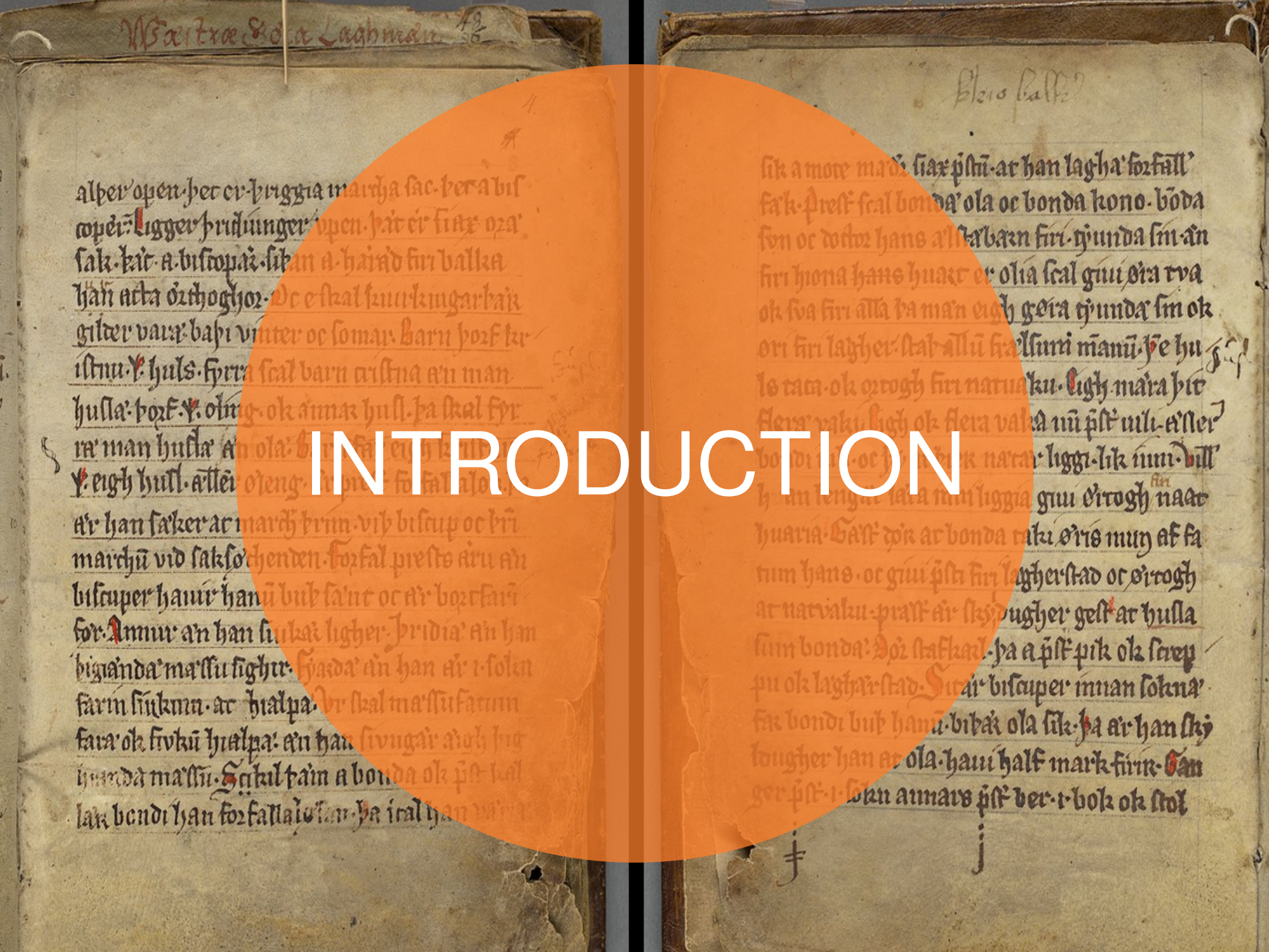
Exploring Norma for Historical Text Normalization

