Unsupervised Regularization of Historical Texts for POS Tagging

Fabian Barteld
fabian.barteld@uni-hamburg.de

Ingrid Schröder
ingrid.schroeder@uni-hamburg.de

Heike Zinsmeister
heike.zinsmeister@uni-hamburg.de

Corpus-based Research in the Humanities
December 10, 2015; Warsaw, Poland

Reference Corpus Middle Low German/Low Rhenish

- Funded by the German Research Foundation (DFG)

Reference Corpus Middle Low German/Low Rhenish

- Funded by the German Research Foundation (DFG)
- 2013–2018

Reference Corpus Middle Low German/Low Rhenish

- Funded by the German Research Foundation (DFG)
- 2013–2018
- Middle Low German (GML)
  - Group of dialects
  - Time: 1200-1650
  - Region: Northern Germany

Reference Corpus Middle Low German/Low Rhenish

- Funded by the German Research Foundation (DFG)
- 2013–2018
- Middle Low German (GML)
  - Group of dialects
  - Time: 1200-1650
  - Region: Northern Germany

Name	Year	Domain	Tokens	Types
Johannes	~1480	religious texts	19645	2305
Griseldis	1502	literature	9062	2251
OldenbSSP	1336	law	21800	2731
SaechsWeltchr	1st half 14th c.	arts	18215	3255
		Sum	68722	10542

Overview

- Spelling Variation
- Modernization
- Regularization
- Unsupervised, Language-independent Regularization
  - The Approach
  - Evaluation
- Conclusion

Spelling Variation

Appears in non-standard texts: historical texts, user-generated content, ...

Spelling Variation

Appears in non-standard texts: historical texts, user-generated content, ...

vorwar vorwar segge ik iu
vorwar uorwar segge ik iu
uorwar vorwar segge ik iu
Uorwar vorwar segge ik iu
uorwar uorwar segge ik ju
vorwar uorwar segge ik iw
UOrwarvorwar segge ik iw
in truth in truth tell I you

'I tell you in truth'
(Johannes)

Spelling Variation

Appears in non-standard texts: historical texts, user-generated content, ...

vorwar vorwar segge ik iu
vorwar uorwar segge ik iu
uorwar vorwar segge ik iu

uorwar uorwar segge ik ju
vorwar uorwar segge ik iw
uorwar vorwar segge ik iw
in truth in truth tell I you

'I tell you in truth'
(Johannes)

Spelling Variation

Appears in non-standard texts: historical texts, user-generated content, ...

vorwar vorwar segge ik iu
vorwar uorwar segge ik iu
uorwar vorwar segge ik iu

uorwar uorwar segge ik ju
vorwar uorwar segge ik iw
uorwar vorwar segge ik iw
in truth in truth tell I you

{vorwar, uorwar}
{iu, ju, iw}

'I tell you in truth'
(Johannes)

Spelling Variation

Appears in non-standard texts: historical texts, user-generated content, ...

vorwar vorwar segge ik iu
vorwar uorwar segge ik iu
uorwar vorwar segge ik iu

uorwar uorwar segge ik ju
vorwar uorwar segge ik iw
uorwar vorwar segge ik iw
in truth in truth tell I you

{vorwar, uorwar}
{iu, ju, iw}

'I tell you in truth'
(Johannes)

Problems

- Limits the usefulness of unannotated corpora [Baron et al., 2009]

Problems

- Limits the usefulness of unannotated corpora [Baron et al., 2009]
- Makes (automatic) annotation harder

> *"Non-standard words are present in many text genres, including advertisements, professional forums, and SMS messages. They can be the cause of reading and understanding problems for humans, and degrade the accuracy of text processing tools [...]."*

[Baldwin et al., 2015]

Solutions in Computational Linguistics

- Automatically preprocess the texts
  - Historical texts
    - Normalization [Bollmann et al., 2012]
    - Canonicalization [Jurish, 2013]
    - Modernization [Scherrer and Erjavec, 2016]
  - User-generated content
    - Standardization [Ljubešić et al., 2014]
    - Normalization [Saloot et al., 2015]

Solutions in Computational Linguistics

- Automatically preprocess the texts
  - Historical texts
    - Normalization [Bollmann et al., 2012]
    - Canonicalization [Jurish, 2013]
    - Modernization [Scherrer and Erjavec, 2016]
  - User-generated content
    - Standardization [Ljubešić et al., 2014]
    - Normalization [Saloot et al., 2015]

Overview

- Spelling Variation
- Modernization
- Regularization
- Unsupervised, Language-independent Regularization
  - The Approach
  - Evaluation
- Conclusion

Modernization

Definition

Mapping historical words to the equivalent modern form.

Modernization

Definition

Mapping historical words to the equivalent modern form.

(non-standard) (standard)

Modernization

Definition

Mapping historical words to the equivalent modern form.

(non-standard) (standard)

vorwar	vorwar	segge	ik	iu
vorwar	uorwar	segge	ik	iu
uorwar	vorwar	segge	ik	iu
uorwar	uorwar	segge	ik	ju
vorwar	uorwar	segge	ik	iw
uorwar	vorwar	segge	ik	iw

in truth	in truth	tell	I	you

Modernization

Definition

Mapping historical words to the equivalent modern form.

(non-standard) (standard)

vorwar	vorwar	segge	ik	iu	
vorwar	uorwar	segge	ik	iu	
uorwar	vorwar	segge	ik	iu	
uorwar	uorwar	segge	ik	ju	
vorwar	uorwar	segge	ik	iw	
uorwar	vorwar	segge	ik	iw	
↓	↓	↓	↓	↓	
fürwahr	fürwahr	sage	ich	euch	Modern German
in truth	in truth	tell	I	you	

Benefits and Requirements of Modernization

Benefits

1. Variation/noise in the data is reduced
2. Tools developed for the modern variant can be used (with reasonable accuracy)

Benefits and Requirements of Modernization

Benefits

1. Variation/noise in the data is reduced
2. Tools developed for the modern variant can be used (with reasonable accuracy)

Minimal requirement

Some definition of the modern target language (dictionary, list of target forms, corpus)

Problems with Modernization

Target language not always obvious, e.g. Middle Low German

Problems with Modernization

Target language not always obvious, e.g. Middle Low German

- (Modern) Low German
  - Not standardized
  - Low-resourced

Problems with Modernization

Target language not always obvious, e.g. Middle Low German

- (Modern) Low German
  - Not standardized
  - Low-resourced
- Modern German
  vorwar $\rightarrow$ *fürwahr*

Problems with Modernization

Target language not always obvious, e.g. Middle Low German

- (Modern) Low German
  - Not standardized
  - Low-resourced
- Modern German
  vorwar $\rightarrow$ *fürwahr*
  ju $\rightarrow$ *euch*

Problems with Modernization

Target language not always obvious, e.g. Middle Low German

- (Modern) Low German
  - Not standardized
  - Low-resourced

- Modern German
  vorwar $\rightarrow$ *fürwahr*
  ju $\rightarrow$ *euch*

- Modern English
  ju $\rightarrow$ *you*

Problems with Modernization

Target language not always obvious, e.g. Middle Low German

- (Modern) Low German
  - Not standardized
  - Low-resourced

- Modern German
  vorwar $\rightarrow$ *fürwahr*
  ju $\rightarrow$ *euch*

- Modern English
  vorwar $\rightarrow$ *in truth*
  ju $\rightarrow$ *you*

Problems with Modernization

Target language not always obvious, e.g. Middle Low German

- (Modern) Low German
  - Not standardized
  - Low-resourced

- Modern German
  vorwar $\rightarrow$ *fürwahr*
  ju $\rightarrow$ *euch*

- Modern English
  vorwar $\rightarrow$ *in truth*
  ju $\rightarrow$ *you*

[Sanders, 1982]

Overview

- Spelling Variation
- Modernization
- Regularization
- Unsupervised, Language-independent Regularization
  - The Approach
  - Evaluation
- Conclusion

