

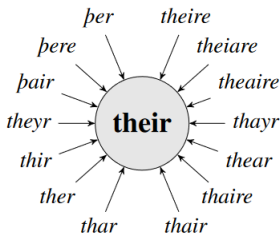
Exploring Norma for Historical Text Normalization

Tova Erbén • Project for Language Technology Resources •

Text normalization

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

...historical texts tend to contain a lot of variant spellings



(Bollmann 2019)

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

lik a more mardī siar pīstī. ar han lagha forfāll
fak prest skal bonda ola or bonda homo. bōda
son or docto hans all skabarn fir. giunda sin an
fir hiona hans huart er olia skal giu oia rva
ok soa fir alla ba man eigz geia giunda sin ok
ori fir lagher stad allū frallinn mianū. Ye hu
lo tam. ok orrogh fir natuakū. ligh mārā hie
flera vaki. ligh ok flera vaka nū pīst mli. eisse
bondi uli. or hī ar biek narar liggi. lik him. bill
h au lengar lata minn liggia giu orrogh naar
huaria. Galt dōr ar bonda taki oris mung af fa
tum hans. or giu pīstī fir lagherstad or orrogh
ar natuakū. prast ar skydugher gest ar hulla
sin bonda. Dōr skafkal. ha a pīst pik ola ferey
pu ok laghar stad. Sicar bisaupe imnan lokna
far bondi bue hanū. bibak ola lik. ha ar han sky
tougher han ar ola. han half mark firn. Gan
ger pīst. i lokn annars pīst ver. i bok ok stol

alþer' open þer er þriggia manna fac þer a biskop
 opir. **L**igger þridungir open þar er sin or
 sak. þar a biskopar sit in a haind sin balla
 hain acta orhoghor. **O**c e skal kunnungar þar
 gilder varir bair vinter oc somar þar þor þer
 istnu. **H**uls. þerra skal barn arstna an man
 hulla þor. **O**lm. ok annas hult. þa skal þer
 ra man hulla an ola. þar þar eigh þer
Eigh hult. atlei oleng. ar prest forþan þar
 ar han sakar ar mark þrin. vi þiscup oc þri
 marchu vid saksothenden. forþal prests aru an
 biskuper haur hanu bu þat oc ar þor þar
 for. **A**nnur an han siukar ligher. þridia an han
 biganda ma'ssufigher. þranda an han ar i solen
 faru siukma. ar hialpa. **O**c skal ma'ssufatun
 fara ok siuk hialpa. an han siungar aroþ þer
 henda ma'ssufatun. **S**kal þar a bonda ok þer
 lak bondi han forfalla lofan. þa skal han þar

lik a more madi. siar þsi. ar han lagha forfall
 sak. prest skal bonda ola oc bonda hono. bōda
 fon oc dothor hans allabarn fir. giunda sin an
 fir hiona hans huar er ola skal giu ora rva
 ok þa fir alla þa man eigh gera giunda sin ok
 ori fir lagher. skal allu siar siur mianu. þe hu
 lo tam. ok orogh fir natu. **L**igh mair þer
 þera vaku. ligh ok þera vaku nu þer mli. a'ler
 þer mli. oc þe ar þer natu. ligh. lik mli. **D**ull
 þe mli. ar lara mli ligha giu orogh naar
 huaria. **S**an þer ar bonda mli oris mli af fa
 tum hans. oc giu þsi fir lagherstad oc orogh
 ar natu. þar ar siar ligher. gest ar hulla
 sin bonda. **D**er siar. þa a þer pik ok siar
 þu ok lagherstad. **S**ur biskuper mli lokna
 þer bondi bu þar. þer ola sik. þa ar han siar
 ligher han ar ola. haur half mark. **S**an
 ger þer i. þan annars þer þer i. þer ok stol

MAPIR Trees

1. Abota

• ca 1448

• 540 tokens

2. Avg/

• Elder Westrogothic law

• ca 1220/1280

• 1228 tokens

3. Moses-b

• Paraphrases of the five books of Moses

4. Ogl-a

• ca 1330/1526

• 7049 tokens

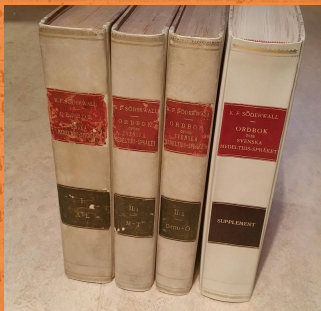
5. Tungulus

• 1457

• 2814 tokens

MABiR Words

- Ordbok över svenska medeltidsspråket (Söderwall 1891-1918)
- Lemma, pseudolemma, parenthetical form, POS (coarse), POS (finegrained)



o	o	oo	interj	interj	
ofnär	ofnär	ooffhneer	prep	prep	prep
ok	ok	aag	konj	konj	
ok	ok	oc	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	

十

- Normalization tool (written in C++)
- Lookup-based, rule-based, distance-based (WLD)
- Isolated or in combination
- Require target lexicon