Tova Erbén • Project for Language Technology Resources •

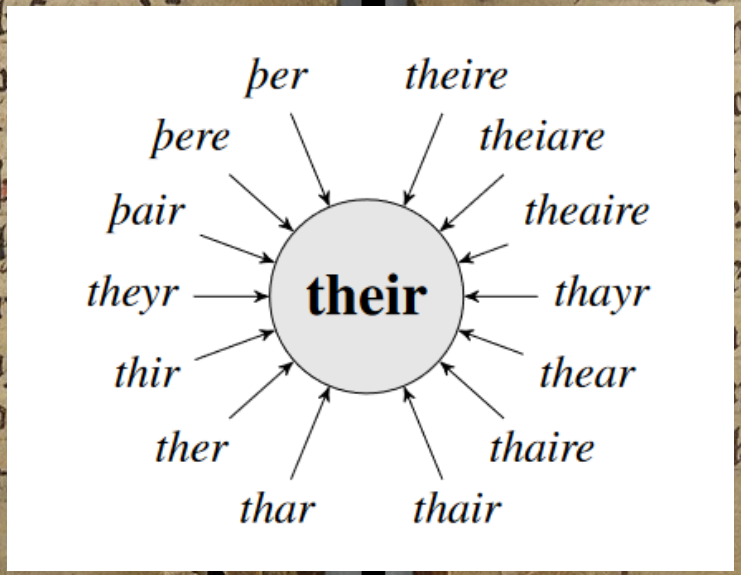


INTRODUCTION

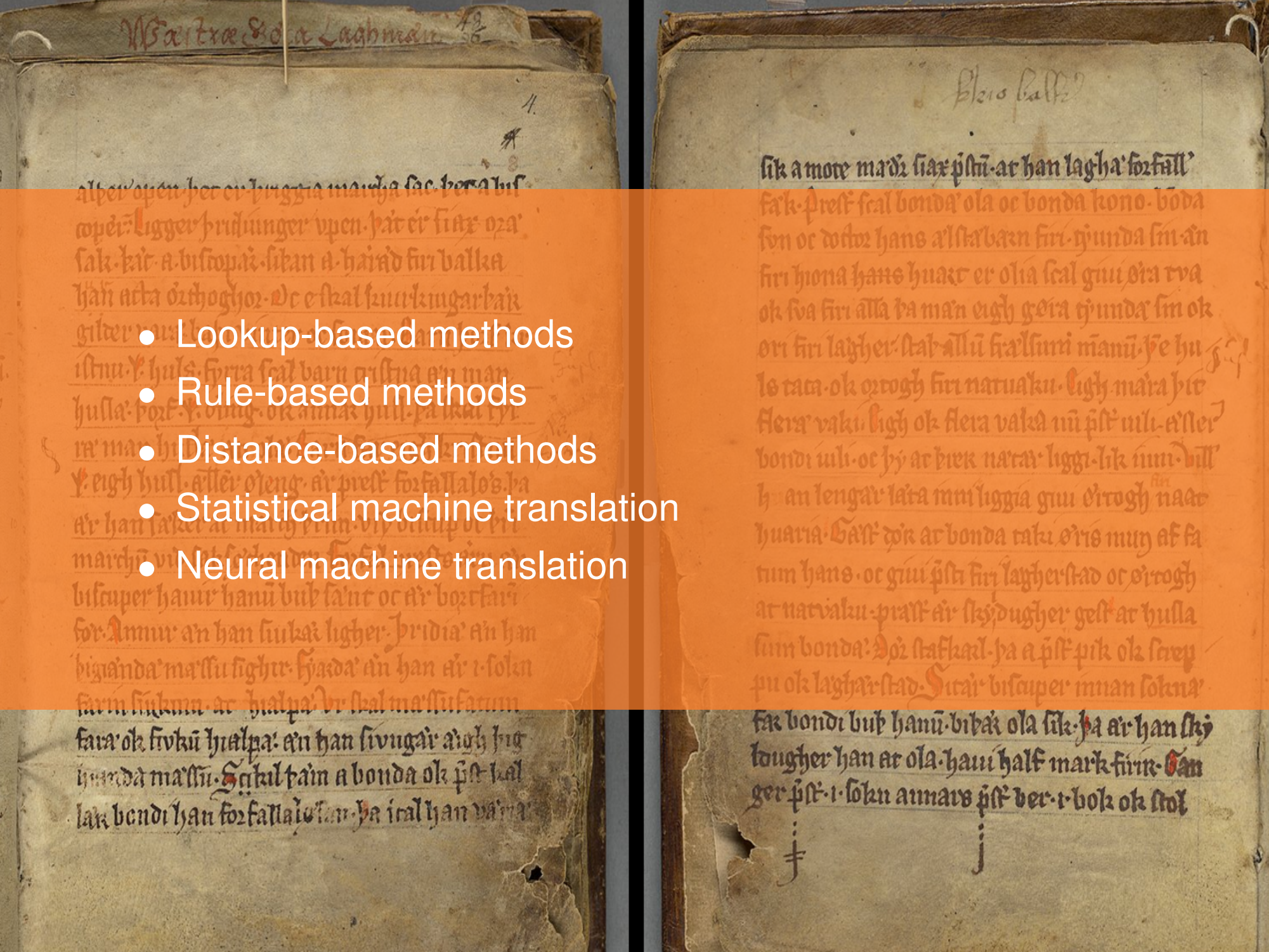
Text normalization

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

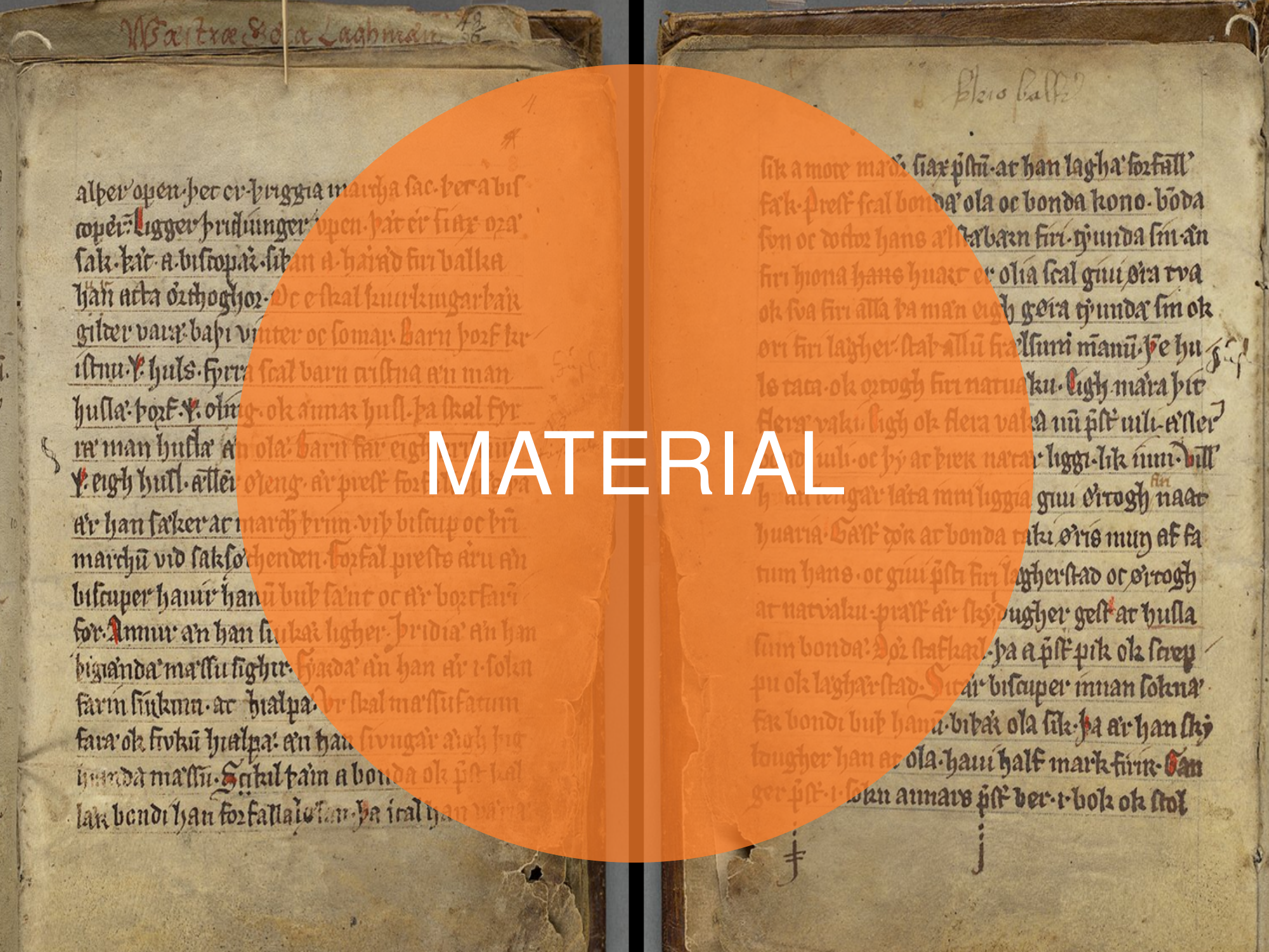
...historical texts tend to contain a lot of variant spellings