Alternative to Modernization: Regularization

Definition

Conflating spelling variants into conflation sets.

vorwar	vorwar	segge	ik	iu
vorwar	uorwar	segge	ik	iu
uorwar	vorwar	segge	ik	iu
uorwar	uorwar	segge	ik	ju
vorwar	uorwar	segge	ik	iw
uorwar	vorwar	segge	ik	iw
↓	↓	↓	↓	↓
fürwahr	fürwahr	sage	ich	euch
in truth	in truth	tell	I	you

Alternative to Modernization: Regularization

Definition

Conflating spelling variants into conflation sets.

vorwar	vorwar	segge	ik	iu
vorwar	uorwar	segge	ik	iu
uorwar	vorwar	segge	ik	iu
uorwar	uorwar	segge	ik	ju
vorwar	uorwar	segge	ik	iw
uorwar	vorwar	segge	ik	iw
↓	↓	↓	↓	↓
fürwahr	fürwahr	sage	ich	euch
in truth	in truth	tell	I	you

Alternative to Modernization: Regularization

Definition

Conflating spelling variants into conflation sets.

vorwar	vorwar	segge	ik	iu
vorwar	uorwar	segge	ik	iu
uorwar	vorwar	segge	ik	iu
uorwar	uorwar	segge	ik	ju
vorwar	uorwar	segge	ik	iw
uorwar	vorwar	segge	ik	iw
↓	↓	↓	↓	↓
fürwahr	fürwahr	sage	ich	euch
{uorwar,vorwar}	{uorwar,vorwar}	segge	ik	{iu,ju,iw}
in truth	in truth	tell	I	you

Benefits and Requirements of Regularization

Benefits

1. Variation/noise in the data is reduced
2. ~~Tools developed for the modern variant can be used (with reasonable accuracy)~~

Benefits and Requirements of Regularization

Benefits

1. Variation/noise in the data is reduced
2. ~~Tools developed for the modern variant can be used (with reasonable accuracy)~~ not always necessary, e.g. keyword statistics

Benefits and Requirements of Regularization

Benefits

1. Variation/noise in the data is reduced
2. ~~Tools developed for the modern variant can be used (with reasonable accuracy)~~ not always necessary, e.g. keyword statistics

Minimal requirement

~~Some definition of the modern target language (dictionary, list of target forms, corpus)~~

Overview

- Spelling Variation
- Modernization
- Regularization
- Unsupervised, Language-independent Regularization
  - The Approach
  - Evaluation
- Conclusion

Existing Regularization Approaches

Two existing approaches for historical texts
[Logačev et al., 2014, Kestemont et al., 2010]

Existing Regularization Approaches

Two existing approaches for historical texts
[Logačev et al., 2014, Kestemont et al., 2010]

Both use annotated texts to train the model (POS or Lemma).

Existing Regularization Approaches

Two existing approaches for historical texts
[Logačev et al., 2014, Kestemont et al., 2010]

Both use annotated texts to train the model (POS or Lemma).

Aim: Regularization without existing annotations

Properties of Spelling Variants

Most spelling variants ...

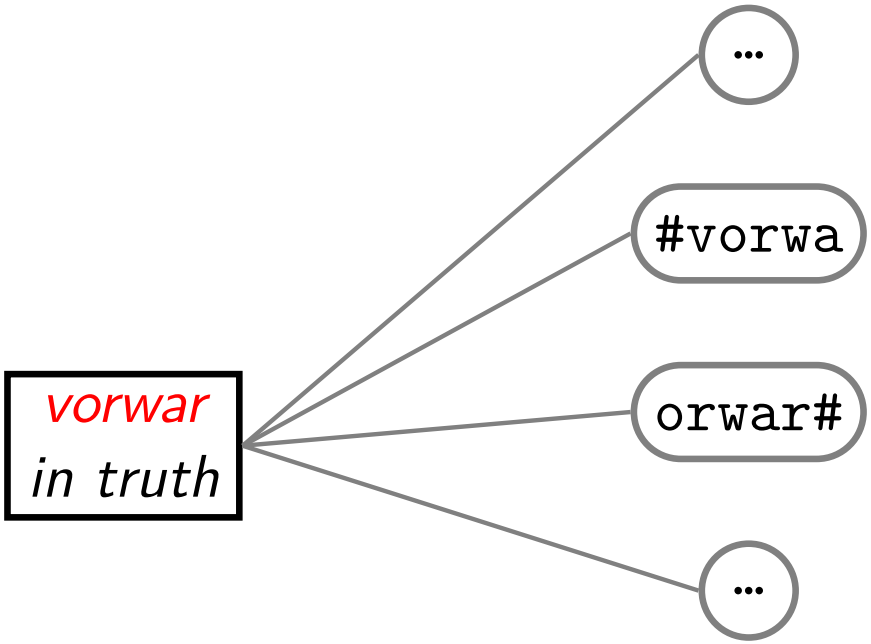
1. are formally similar
2. appear in similar contexts

Properties of Spelling Variants

Most spelling variants ...

1. are formally similar → Proxinette [Hathout, 2014]
2. appear in similar contexts

Proxinette



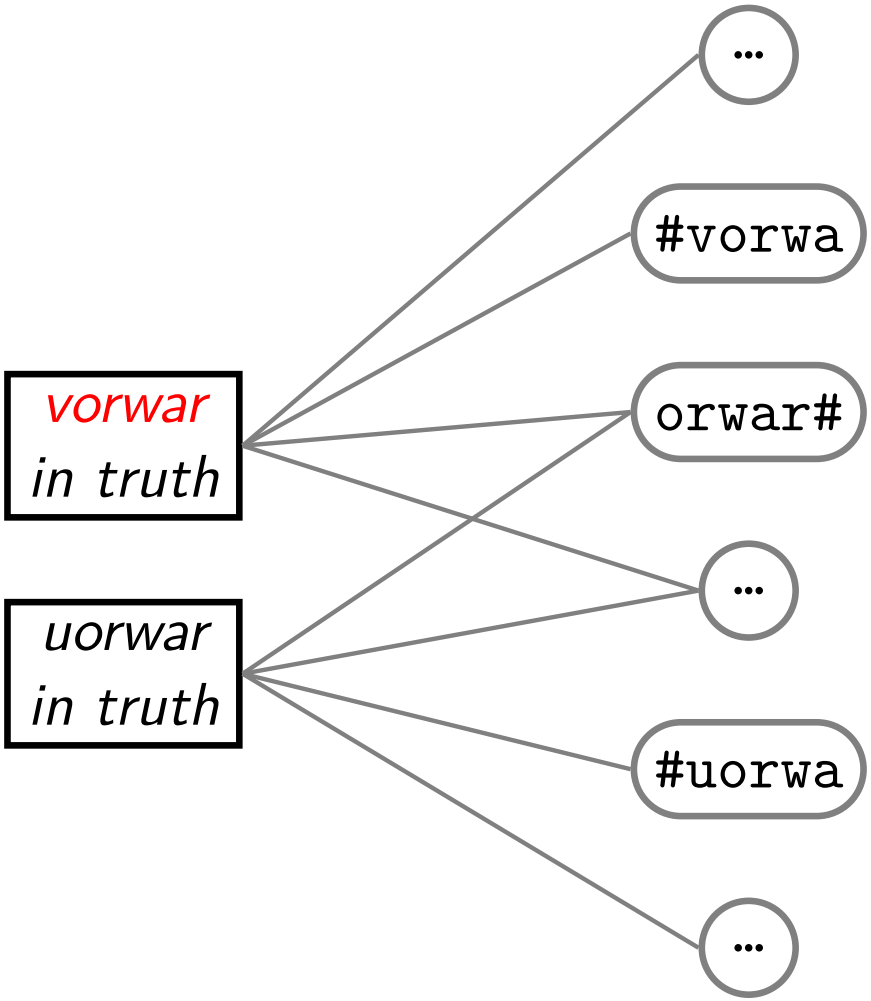
Legend:

