Exploring Norma for Historical Text Normalization

Tova Erbén • Project for Language Technology Resources •

lik amore made liax plui-ar han lagha' forfull'

alber open ber er briggig manha fic berabit oper blit oper bei er bitte oza fak kar a bittopak fikm a baind fir balka ljah acha dihoghor oc e fizal kunikungarbak gitter vann bah vinter or foman karn boef kriftin film finds foren fall varn arthug an man hulla boef k oling okaman hull balkal for ra man hulla an ola f

Y eigh hull atter

Modertracko a Laghman 18

INTRODUCTION

At han faker ar mard trim vid bling or bit marchit vid faklo den en forfal prefe aru an bifuper hauer han fullat ligher bridia an han bigainda maiffu figher hava un han air toka farin linkum ar hialpa ar han fivugar and har fara ok fivku halpa an han fivugar and bus ipanda maiffu fighal fain a bouda ok per bal lan bendi han forfalla of in ha fial han fall Dreff feat bonda' of a oc bonda kono boda fon or who have all tabarn fire winton fin an fire mona hans huast et olia feal guil dia tva oh foa firi alla ba man eigh gora grunda fin ok Muni manu Fe hu aku ligh mara bit muplt uil-aner · hage-lik min bill gun dirogh naar the orromun af fa appershad or o'rogh ougher gelf ar hulla I pa a pit pik ok lacy wair bilaper innan lohna u. bibar ola lik ha ar han lhi ar ola haur half mark firm Can wan armars off ber-t-bols ok fol

alberropen per cr-programanha fac Text normalization

....s about mapping spelling variants (or misspellings) to a standard form

historical texts tend to contain a lot of variant spellings

freigh buil auer ofeng ar piele torianalos pa

a'r han faker ar march frin vif l marchil vid fakforhenden fozfal p bufruper hanur hanu bult faint oc for Amnur an han furbat ligher J bigiainda marllu fighur foada' an l farm finknin ac hialpa' yr lbal n fara o'r fivku hialpa' an han fivn i sanda marllu Sofull fain a boud lan bondi han forfallale I. n- pa i per theire
pere theiare
theyr their theyr
thir thear
thaire
thaire
thaire

han lengar lara mm liggia giui dirogh maar

t on ar bonda ralei oris mun af fa or giui pilu firi laphershad or orisogh epast sir skrjougher gest ar husla edo stafkad ha a pist pik ola siren ersar busuper muan soknar ut hanu bibar ola sik ha ar han skrj n ar ola haur halt mark firin san an annare sist berer bola ok stol

(Bollmann 2019)

alben'apen ber en lunggia maniha fac deur aluf open luggen bridningen upen hat et fida 1920' fak kit a bifupak likan a bahad fin balka liki uch ekspeljer ab ologia kundamanki

- Lookup-based methods
 - Rule-based methods
 - Distance-based methods
- Statistical machine translation
- Meural machine translation

for Annur an han liukat ligher Fridia' dir han bigianda martii lightir figida' an han ar i loka

farm infance ar hatpa to bel mailtraining fara ok fisku hidlpa: ein han fisingar abok hig hidrada mailti. Sigkul hain a bonda ok fie kel lak bendi han fosfallalotar ha iral han sarrar lik a more made liax plui-ar han lagha forfall'

faction of the hand bibat old like ha ar han the longher han ar old hand half mark firm Can gor fit to lot annave fift bere bole ok trot

alber open per cu-progra marcha fac bera bit mper ligger briffinger vuen bar er fing oza Tak kat a bistopat sikan a hajad fir balka Hali neta demochor De oftal fentelengarbaix gitter vary babi vinter of fomay Barn box for ilinu. L'huls form feat barn arthug an man hulla box to oling oh anna hull ba hal for