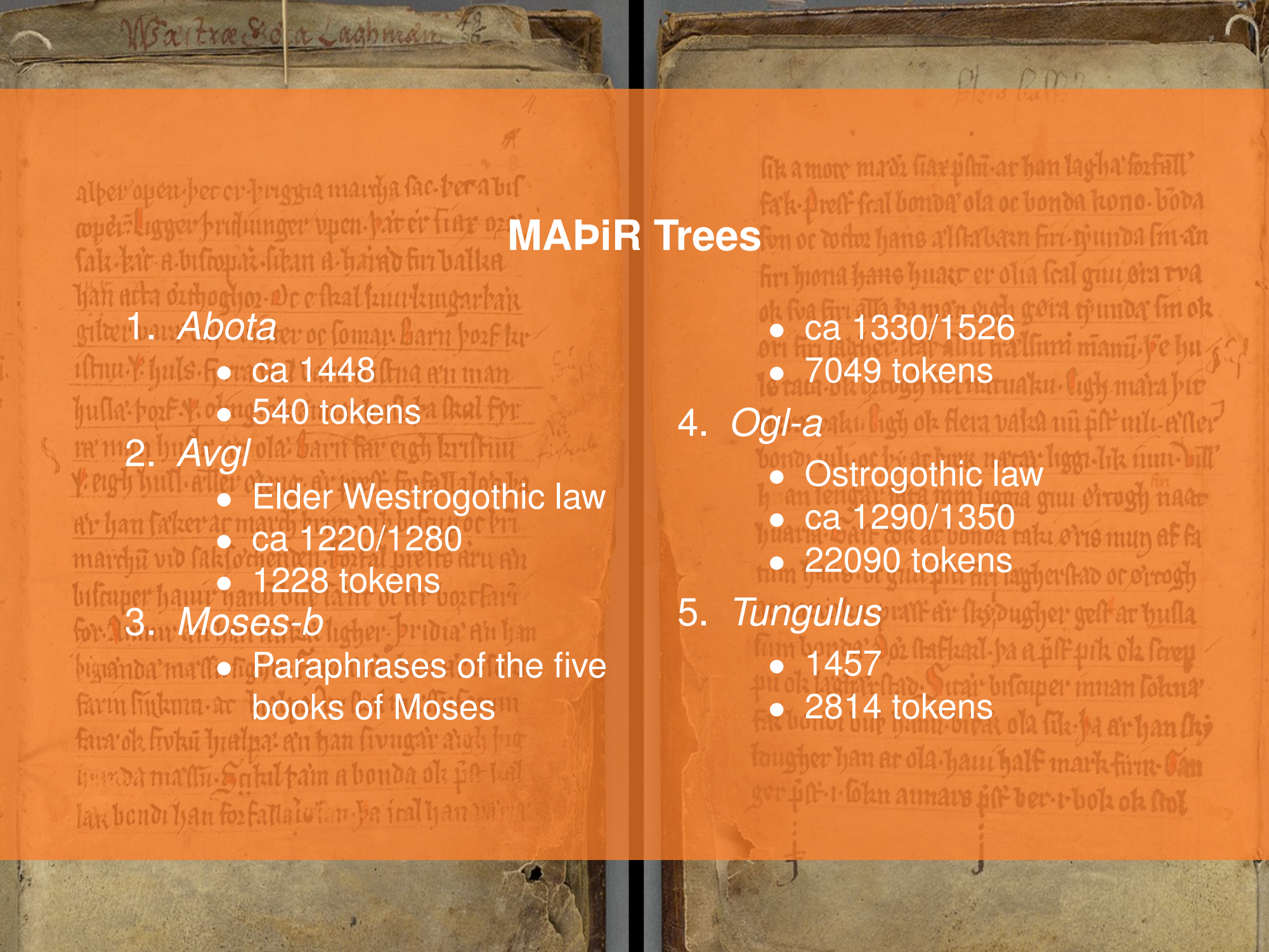


(Bollmann 2019)



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


MATERIAL

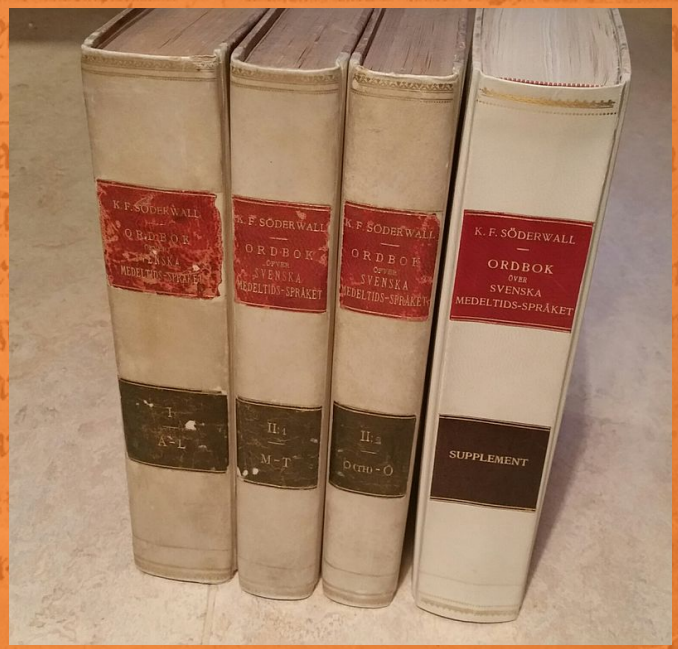


MAÞiR Trees

1. *Abota*
   - ca 1448
   - 540 tokens
2. *Avgl*
   - Elder Westrogothic law
   - ca 1220/1280
   - 1228 tokens
3. *Moses-b*
   - Paraphrases of the five books of Moses
   - ca 1330/1526
   - 7049 tokens
4. *Ogl-a*
   - Ostrogothic law
   - ca 1290/1350
   - 22090 tokens
5. *Tungulus*
   - 1457
   - 2814 tokens