Type

`Character-Ngram`

- Word boundary

Proxinette



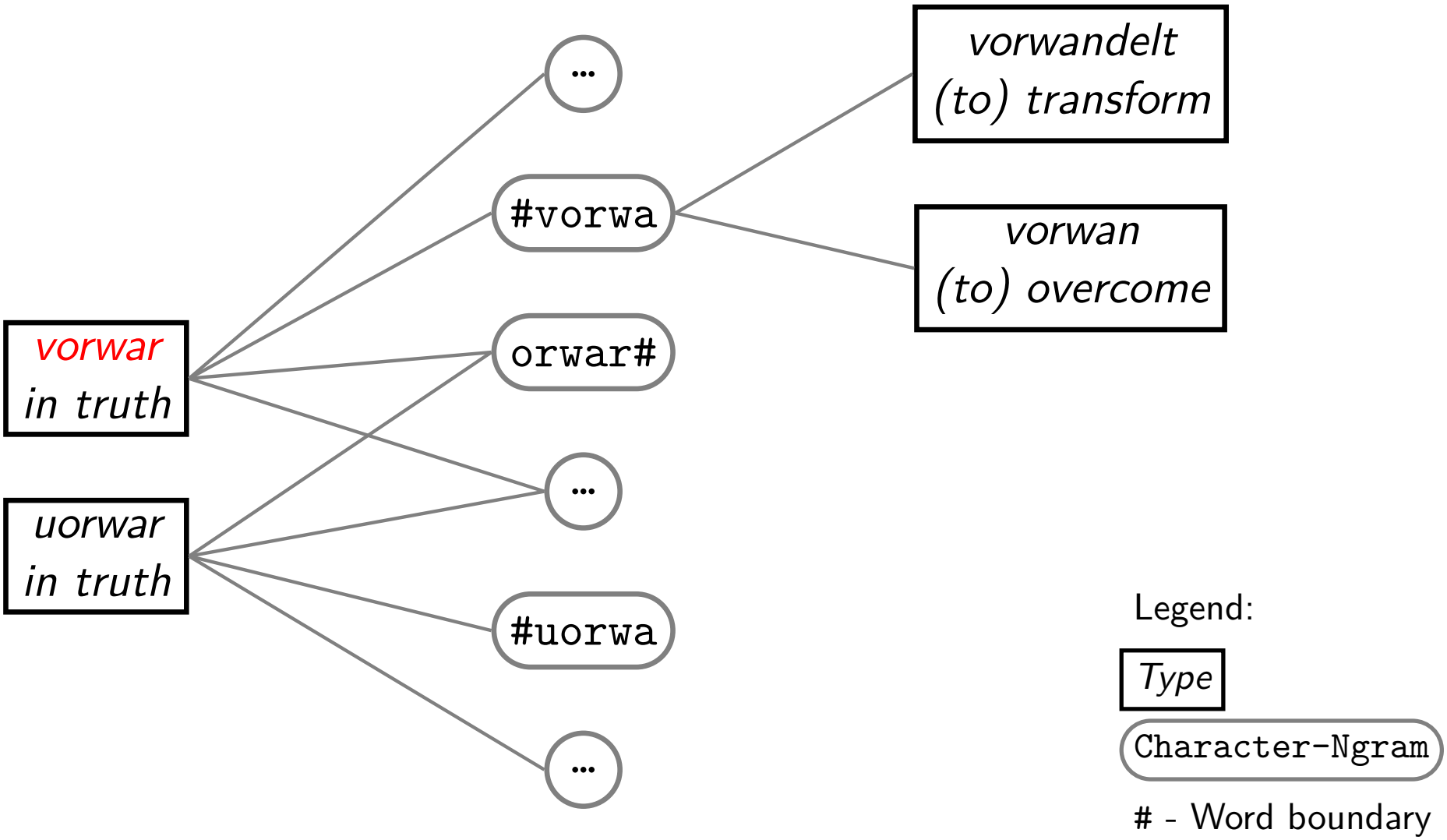
Legend:

Type

`Character-Ngram`

- Word boundary

Proxinette


vorwandelt
(to) transform
...
#vorwa
vorwan
(to) overcome
vorwar
in truth
orwar#
...
uorwar
in truth
#uorwa
...
Legend:
Type
Character-Ngram
- Word boundary

Properties of Spelling Variants

Most spelling variants ...

1. are formally similar → Proxinette [Hathout, 2014]
2. appear in similar contexts

Properties of Spelling Variants

Most spelling variants ...

1. are formally similar → Proxinette [Hathout, 2014]
2. appear in similar contexts → Brown clusters [Brown et al., 1992]

Brown Clusters

Brown Clusters

<BOS> vorwar vorwar segge ik iu

left context of x	x
<BOS>	vorwar

Brown Clusters

<BOS> vorwar vorwar segge ik iu

left context of x	x
<BOS>	vorwar
vorwar	vorwar

Brown Clusters

<BOS> uorwar vorwar segge ik iu

left context of x	x
<BOS>	vorwar
vorwar	vorwar
uorwar	vorwar

Brown Clusters

left context of x	x
<BOS>	vorwar uorwar
vorwar	vorwar uorwar
uorwar	vorwar uorwar