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

alþer' open þer er þriggia manna fac þer a biskop
 opir. **L**igger þridungir open þar er sin 92a
 sak. þar a biskopar sit in a haind sin ballka
 hain acta oðhoghor. **O**c e skal kunnungar þar
 gilder varir bair vinter oc somar. þar þorð þer
 istnu. **H**uls. þerra skal þar arstna an man
 hulla. þorð. **O**lm. ok annas hult. þa skal þer
 ra man hulla an ola. þar þar eigh k. **E**
E eigh hult. atlei oleng. ar prest forfall
 ar han sakar ar mark þrin. viþ biskup oc þri
 marchu vid saksoþenden. forfall prests aru an
 biskuper haur hanu þuð saine oc ar þorð þar
 for. **A**mmur an han siukar ligher. þridia an han
 þiganda maðli fighir. þada an han ar i solen
 þarm siukma. ar þialpa. **O**c skal maðli farum
 fara ok siuk hialpa. an han siungar aroþ þu
 þanda maðli. **S**kal þain a bonda ok þu þal
 lak bondi han forfalla loðan. þa skal han þu

lik a more maðli siar þuð. ar han lagha forfall
 sak. prest skal bonda ola oc bonda hono. bōda
 þon oc dothor hans all þabarn þri. giunda sin an
 þri hona hans huar er ola skal giuð ora rva
 ok þu þri alla þa man eigh gera giunda sin ok
 ori þri lagher. skal allu siar þuð manu. þe hu
 lo tam. ok girogh þri natuðku. **L**igh maðli þu
 þera vaku. ligh ok þera vaku nu þuð mli. ælter
 þuð mli. oc þu ar þuð natuð. ligh. lik þuð. **D**ull
 þuð lengar lata þuð ligha giuð girogh naar
 huaria. **S**an þu ar bonda þuð oris mly af fa
 tum hans. oc giuð þuð þu lagherstad oc girogh
 ar natuðku. þu ar siar þuð. gest ar hulla
 þu bonda. **D**oð siar þuð. þa a þuð þu ok siar
 þu ok lagherstad. **S**uð biskuper innan lokna
 þu bondi þu þu. þuð ola siar. þa ar han siar
 þuð þu ar ola. þu half mark þu. **S**an
 ger þu i þu annars þuð þu i þu ok siar

Preprocessing I

1. mathir_tokens.txt

2. mathir_train.txt

hær

hær

här

sigx

sigx

sighia

aff

aff

af

abotum

abotum

abbote

allum

allum

alder

Script for extracting tokens and token-lemma pairs
from MAPiR Trees

- **Configuration:** normalizers=Mapper

- "Training" data: `sdw_train.txt`, which produces a file called `sdw_train.Mapper.mapfile`

- Input data: mathir_tokens.txt

- Output data: list of normalized tokens

Norma

lik a more mark. fies pſt ar han lagha forfall
fak. preſt ſcal bonda ola oc bonda hono. boda
fon oc doctoz hans a lſta barn fir. giunda ſin an
fir hona hans huac er ola ſcal giu ora rva
ok ſoa fir alla þa man eigh geira giunda ſin ok
ori fir lagher ſtat allu ſia lſiri manni. þe hu
pper
which produces a file
file
þa an tengar lara man liggia giu oitrogſ naac
huaria. Gaſt don ar bonda taku ois muſ af ſa
tun hans. oc giu pſt fir lagher ſtat oc oitrogſ
ar nat valku. praſt ar ſkydugher geſt ar huſla
ſum bonda. Doz ſtafhael. þa a pſt pik ola ſrey
pu ok lagha ſtat. Sita. biſcoper innan ſokna
fa bondi buſ hamu. bita ola ſik. þa ar han ſky
tougher han ar ola. hamu half mark firre. Gan
ger pſt. i. ſokn amars pſt ver. i. bok ok ſtol

Baseline and Evaluation

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

[illegible]

alþer' open þer er þriggia manna fac þer a biskop
 er. **L**igger þridungur open þar er sin 92a
 sak. þat a biskopar sit in a haind sin ballka
 hain acta oðhoghor. **O**c e skal kunnungar þar
 gilder varv þar vinter oc somar þar þorð þer
 istnu. **H**uls. þerra skal þar cristna en man
 hulla þorð. **O**lm. ok annar hult þa skal þer
 ra man hulla en ola þar þar eigh þer
Eigh hult. attel oðeng. ar prest forfalla
 ar han sakar ar mark þrin. vi þiscup oc þri
 marchu vid saksoðenden. forfall prests aru an
 biskuper haur hanu þuð salu oc ar þorð þar
 for. **A**mmur an han siukar ligher. þridia an han
 þiganda maðli figher. þada an han ar i solen
 þar siukma. ar þialpa. **O**c skal maðli faturu
 fara ok siuk hialpa. an han siungar aroð þer
 þanda maðli. **S**kal þain a bonda ok þat þat
 lak bondi han forfalla loðan þa ic al han þar