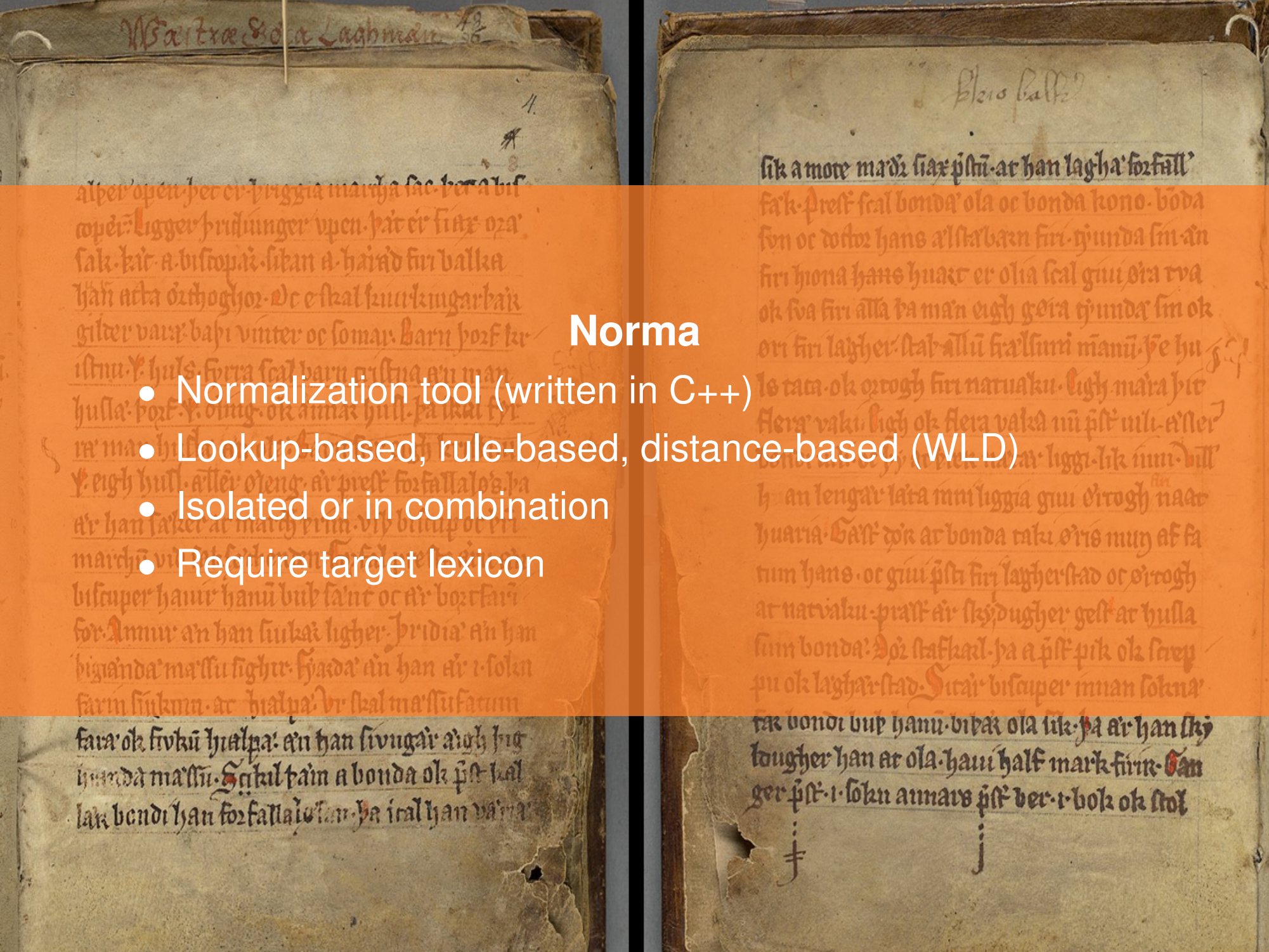
MAÞiR Words

- 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
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	