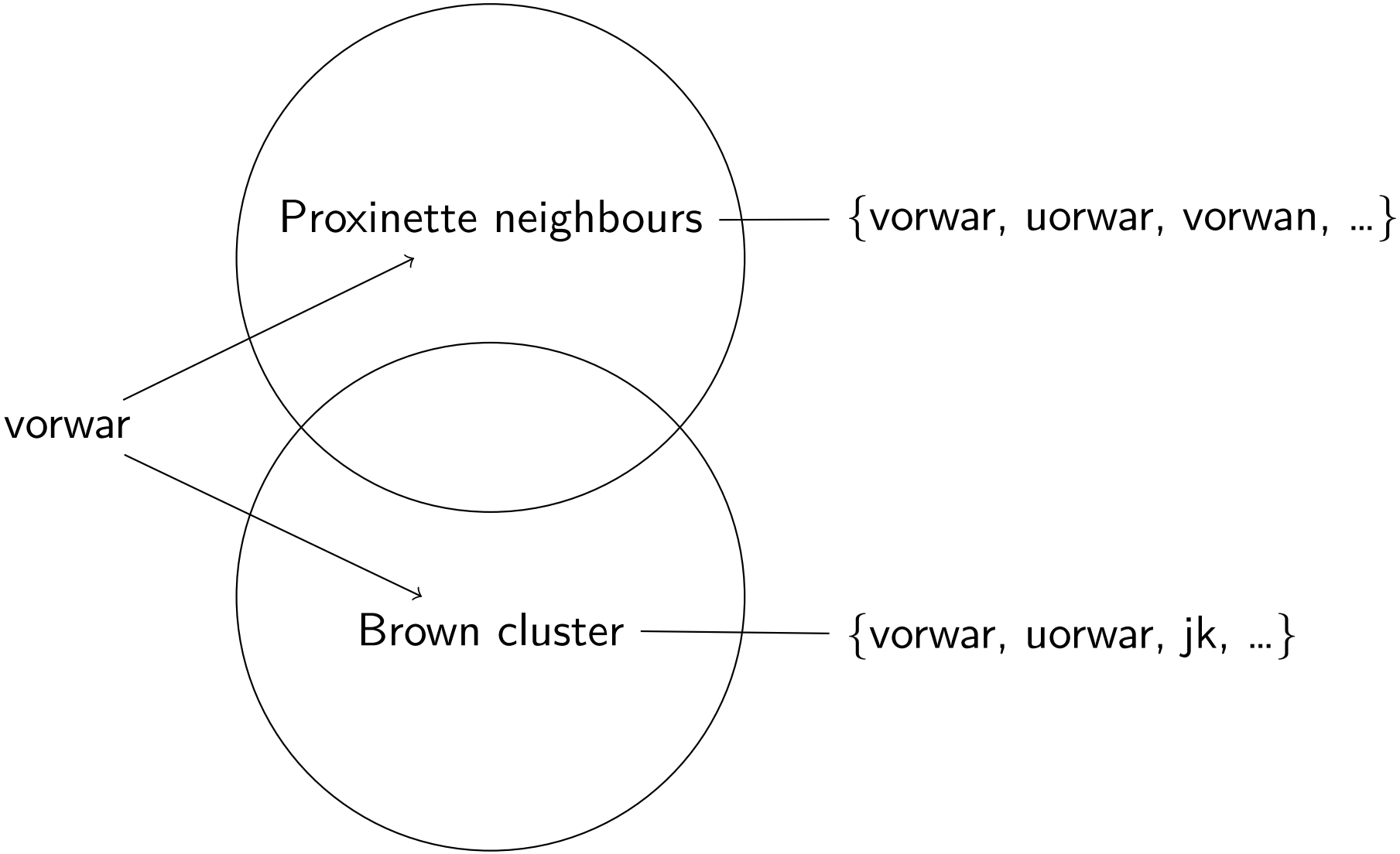
Brown Clusters

left context of x	x
<BOS>	vorwar uorwar jk ik ick ...
vorwar	vorwar uorwar segge ...
uorwar	vorwar uorwar segge ...

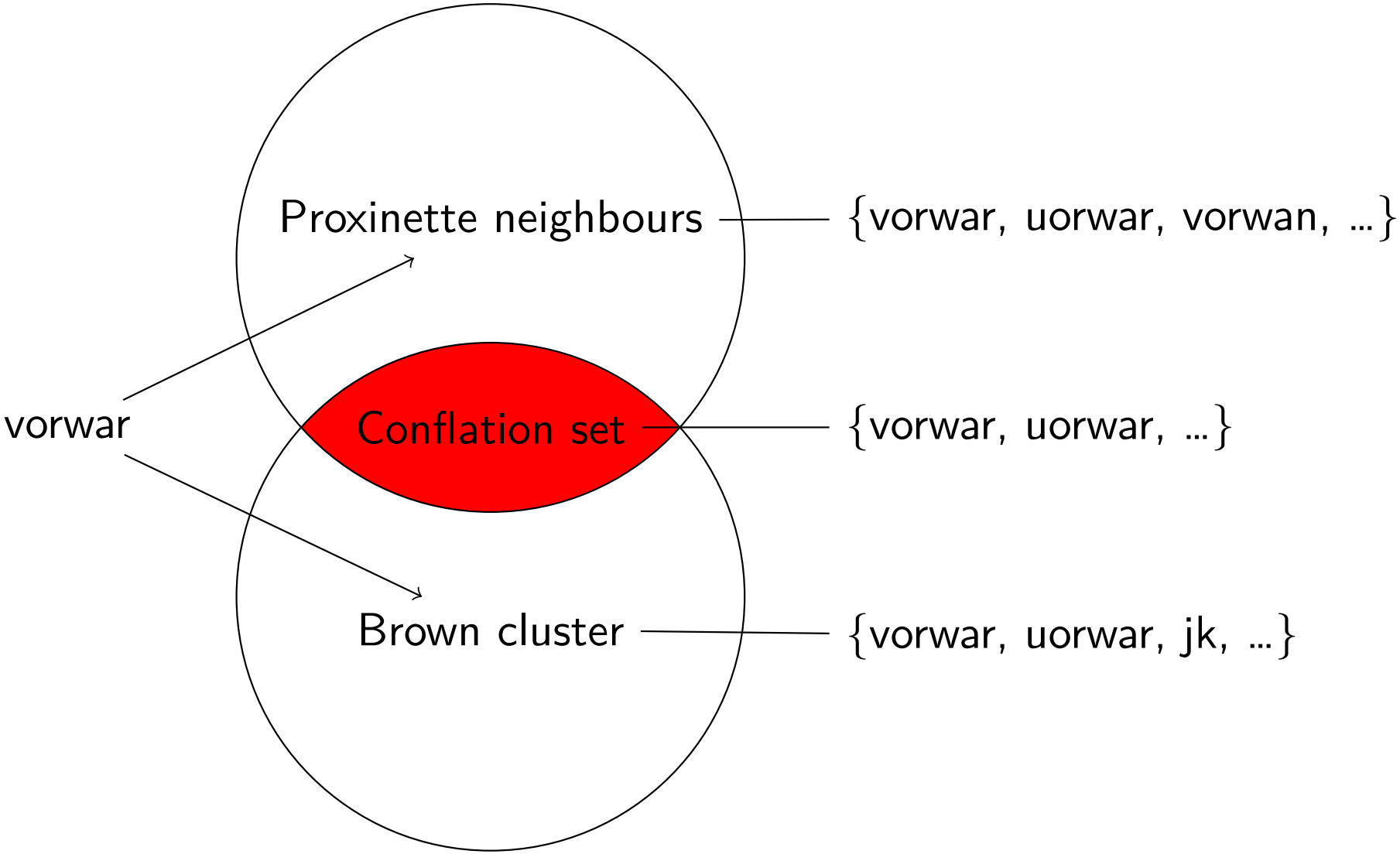
Regularization with Proxinette and Brown clusters

vorwar

Regularization with Proxinette and Brown clusters



Regularization with Proxinette and Brown clusters



Evaluation with POS Tagging

- Train a POS tagger on one text, tag another one.

Evaluation with POS Tagging

- Train a POS tagger on one text, tag another one.
- Unknown words might be spelling variants of known words:

Evaluation with POS Tagging

- Train a POS tagger on one text, tag another one.
- Unknown words might be spelling variants of known words: Substitute unknown types with a randomly chosen known type from the conflation set before applying the tagger.