Wastra So a Laghman 13

ar han faker ar mardy brim vib bilaup or bit marchit vid fakfortenten forfal prette aru an befruper hance hand but faint or ar box fair for Amur an han linker hoher bridia an han bigianda ma flu fighte fraca an han ar e-folia farm finkmn ac halpa fara oh fivku hedha! en han fivngar arch ho hundamallu. Sahiltain a bond lak bendi han forfallalo fan ha iral har

lik amore made liax plui-ar han lagha' forfull' fack-Arest-feat bonda of ola oc bonda kono. boda fon or who have all tabarn fire winton fin an fire mona hans huast et olia feal guil dia tva oh foa firi alla ba man eigh gora grunda fin ok on fin langer Cal alli fiallimi manil fe hu le tam ok orweh fire natuaku ligh mara bit aki hon ob flera vaka nu pli mili erler na man hulla anola barn fav eigt via I E A unit och ar her nacar liggt-lik mini-bill keigh hull after ofeng ar prett far to grand macer gun dirogh naar aki o'iro mun af fa appershad or o'rogh Jugher gelf ar hulla I pa a pit pik ok lacy mok laghar flad Swair bilauper innan lokna? ar bonde but hand bitar ola like ha ar han the ar ola ham half mark firm Can wan annave off ber to bole ok that



aper oper profittinger upon par er fing of MADIR Trees on or who have allarban fire gunda fin an

- gitari na Abota er og soman Barn boz fizr
- ilmu & huls foracal 1448 (Ima en man
- hulla box 140 m540 ntokens a faul for
- ra'm2. hrAvglola barn far eigh kriffing
 - Elder Westrogothic law
- er han faker ar man 1220/1280
- biling haur 1228 tokens
- 601 3. Moses-blocker bridge an han
- Paraphrases of the five
- fara ok fivkú hrálpa! an han fivngar arch h hranda ma fir Sokul pain a bonda ok pob

- ca 1330/1526 gen gunda imok
- 7049 tokens
- 4. Ogl-a aku hogh ole flera valen mi pli mil atter
 - Ostrogothic law
 - ca 1290/1350
 - 22090 tokens
- 5. Tungulus nall ar flestougher gelt ar hulla
 - 1457 or harbart pa a filt pik ole freque
 - 2814 tokens of file of the han fix

alber open ber cir briggia mariha lac bera MADIR Words

- Ordbok över svenska medeltidsspråket der hans affatsan för gjunda findall hans og det gjunda findall hans blade er old fra gjunda red
 - Lemma, pseudolemma, parenthetical form,
 POS (coarse), POS (finegrained)



ì	0	0	00	interj	interj
ı	ofnär	ofnär	ooffhne	er	prep prep
	ok	ok	aag	konj	konj
	ok	ok	ос	konj	konj
	ok	ok	och	konj	konj
	ok	ok	ock	konj	konj
	okar	okar	okkar	pron	pron poss
	okar	okar	ukar	pron	pron poss
	ovan	ovan	affwan	prep	prep
	ovan	ovan	afwan	prep	prep
	ovan	ovan	offwon	prep	prep
	ovan	ovan	ofwen	prep	prep
	ovan	ovan	vuan	prep	prep

alber open ber er briggig manha fic berabit oper blit oper het er fint ozar fakt kir a bringer hen bar er fint ozar fakt kir a bringer ber balka ljah acta ozhoghor de e fizat kurrkingarbak giter vank bah vinter er fomar barn bozf tariftur fints forra fat varn cultug an man hulla bozf fe oling oktamak hull ba ikal forra man hulla bozf fe oling oktamak hull ba ikal forra man hulla bozf fe oling oktamak hull ba ikal forra man hulla bozf fe oling oktamak hull ba ikal forra man hulla bozf fer oling oktamak hull ba ikal forra man hulla bozf fe oling oktamak hull ba ikal forran man hulla bozf fe oling oktamak hull ba ikal forran man hulla bozf fe oling oktamak hull ba ikal forran man hulla bozf fe oling oktamak hull ba ikal forran man hulla barn fe oling olin