SOFTWARE



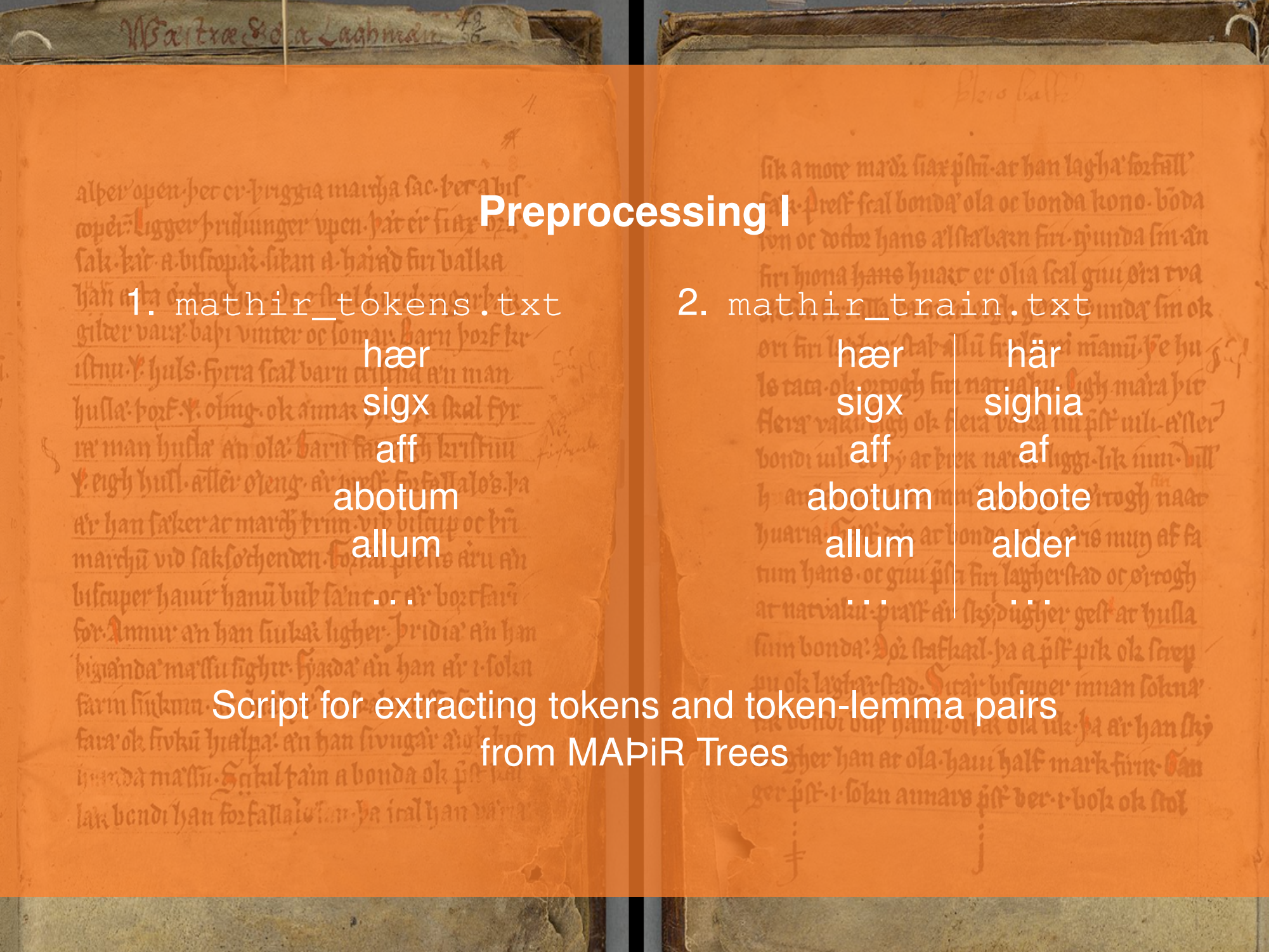
Norma

- 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

METHOD



Preprocessing I

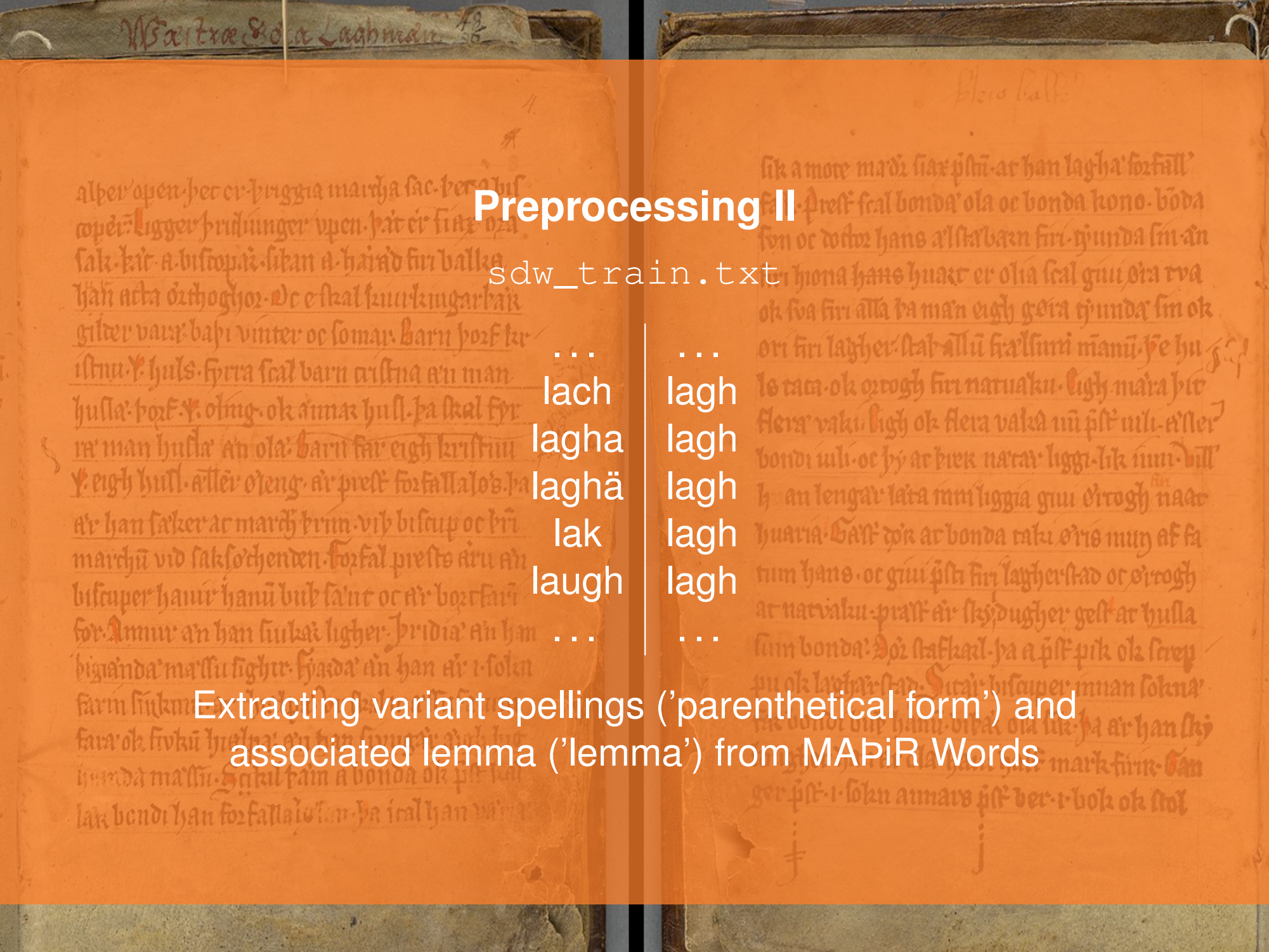
1. `mathir_tokens.txt`

hær
sigx
aff
abotum
allum
...

2. `mathir_train.txt`

hær	här
sigx	sighia
aff	af
abotum	abbote
allum	alder
...	...

