglejmo si primer za slovenski jezik, na primer slovenski jezik vsebuje po 4 osnovne sklanjatvene vzorce za vsak spol, 6 sklonov ter tri števila, skupaj 4\*6\*3 = 72 enostavnih pravil, ki le zamenjujejo konˇcnice« ne drži – to so res samo osnovne paradigme, upoštevat moraš še vse možne premene končnic in osnov, njihove kombinacije, in izjeme, situacije kjer ima ena beseda več osnov (človek / ljudje; iti / grem), itd.

zbrisal, sem delno opisal v drugem poglevju

JERNEJNOVO TO POSEBEJ OBDELAJ: Tu tudi ni mesto, da o tem diskutiraš, to bi moral v relevantnem poglavju, in tam bolj na dolgo, ker delaš precej poenostavitev – kar sicer lahko, samo moraš priznat, da jih in argumentirat zakaj. Posebej glede na to , da obstaja CLOG, ki čisto dobro lematizira slovenščino, pa tudi generirat besedne oblike iz leme zna.

1. »Besedi s pomoˇcjo krnenja« Čudno povedano, in ni jasno kaj s tem hočeš - pa spet, to itak nima kaj iskati v zaključkih, pač pa v dotičnem poglavju.

premisli kako bos spremenil poglavje

spremenil

1. 8.1.4 »Za preizkus metode je bil uporabljen korpus« to se nanaša na »4.4.1.1 Izdelava paradigem« Tam ni nobene evalvacije, čeprav bi jo lahko naredil. Tisto poglavje je itak precej na kratko in nejasno napisano.

premisli kako bos spremenil poglavje

dodal evalvacijo in spremenil poglavje

1. »Naloga raziskave ocenjevanja pravil strukturnega transferja je razdeljena na tri podroˇcja:« tu nekako še enkrat pišeš kazalo doktorata, kar je nepotrebno. Koncentriraj je na važne stvari in v prozi opiši kaj je novega.

premisli kako bos spremenil poglavje

Sem spremenil

1. »Raziskava možnih uporab ocenjevanja pravil« to verjetno paše v nadaljnje delo, čeprav v resnici sploh ne razumem kaj tu hočeš povedat.

Prestavil v opis področja

1. »V okviru doktorskega dela bo izdelana posebna metrika« No, samo da bo.

popravil

1. »Pri snovanju ter postavljanju sistema so bili izdelani napotki za hitro izdelavo podobnih sistemov z drugimi jezikovnimi pari.« To si že povedal, pa, kolikor jaz vidim, še res ni. Kje točno je to opisano?

JERNEJNOVO ???

1. »8.2 Prevajalni sistem GUAT« To bi bilo treba predsatvit prej, in bolj podobno, saj je ni lahko bil eden oz prispevkov doktorata, tako pa GUAT srečamo šele v zaključkih. Ali ni GUAT tudi na spletu?

Prestavil v pregled področja

1. »Poglavje A«?

popravil, stil je poznal le poglavja, jaz pa nisem pogleda

1. Teh pravil nisem gledal, ker, kot že omenjeno, pričakujem boljšo razlago formata, in boljše labele.

Uredil, „poslovenil“ sem oznake in opisal format

1. A.2 Primeri prevodov: to očitno še vse manjka.

uredil

1. Literatura: šlampasto stavljena, npr. »the case for english-swedish«, različno veliki fonti, manjka mesto objave itd. itd.

Sprehodil sem se skozi vse zapise v .bib. Stil uporablja tri oblike: osnovna, kurzivna za naslove ter Courier za url. Mesto objave: na primer Ljubljana? Ali ime publikacije? Imena publikacij bi morala biti pri vsakem zapisu, mesta pa nimam praktično pri nobenem zapisu, pa večino zapisov sem pobral že v bibtex obliki.

1. Zakaj ni jos nikjer omenjen, saj ima npr. drevesnico; velik tagiran korpus (10x večji kot 1984) in spletni lematizator.

Zdaj je opisan