Modertracko a Laghman 18

na man hutla ar ola barn far SOFTWARE

At han faker ar nard frin vid bilup or bit marchi vid laklo denten forfal pielle arn an biluper hant han linkat ligher britia an han biganda mallit light. Hada in han ar i foka farin linkum ar halpa or balma littamin fara ok livku halpa an han livngar and ha harda mallit. Soful fain a bonda ok por bal lan bendi han forfalla lotar ha iral han

lik amore made liax plui-ar han lagha' forfull' fack-Arest feat bonda of ola of bonda kono boda fon or who have all tabarn fire winton fin an fire mona hans huast et olia feal guil dia tva oh foa firi alla ba man eigh gora grunda fin ok on fin langer Cal alli fiallimi manil fe hu le tam ok orweh fire natuaku ligh mara bit Hera vakanü pit uili aller · hage-lik min bill gru cirogh naac aki o'iro mun af fa appershad or o'rogh ougher gelf ar hulla I pa a pit pik ok lacy mair bilaper innan lohna ni bibar ola lik ha ar han lhi ar ola ham half mark firm Can wan annave off ber to bole ok that

alben'open-ber ei Arnggia maniga die ber albu copen-liggen bridungen voon haven film oza fals-kie-a-victopas-likan a-bain'o tim balka ljäh atta ozibogijos-de e-fizal kuurkinganhais sitten vasik baki vuiter ee lomas kann besekter

Wartrago a Laghman 18

Norma

- Normalization tool (written in C++)
- Lookup-based, rule-based, distance-based (WLD)
- Isolated or in combination
- narchi vi Require target lexicon
 ufuper haur hand but fant och berfan
 vi Ammur an han furfat ligher bridia an han
 imanna martu fictur fracot in han eir i folia

fara ok fivku hadha: an han fivugar atok hu handa madhi. Salul hain a bonda ok fir Lal lak bendi han forfallalofan ha iral han mara lik amore made liax ploi-ar han lagha forfull'

fon or who have allalvan firi hunda fin a firi hona have huare et olia feal giut ora eve ah foa firi alla bamen eich gota evinda fin o

s ram ok oxogh firi natuaku. Ugh mam þ s m vaku hog ok flein vaka nú filt nili-ei

han lengar lara mu liggia gun o'rrogh naa huaria. Gall on ar bonda take o're mun af k tum hane or gun glu fin lig lagherstad or o'rrogh ar narvaku-prast eir skydugher gest ar husla siim bonda: Dod stafkad ha a gist pik vik ok sney

tar bonor but hann breat old the 1st ar han the lougher han ar old hand half mark firm ban ger fit-t loten annave fift bere bole ok trot

Docker



- Platform as a Service (PaaS)
- Applications in an isolated environment known as container
- Containers contain all necessary dependencies and configurations
- Docker has a public repository for such containers
- Run the container of an application via a Docker image

alber open ber er briggig marcha far berabit oper bligger bridinger inen bater fint oza fak-kit a bilopak-litan a band fin balka litan acho find balka litan acho find barn bor karibak giter vark bapi vinter or foman Barn bor kriftin. Khuls forta fait barn uilina an man hulla box Koling okaman hulla ba kal fortan man hulla box Koling okaman hulla ba kal fortan man hulla an man balkal fortan man hulla an man barn far ugh k

Modertracko a Laghman 12

ne man huller an ola barn far eigh k W ETHO milioc pri ar piere feigh hull either oleng ar piere feifall W ETHO milioc pri ar piere

ar han faker ar march rum ord beliup or bri marchi vid lakforden on forfal preto are an beliuper haure han linkat ligher bridge and ha biganda marlie ligher brown on han ar rocke farm linkam ar halpa brillal marlie tum tara ok livka halpa: an han livngar and ha iserda marlie Social bain a bonda ok pretotal lan bendi han forfallal of ar ba fall han

bendi han forfallalulan ha iral han wind will

lik a more made liax pilni-ar han lagha forfull' fack-Arest-feat bonda of ola oc bonda kono. boda fon or who have all tabarn fire winton fin an fire mona hans huast et olia feal guil dia tva oh foa firi alla ba man eigh gora grunda fin ok on fin langer Cal alli fiallimi manil fe hu le tam ok orweh fire natuaku ligh mara bit vaku high ob flera valia mi plt mit erler navar liggi-lik min bill gun dirogh naar aki o'iro mun af fa appershad or o'rogh Jugher gelf ar hulla I pa a piff pik ok facy mok laghar flad Swair bilauper innan lokna? as bonde but hand bitai ola file pa ar han thi ar ola ham half mark firm Can wan annave off ber to bole ok that