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

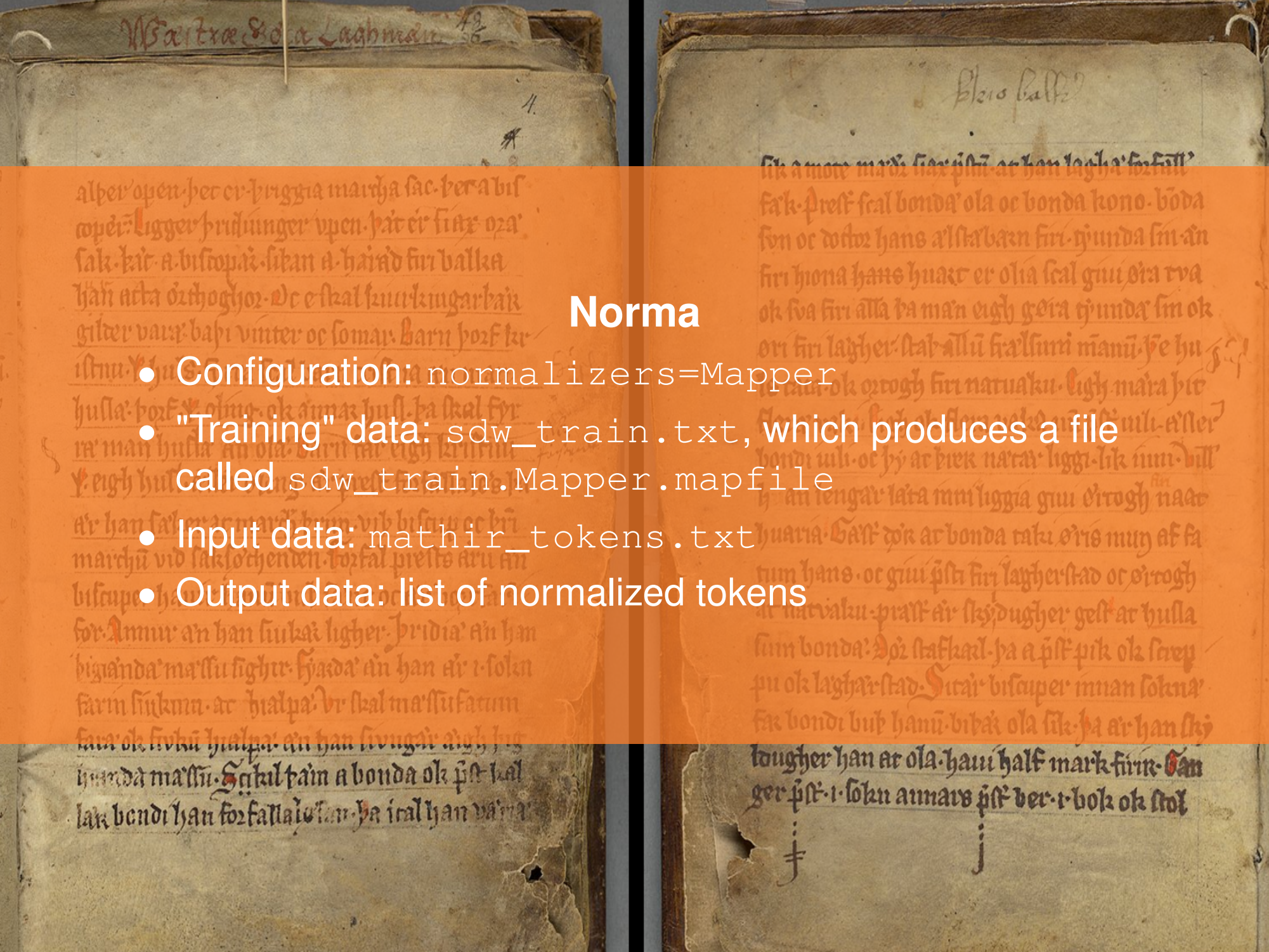


Preprocessing II

`sdw_train.txt`

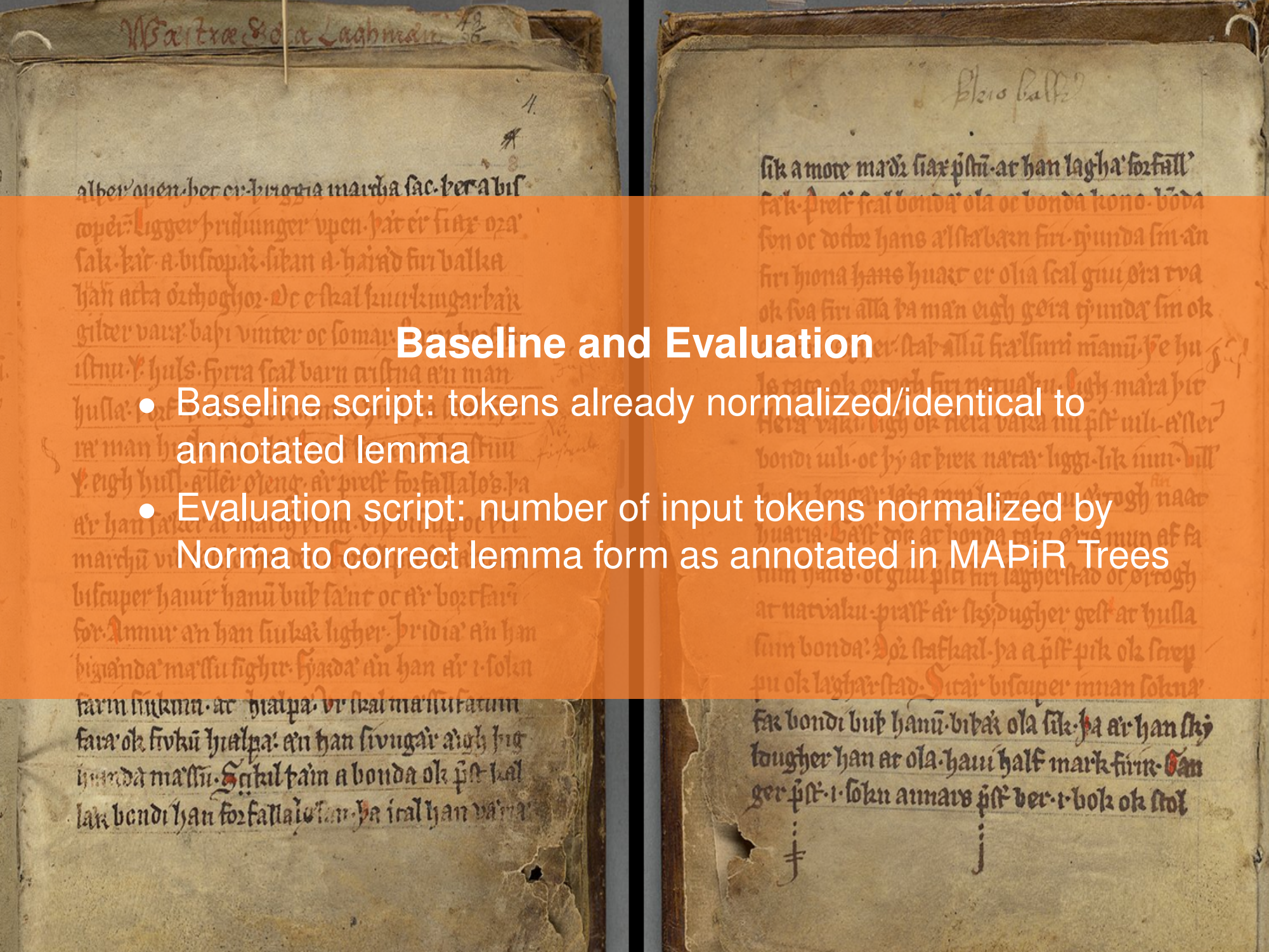
...	...
lach	lagh
lagha	lagh
laghä	lagh
lak	lagh
laugh	lagh
...	...