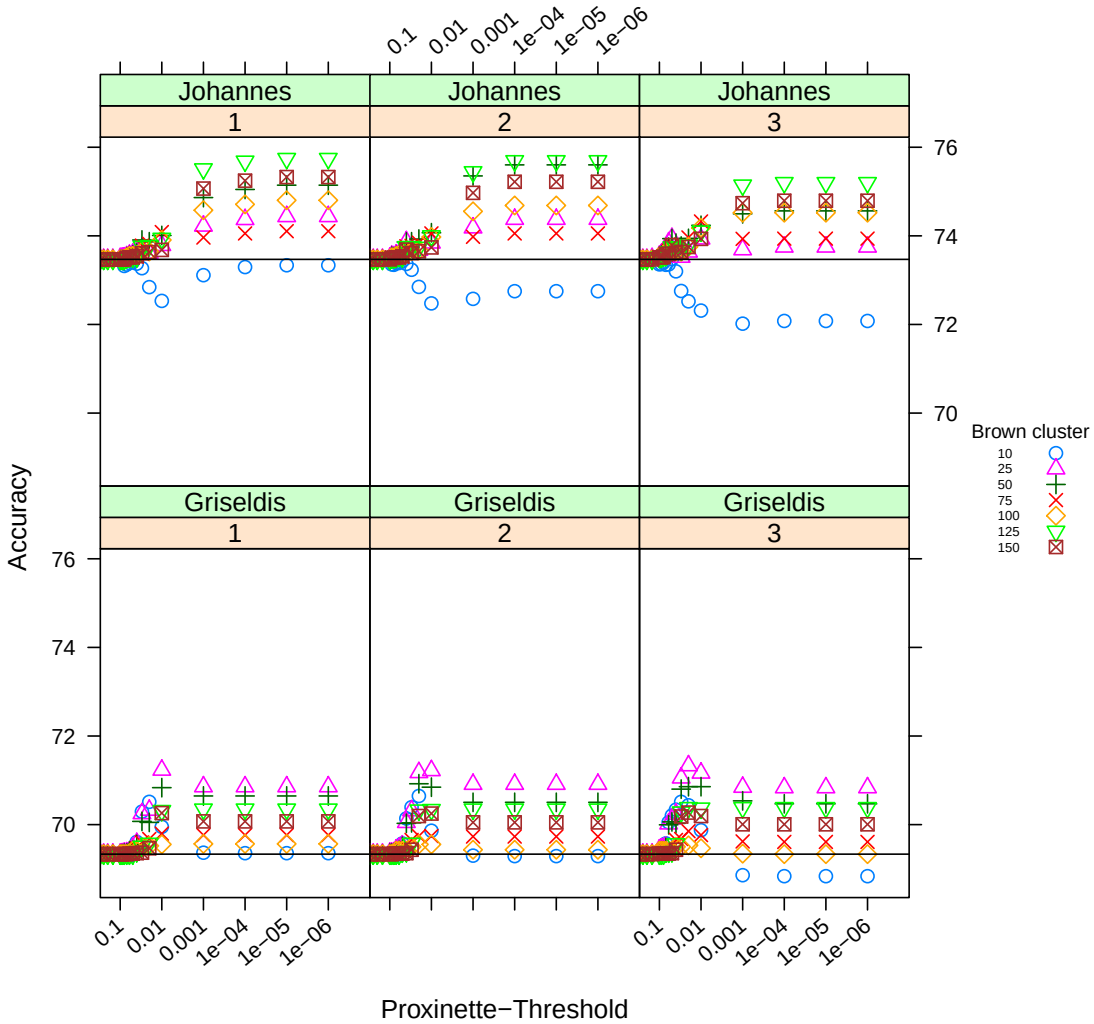
Evaluation with POS Tagging

- Train a POS tagger on one text, tag another one.
- Unknown words might be spelling variants of known words:
  Substitute unknown types with a randomly chosen known type from the conflation set before applying the tagger.
- Expectation:
  Accuracy will increase, if the conflation sets are sets of spelling variants.

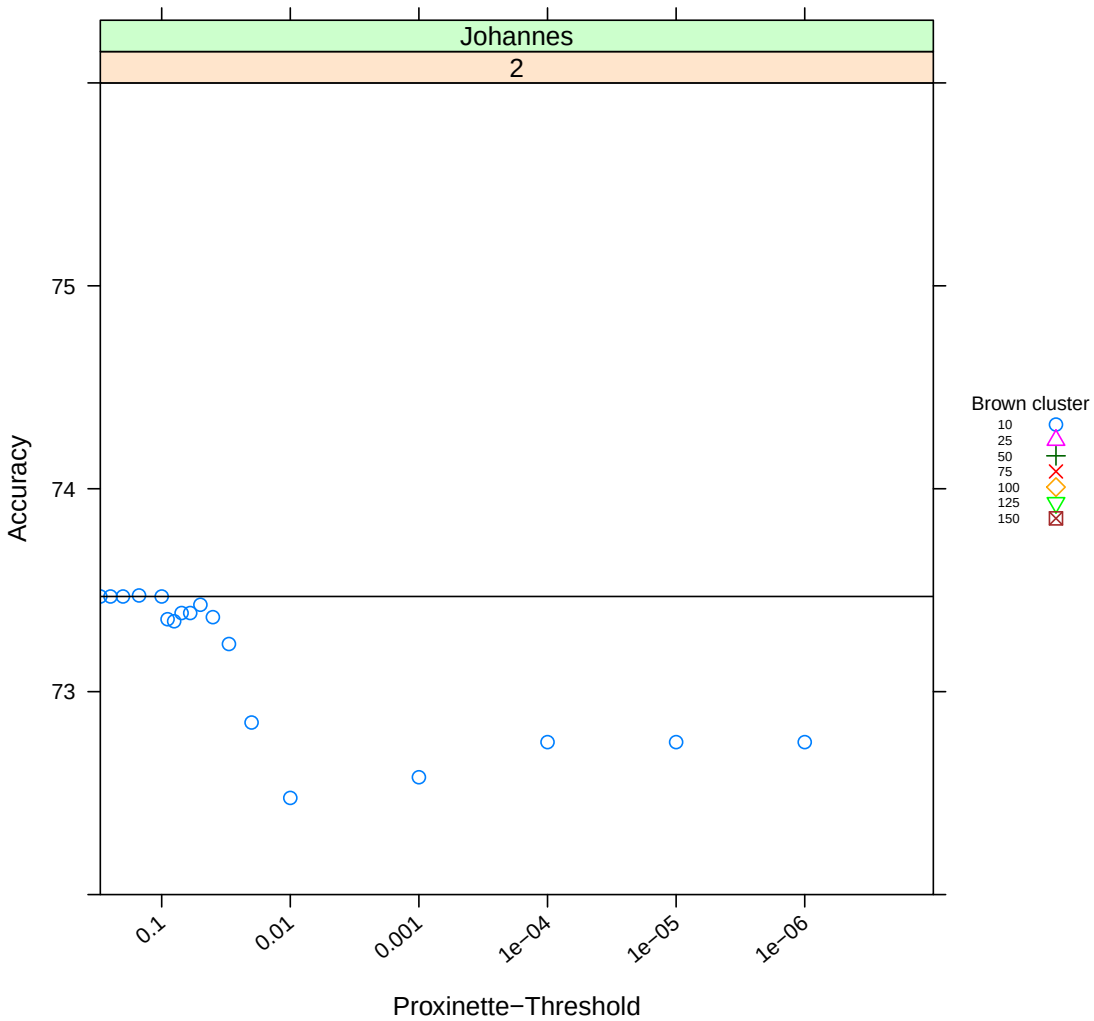
Experiment

- Tagger: RFTagger [Schmid and Laws, 2008]
- Parameters
  - Proxinette: Minimal similarity, Minimal size of character-ngram
  - Brown clusters: Number
  - Both: Texts used to compute the network/clusters