alber open ber or briggia marcha facter out processing I

1. mathir tokens txt

Wastracko a Laghman 3

gilter varre bahr vinter of formar harn boof kr ilmu k huls forma fail barn all man man hulla book k oling ok anna Sigx kal for

m man hulla anola bari haffi kulliu Pengh hull alla oleng ar abotum

de han feker ar march from von bilath or bri marchii vid faktochenten fo allum o den en bilather hane hani but faine or de borran 2. mathirmtrain txtumorimok

on fine hærtal blut här mannige but signing si

The state of the s

lak bendi han forfallalulan ba ical han nama a de ger pit-t loku annam pit ber t bok ok in

Preprocessing II fak kar a biliopat ditan a harab fin balka dw_train.txt i hona hane huac er olia led giut ora roa lach lagh lagha lagh laghä lagh lagh laugh lagh

Maitra 80 a Laghman 18

Extracting variant spellings ('parenthetical form') and spellings ('parenthetical form') and spellings ('parenthetical form') and spellings ('lemma') from MAPiR Words ('lemma') from M

alber open ber er briggig mardja fac ber abif wydr lligger bridiinger vpen bâret fi fir oza fak kit a bifwpat litan a baind firi balla fah atta dithoglor Oc e fizal luurkingarbak

Norma

Ex amon mak tracille ration tooks to fail?

fak frest feal bonda ola oc bonda kono boda fon oc order hans alstabarn for ginnda fin a for hona hans huar er olia seal gini dia roa ok soa fin alla da man eigh gesta thunda sin o

- Configuration: normalizers=Mapper
- "Training" data: sdw_train.txt, which produces a file called sdw_train.Mapper.mapfile
 - Input data: mathir_tokens.txt/wara Can arbonda mku one mmy af
- input not use of cut to far ment out for formalized tokens evaluate or group fin for larger flat or or or or of the normalized tokens evaluated in the large of the final and manufactured tokens evaluated in the property of the first pullation of the part of the property of the first pullation of the part of the property of the part of the p

and of the highest and han the near a see high heards mathe Antal fain a bonds of fire hal lak bendi han for fallalofan ha fral han ballar

tougher han ar ola haut half mark firm Can ger phi to loku annave phi ber to bok ok fol

lik amore made liax piniar han lagha forfall'

alborranon bor or brown martha fac berabil

operalugger brithunger ropen bit ét finz oza; falt kit a biltopak likan a haind fin balka liah atta ozdoorpoz De e feal faurkmoarbaix

other vary bale vinter or forman Baseline and Evaluation of that all if fire firm manifye hu

- Baseline script: tokens already normalized/identical to
- Evaluation script: number of input tokens normalized by Norma to correct lemma form as annotated in MAÞiR Trees

bilinper hanir hanil bub la'ne oc ar barfail binanna an han linkar ligher, bridia an han binanna ma llinlighir binaar an han ar i-foku

tarm hitannear hatpa ve teal marituratum fararok fisha hudha: an han fisingar atoh hug humda mariti. Ankil hain a bonda ok fit tal lan bondi han forfallalutar ha fral han saira ar narvaku-prati di (kybugher geli ar hulla filim bonba: Aph flathad ha a pli pik ok firey un ok lashar flath Strair bifuner innan fokua

fac bonde but hand bitat old like ha ar han the tougher han ar old have half mark firm ban ger pot took ok tol

alber open ber er briggig marga far bera bil oper bigger bridunger iven hat er fint oza fak kit a brind fint balka han deta dishog joz-de e deal kurikungarban gilter vang bahi vinter op foman barn hoef far ilmuk hulls fort fat barn artina an man hulla boef koling oka mas hull ba ikal for ta man hulla de kal for ta man hulla an man barn boef koling oka mas hull ba ikal for ta man hulla an ola barn far engh m