Extracting variant spellings ('parenthetical form') and associated lemma ('lemma') from MAÞiR Words



Norma

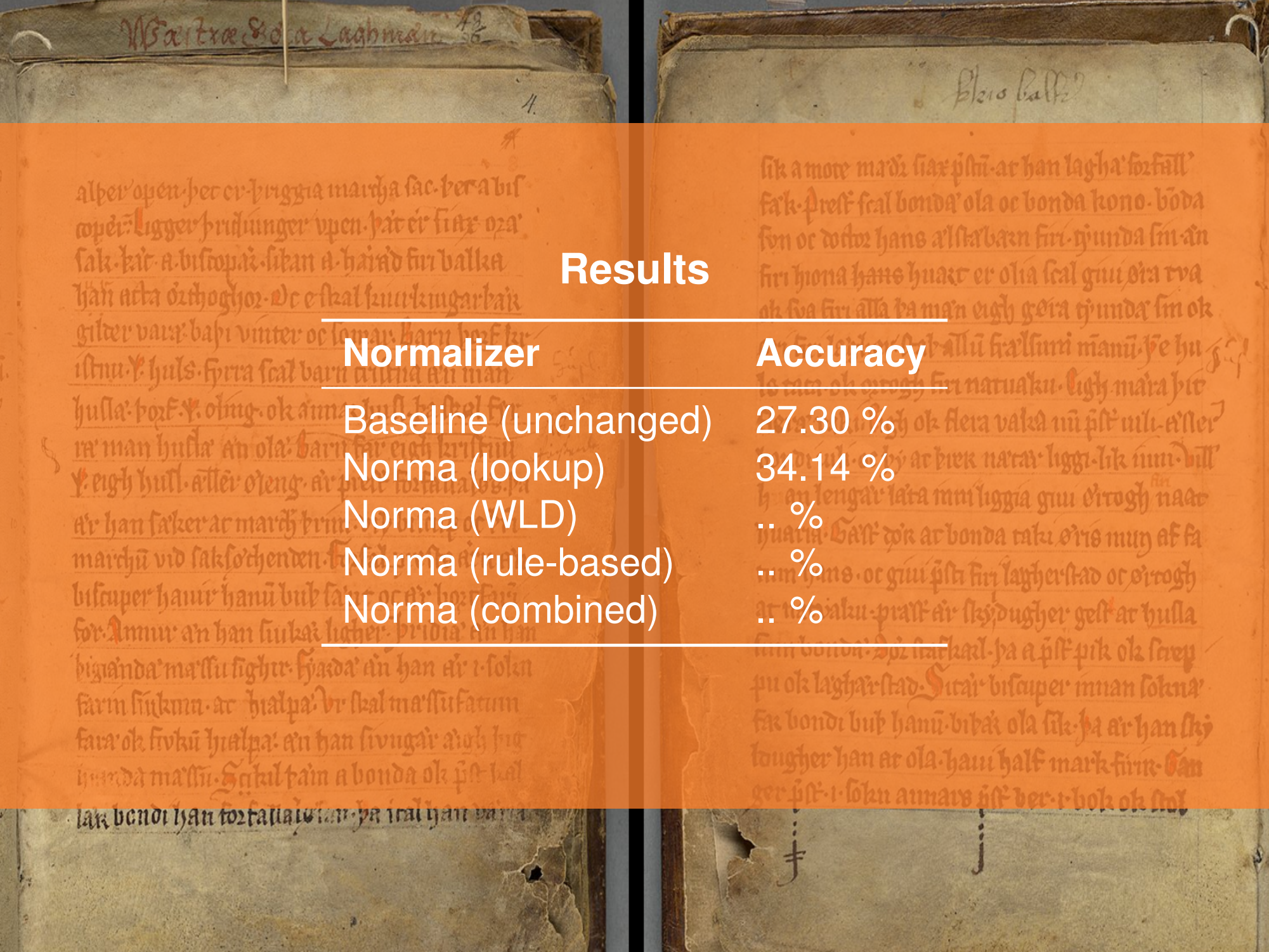
- Configuration: `normalizers=Mapper`
- "Training" data: `sdw_train.txt`, which produces a file called `sdw_train.Mapper.mapfile`
- Input data: `mathir_tokens.txt`
- Output data: list of normalized tokens



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 MAÞiR Trees


RESULTS



Results

Normalizer	Accuracy
Baseline (unchanged)	27.30 %
Norma (lookup)	34.14 %
Norma (WLD)	.. %
Norma (rule-based)	.. %
Norma (combined)	.. %