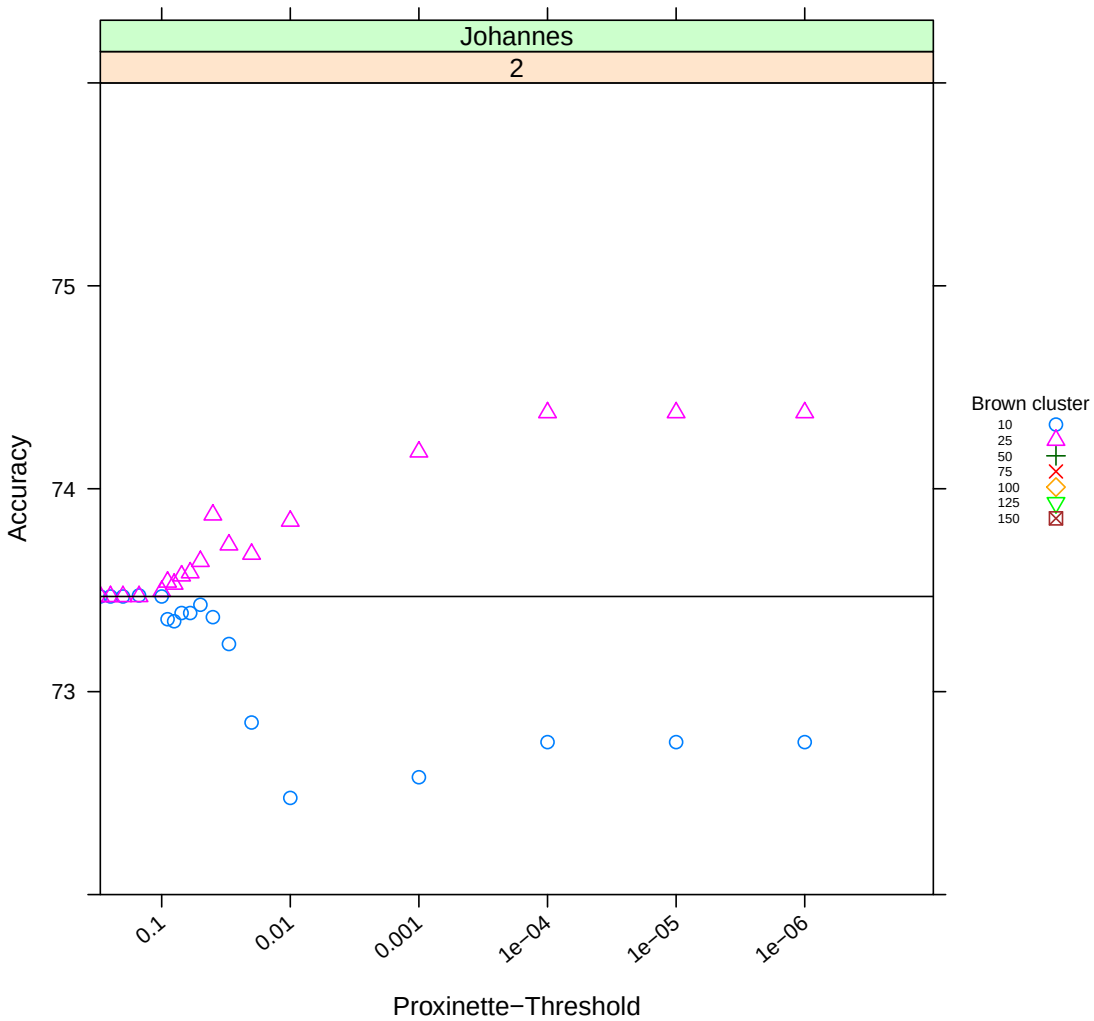
Results


Johannes
1
Johannes
2
Johannes
3
Griseldis
1
Griseldis
2
Griseldis
3
Accuracy
76
74
72
70
0.1
0.01
0.001
1e−04
1e−05
1e−06
Proxinette−Threshold
Brown cluster
10
25
50
75
100
125
150

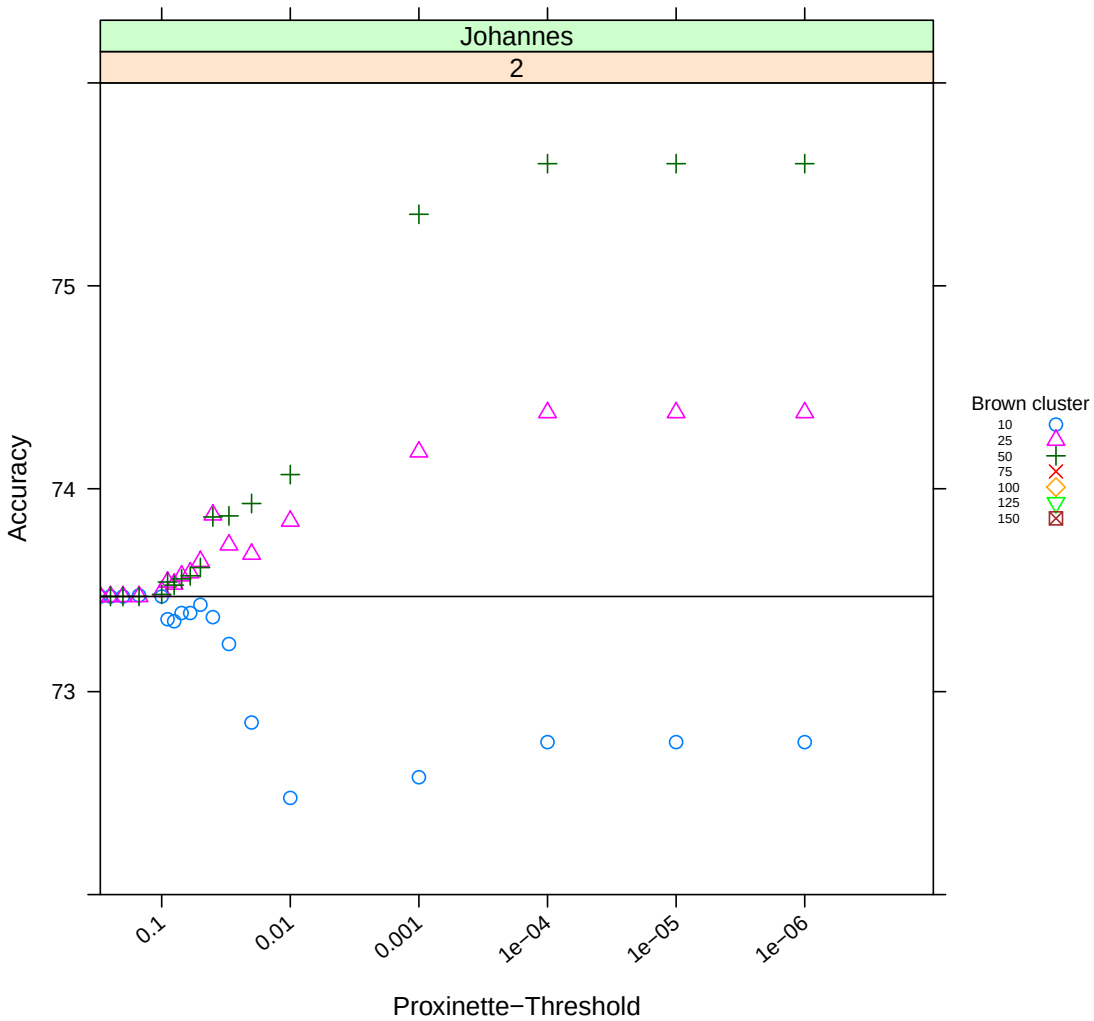
Results


Johannes
2
Accuracy
75
74
73
0.1
0.01
0.001
1e−04
1e−05
1e−06
Proxinette−Threshold
Brown cluster
10
25
50
75
100
125
150