Modertracko a Laghman 12

me man huther an old barn favered to the ESULT be Sulf-or regel huth atter of eng ar prest for the land ESULT be sugar

ar han faker ar mard from ord before order marchi vid fake chonen forfal prete are an belieper hance han faut favor or ar box fair for Amour an han figher had a de han bigainda marffu figher. Had marffu figher had a de han fair i filkan ar fillpar an han five fair a bigainda marffu figher. Had marffu fairm fina ok fivku halpar an han five fair a box or fillpar an han five fair a box or fillpar fillpar an han five fair a box or fillpar fillpar

lik amore made liax plui-ar han lagha' forfull' fack-Arest-feat bonda of ola oc bonda kono. boda fon or who have all tabarn fire winton fin an fire mona hans huast et olia feal guil dia tva oh foa firi alla ba man eigh gora grunda fin ok on fin lather laballi fallini mani be hu le tam ok orweh fire natuaku ligh mara bit the hot or flera valed nu ple uit eller navar liggi-lik min bill gun dirogh naar aki o'iro mun af fa appershad or o'rogh Jugher gelf ar hulla I pa a piff pik ok facy mok laghar flad Swair bilauper innan lokna? ar bonde but hand bitar ola file pa ar han the ar ola ham half mark firm Can wan annave off ber to bole ok that

Normalizer

Results

Accuracy all a feel and manage has

gilter vary babi vinter oc
hulla box F. oling okam
Y eigh hutl-atter ofeng ar
de han faker ar mardi bri
marchit vid faklochenten
bilaper haur hand but f
for Amur an han finkat !
himanna marthe to here fine

Baseline (unchanged) Norma (lookup) Norma (WLD) Norma (rule-based) Norma (combined)

27.30 % 34.14 %

Huarin Bart wir ar bonda taki o'ris mun af fa or or will fir lapher lad or or ogh

At 19/3 rateu - pratt die flassouother gelt ar hulla

lan benot han tortaliate lan pa trai gant paren

Matches input output

Failed Matches input output correct

þän bätre wtan widh

alper open her er briggig manig far bera bit oper ligger bridinger open hat er fine oper fakt kir a biloppit ligen a baind fin balka ligh neta oxidogljor or e haal kuurkingarbak gilter vary bah vinter or fomas karn poet kriftnuk hils forra faal barn ariting an man hulla boek koling ok anna hull balkal for ra man hulla an ola karn t

Wastra 80 a Laghman Fo

ne man hulle an ola barn to DISCUSSION freigh hull after oleng as pur DISCUSSION

lik a more made liax pilni-ar han lagha forfull' fack-Arest feat bonda of ola of bonda kono boda fon or who have all tabarn fire winton fin an fire mona hans huast et olia feal guil dia tva oh foa firi alla ba man eigh gora grunda fin ok on fin langer Cal alli fiallimi manil fe hu lo tata-oh orogh fire natuaku-ligh mara bir era valanü pit uili aller · hage-lik min bill gun dirogh naar the orromun af fa appershad or o'rogh ougher gelf ar hulla I pa a pit pik ok lacy mair bilaper innan lohna u. bibar ola lik ha ar han lhi ar ola haur half mark firm Can wan armars off ber-t-bols ok fol

lak kir a biliopat likan a haind für balka Discussion

- Dictionary-based lookup: hard to handle inflected forms
- By-chance lucky matches like wordo -> varba
- Most successful matches: identical token-lemma pairs
- Some failed matches: improved by rule-based/WLD-based normalization?
- Would be interesting to run Norma in combined mode (cf. Bollmann 2019)

Blaco Balle

tian and orthographe of a Conclusion and Future Work man and com summer final

- A lookup-based method is only as good as its source material
- Docker is the way to go when using dependency-heavy LT applications
- Using Norma's other modes with target dictionary
- cSMTiser (character-based statistical machine translation)

gerwa madu. Ankil pam n bonon ok pa sad lak bendi han kefallalolan ba iral han wira

ger fit i foku annavs fit ber t bol ok fol