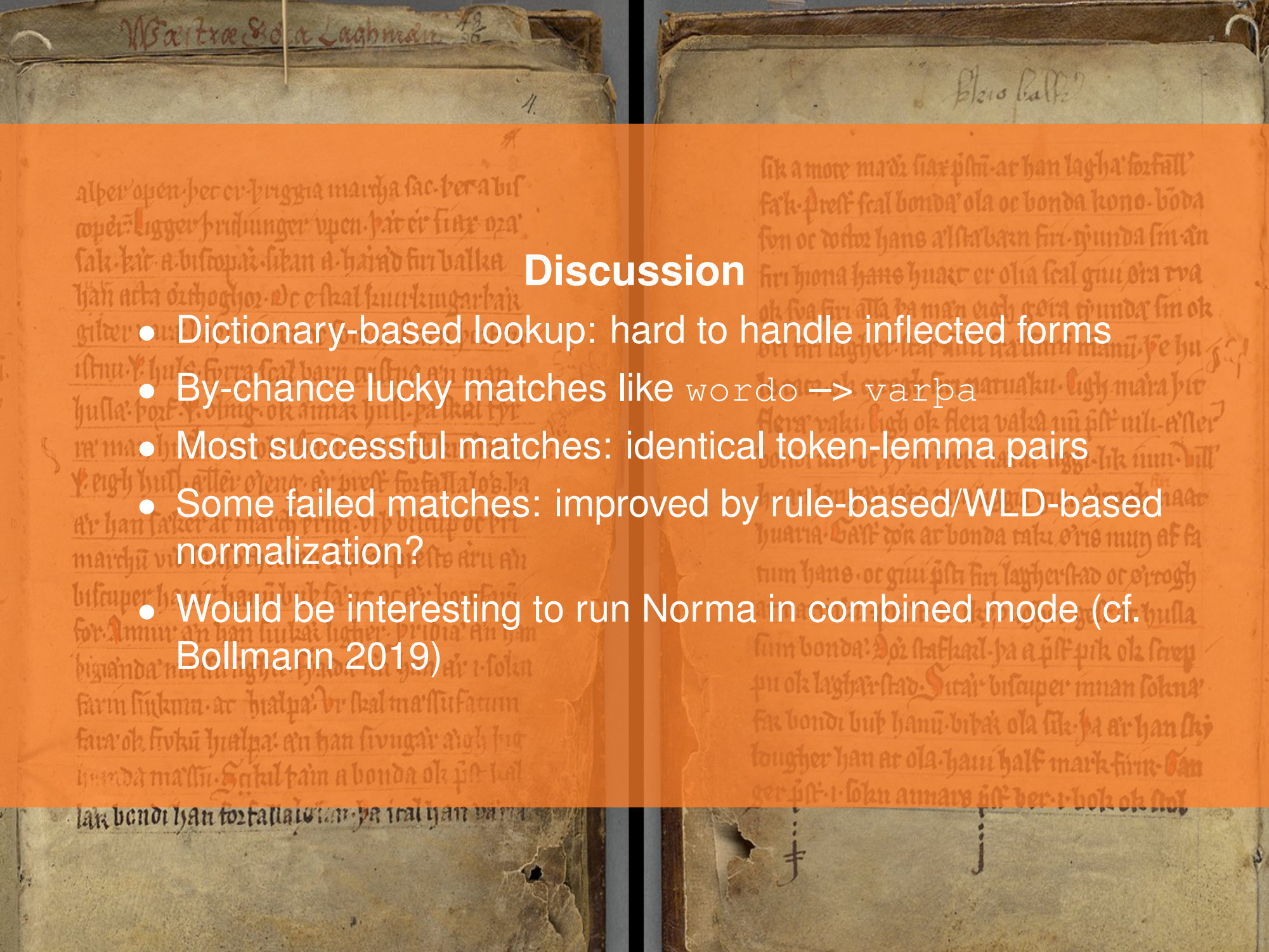
Matches

input	output
…	
oc	ok
uili	vilia
scal	skula
med	mäþ
þessi	þänne
engelin	ängeli
bönder	bonde
wordo	varþa
kyrkiu	kirkia
landbor	landboe
himmerike	himirike
…	

Failed Matches

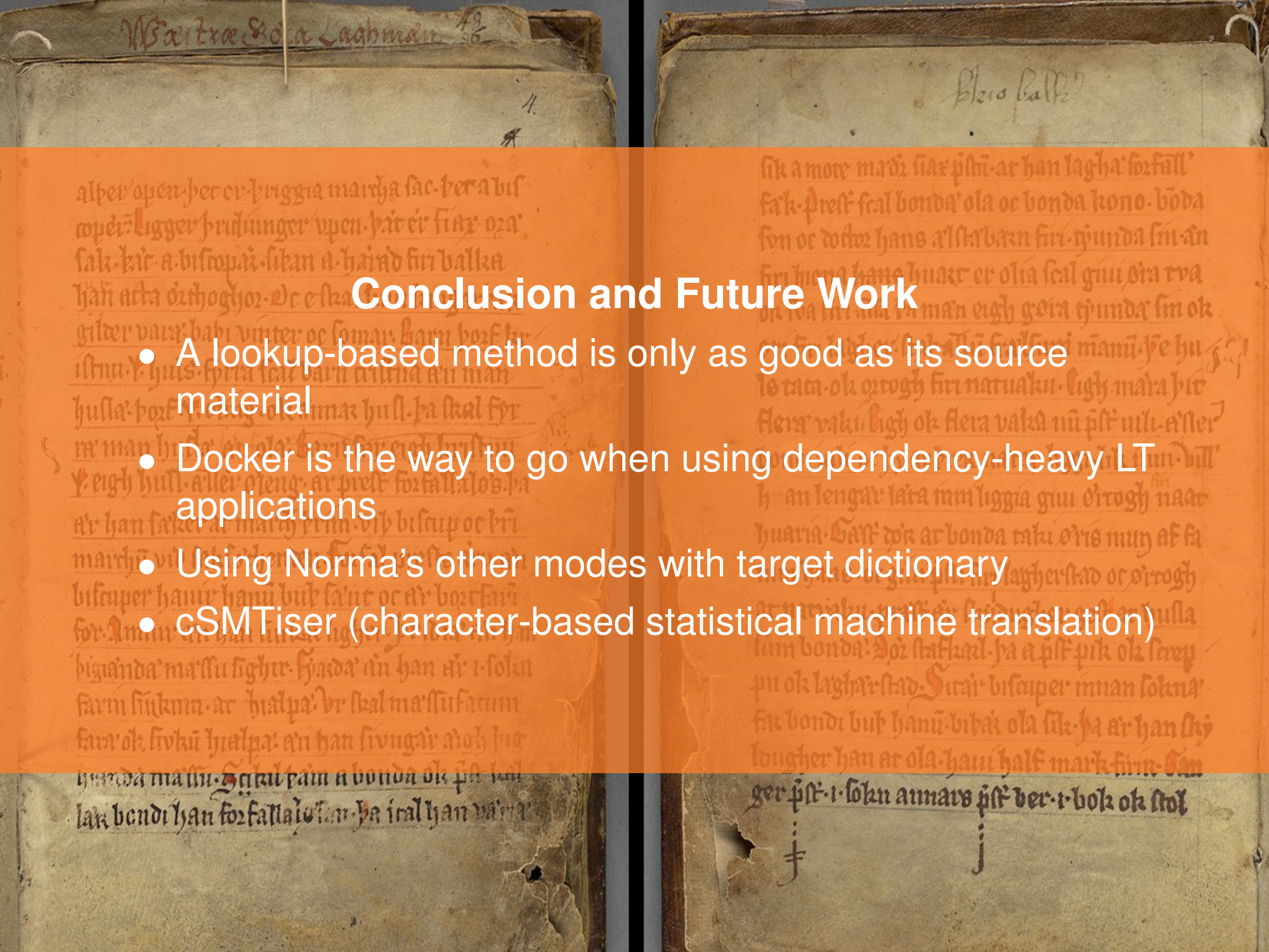
input	output	correct
…		
wt	wt	ut
ær	ær	vara
thet	tea	þän
æter	æter	äta
þa	þiggia	þa
bætri	bætri	bätre
sigher	sigha	sighia
wigyæ	wigyæ	vighia
guðfæþur	guðfæþur	guþfaþir
for budhit	for budhit	forbiuþa
wtan widh	wtan widh	utan viþer
…		


DISCUSSION



Discussion

- Dictionary-based lookup: hard to handle inflected forms
- By-chance lucky matches like `wordo` –> `varþa`
- 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)



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)


THANK YOU