lik a more maðli siar þli. ar han lagha forfall
 sak. prest skal bonda ola oc bonda hono. bōda
 þon oc doðer hans all þabarn þri. giunda sin an
 þri hona hans huar er ola skal giuð ora rva
 ok þa þri alla þa man eigh gera giunda sin ok
 ori þri lagher. skal allu siar þli manu. þe hu
 lo tam. ok oðrogi þri natuðku. **E**igh maðli þer
 þer vaki. ligh ok þera vaki nu þat mli. a þer
 þer mli. oc þy ar þer natuð ligh. lik mli. þall
 þy siar þer lara mli ligha giuð oðrogi naar
 huaria. **S**an þer ar bonda mli oðri mli af þa
 tum hans. oc giuð þli þri lagherstad oc oðrogi
 ar natuðku. þar ar siar þer get ar hulla
 sin bonda. **O**c skal þat þa a þat þat ok siar
 þu ok lagherstad. **S**ar biskuper mli siar
 þat bondi þu þan. þat ola siar. þa ar han siar
 þer þan ar ola. þan half mark þin. **S**an
 ger þat i þan annars þat þer i þat ok siar

Matches

input	output
-------	--------

...

oc	ok
uili	vilia
scal	skula
med	mäp
þessi	pänne
engelin	ängeli
bönder	bonde
wordo	varpa
kyrkiu	kirkia
landbor	landboe
himmerike	himirike

...

Failed Matches

input	output	correct
-------	--------	---------

...

wt	wt	ut
ær	ær	vara
thet	tea	pän
æter	æter	äta
þa	piggia	þa
bætri	bætri	bätre
sigher	sigha	sighia
wigyæ	wigyæ	vighia
guðfæpur	guðfæpur	gupfapir
for budhit	for budhit	forbiupa
wtan widh	wtan widh	utan viper

...

alþer' open þer er þriggia manna fac þer a biskop
ligger þridungur open þar er sin or
fak þar a biskopar sit in a hand sin balla
han acta orhoghor. Or skal þuorkingabarn
gilder varu bafi vinter oc somar barn þorþ þer
istnu. þ huls þvra skal barn arstna an man
hulla þorþ. þ olmg ok annas hult þa skal þvra
ra man hulla an ola barn þer er þvra
þ eigh hult atlei olmg ar þvra þvra
ar han sakar ar mark þvra vif biskop oc þvra
marku vid sakor þenden forþal prests an
biskuper haur hanu þvra þvra oc ar þvra þvra
for. Annur an han siukar ligher þridia an han
þvra anda ma þvra fighr þvra an han ar i solen
þvra siukma ar þvra þvra. Or skal ma þvra þvra
þvra ok þvra þvra an han þvra aroþ þvra
þvra ma þvra. Skal þvra a bonda ok þvra þvra
þvra bondi han forþalla þvra þvra þvra þvra

DISCUSSION

lik a more ma þvra þvra ar han lagha forþall
þvra þvra skal bonda ola oc bonda hono. þvra
þvra oc þvra þvra an þvra þvra þvra an
þvra þvra þvra er ola skal gnu þvra þvra
ok þvra þvra þvra man eigh gera þvra þvra ok
ori þvra lagher skal allu þvra þvra man þvra þvra
þvra ok þvra þvra þvra natu þvra. Eigh ma þvra þvra
þvra þvra ok þvra þvra ma þvra þvra a þvra
þvra þvra ar þvra natu ligg þvra þvra þvra
þvra þvra þvra ma þvra liggia gnu þvra þvra
þvra. Skal þvra ar bonda þvra þvra mgy af þvra
þvra þvra oc gnu þvra þvra lagher stad oc þvra
ar natu þvra þvra ar þvra þvra gest ar þvra
þvra bonda. Skal þvra þvra a þvra þvra ok þvra
þvra ok lagher stad. Skal biskuper man þvra
þvra bondi þvra þvra þvra ola þvra þvra ar han þvra
þvra þvra ar ola þvra þvra mark þvra. Skal
ger þvra i þvra annars þvra þvra i þvra ok þvra

lak bendi han toftallato han ya iai han uai

- Would be interesting to run Norma in combined mode (cf. Bollmann 2019)

lik a more made siar pſti ar han lagha' fořfall'
fa'h- preß ſcal bonda' ola or bonda hono- bōda
fon or doctur hans al ſta barn fir- giunda ſin an
fir hiona hans huat er oia ſcal gni ora rva
oh ſoa fir alla ha ma'a ead geira giunda ſin ok
ori ni lagher ſkad ann natuna manū. Ye hu
do → varpa atuaſku. ligħ māra hie
flera vaka. ligħ ok flera vaka nū pſt mli- aller
al token-lemma pairs
y rule-based/WLD-based
a in combined mode (cf. buſla
ſim bonda'. Dor ſtafkaal- pa a pſt pik ola frey
pu ok lagha' ſkad. Sutar biſcupe imnan ſokna'
ſca bondi buř hamū bibat ola ſile- pa ar han ſki
louger han ar ola- hamū half mark- firm- ſan
eer ſit- i ſoku amars ſit der- i bok ok ſak

Conclusion and Future Work

- 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)

Ad laud. ⁊ ad bñd. vt in communi
vnus uirginis.

THANK YOU