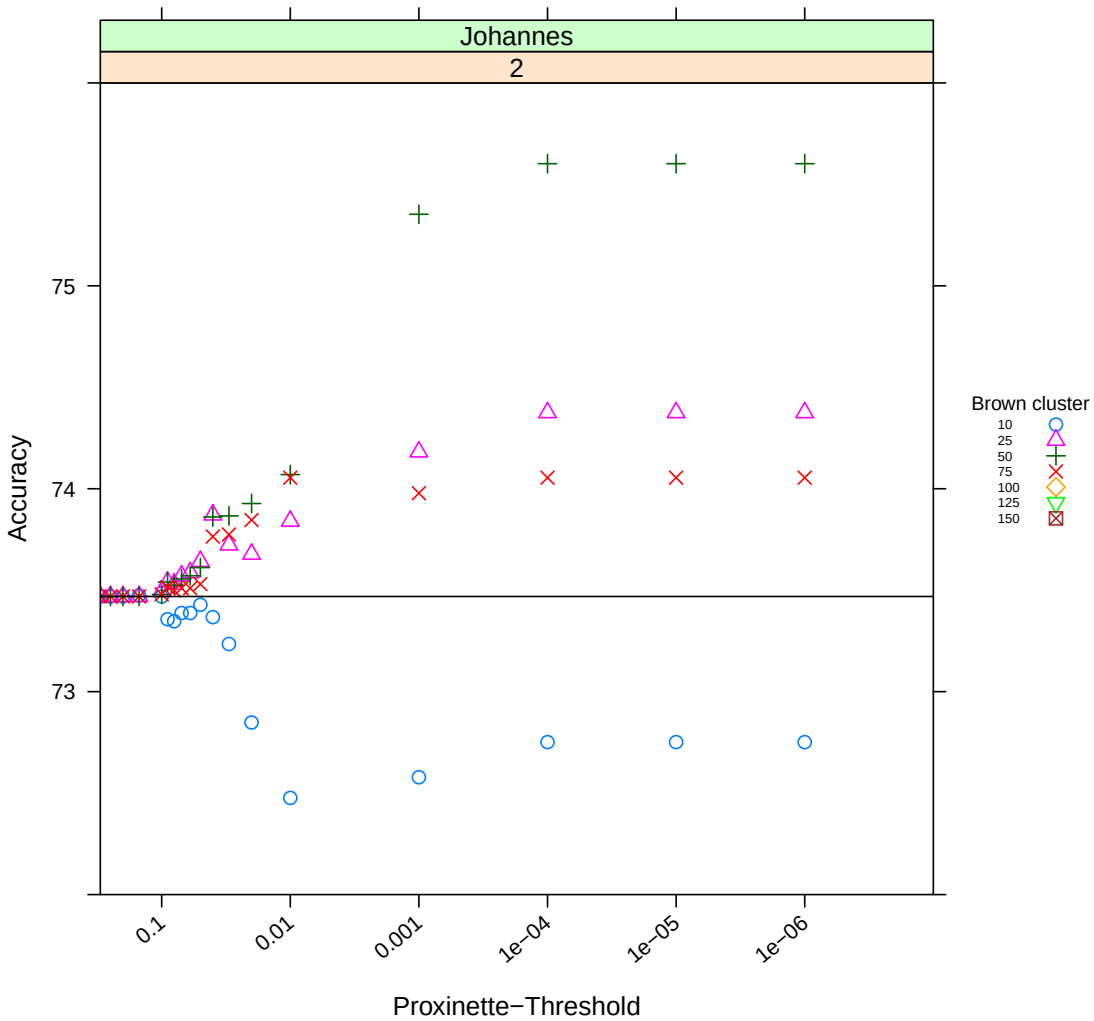
Results


Johannes
2
Accuracy
75
74
73
0.1
0.01
0.001
1e−04
1e−05
1e−06
Proxinette−Threshold
Brown cluster
10
25
50
75
100
125
150

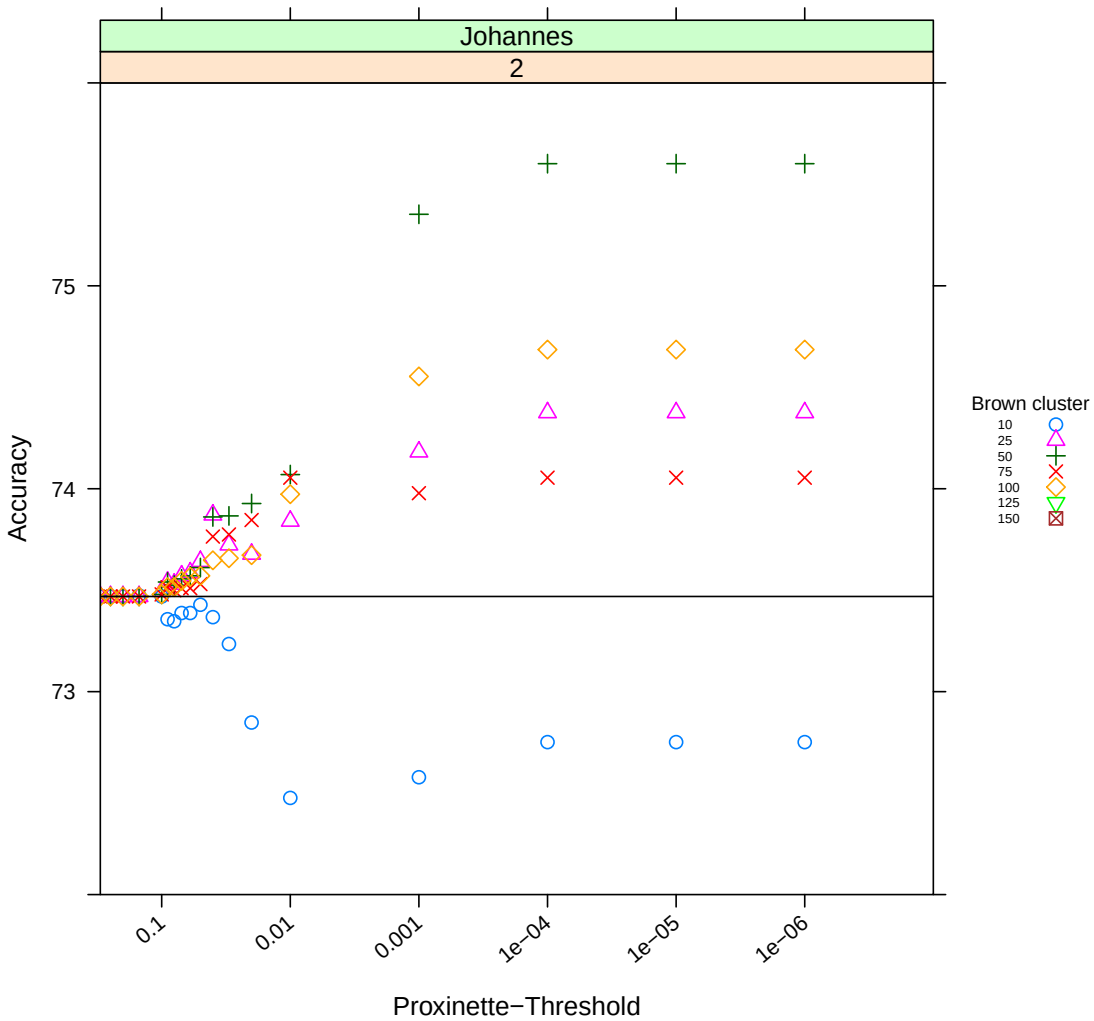
Results


Johannes
2
Accuracy
75
74
73
0.1
0.01
0.001
1e−04
1e−05
1e−06
Proxinette−Threshold
Brown cluster
10
25
50
75
100
125
150

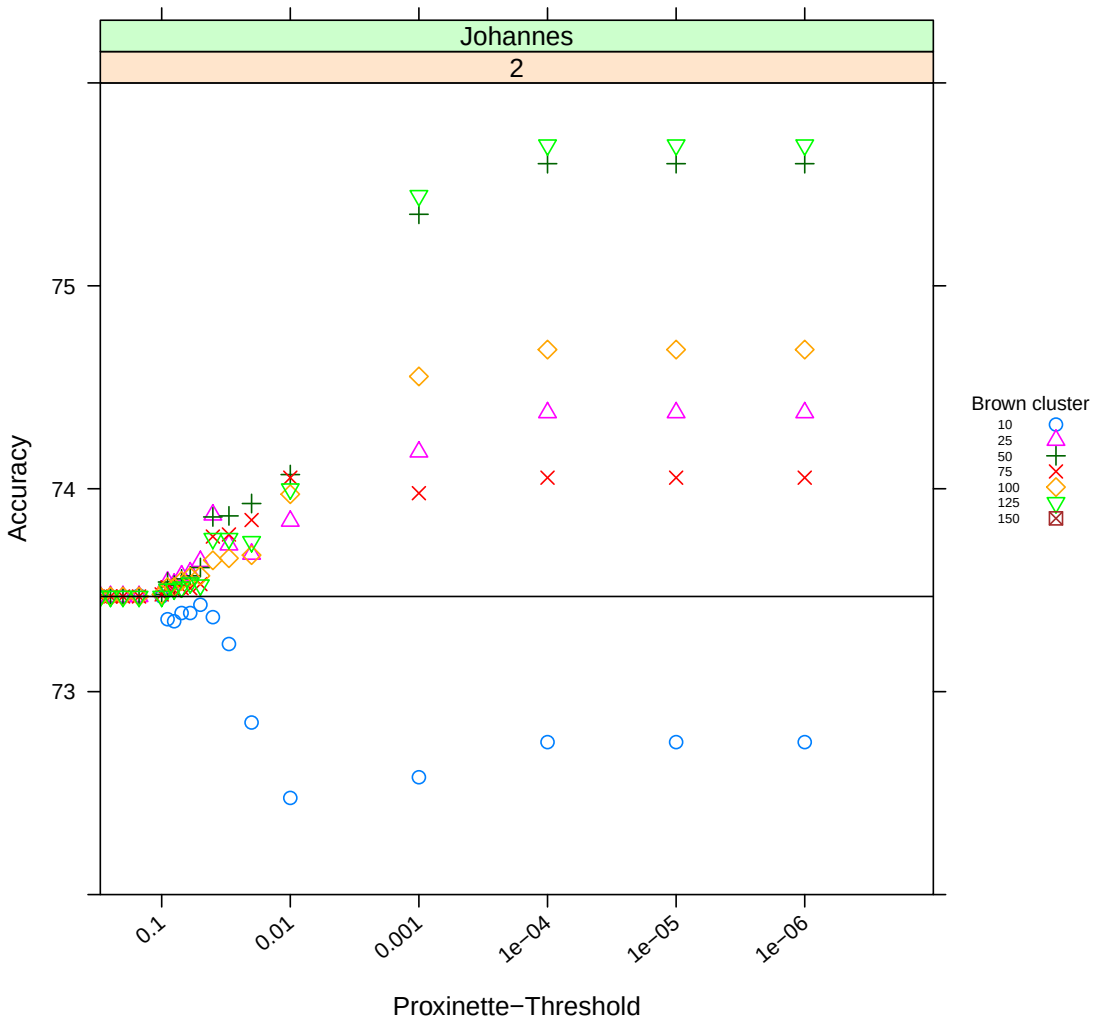
Results


Johannes
2
Accuracy
75
74
73
0.1
0.01
0.001
1e−04
1e−05
1e−06
Proxinette−Threshold
Brown cluster
10
25
50
75
100
125
150

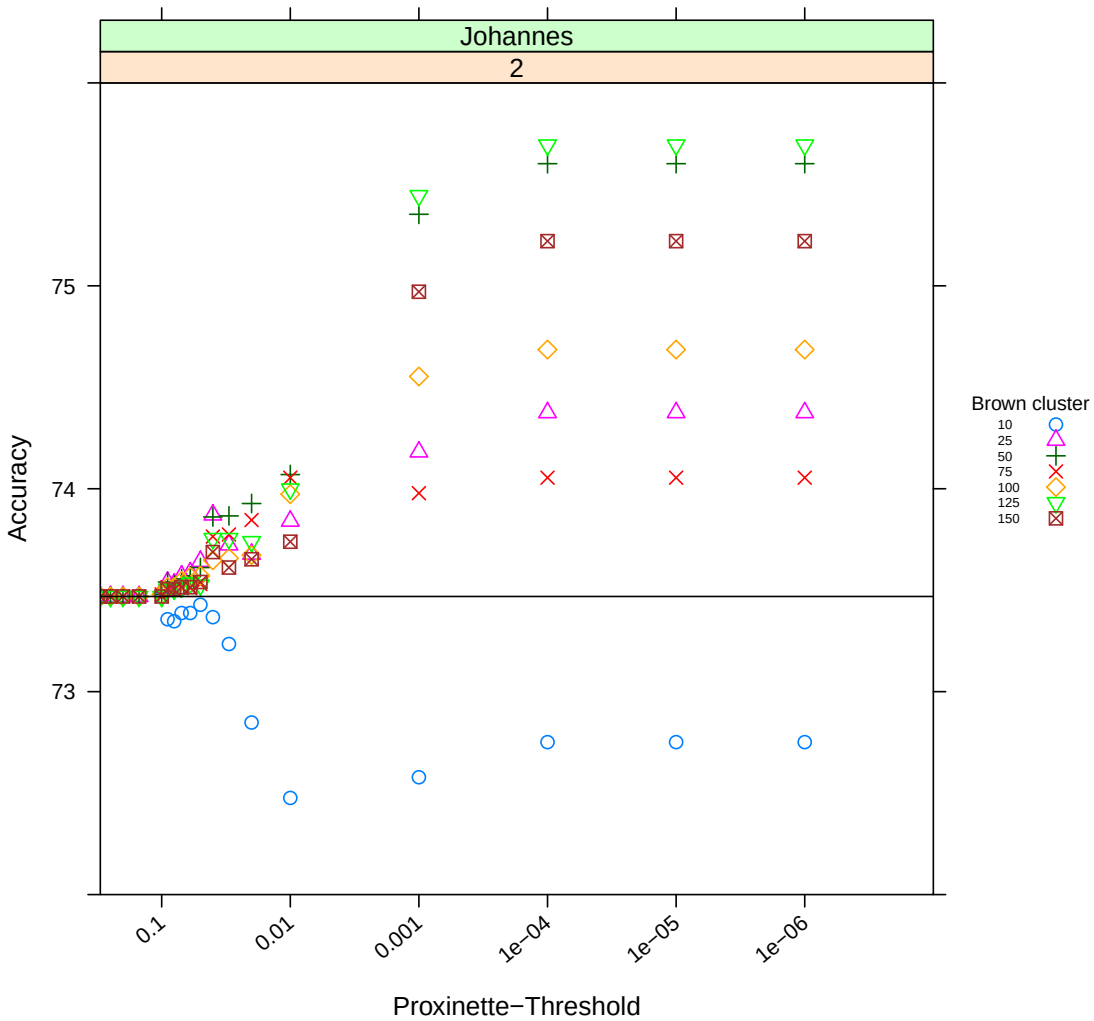
Results


Johannes
2
Accuracy
75
74
73
0.1
0.01
0.001
1e−04
1e−05
1e−06
Proxinette−Threshold
Brown cluster
10
25
50
75
100
125
150

Results


Johannes
2
Accuracy
75
74
73
0.1
0.01
0.001
1e−04
1e−05
1e−06
Proxinette−Threshold
Brown cluster
10
25
50
75
100
125
150

Results


Johannes
2
Accuracy
75
74
73
0.1
0.01
0.001
1e−04
1e−05
1e−06
Proxinette−Threshold
Brown cluster
10
25
50
75
100
125
150

Examples of Conflations from Johannes

Freq.	Type	Conflation	is spelling variant	translation
11	hochtijt	hochtid	x	'celebration'
12	hijr	hir	x	'here'
12	scole	schole	x	'shall'
13	scrift	schrifft	x	'writing'
17	efte	eft	x	'or'
17	sic	sick	x	'herself,himself...'
18	neman	nemande	x	'nobody'
21	uadere	vadere	x	'father'
27	comen	komen	x	'come'
30	lef	leff	x	'beloved'
43	scolen	scholen	x	'shall'
67	uader	vader	x	'father'
14	echter	echtes		'again' (diff. morphology)
15	ghecomen	komen		'come' (diff. inflection)
15	iiij	iii		Roman numerals
15	schare	hare		'cohort'; 'hair'
16	loue	louen		'believe, praise, promise' (diff. inflection)
22	jk	jodoch		'I';'but'
61	ioden	boden		'jews'; 'messengers'

Examples of Conflations from Johannes

Freq.	Type	Conflation	is spelling variant	translation
11	hochtijt	hochtid	x	'celebration'
12	hijr	hir	x	'here'
12	scole	schole	x	'shall'
13	scrift	schrifft	x	'writing'
17	efte	eft	x	'or'
17	sic	sick	x	'herself,himself...'
18	neman	nemande	x	'nobody'
21	uadere	vadere	x	'father'
27	comen	komen	x	'come'
30	lef	leff	x	'beloved'
43	scolen	scholen	x	'shall'
67	uader	vader	x	'father'
14	echter	echtes		'again' (diff. morphology)
15	ghecomen	komen		'come' (diff. inflection)
15	iiij	iii		Roman numerals
15	schare	hare		'cohort'; 'hair'
16	loue	louen		'believe, praise, promise' (diff. inflection)
22	jk	jodoch		'I';'but'
61	ioden	boden		'jews'; 'messengers'

Examples of Conflations from Johannes

Freq.	Type	Conflation	is spelling variant	translation
11	hochtijt	hochtid	x	'celebration'
12	hijr	hir	x	'here'
12	scole	schole	x	'shall'
13	scrift	schrifft	x	'writing'
17	efte	eft	x	'or'
17	sic	sick	x	'herself,himself...'
18	neman	nemande	x	'nobody'
21	uadere	vadere	x	'father'
27	comen	komen	x	'come'
30	lef	leff	x	'beloved'
43	scolen	scholen	x	'shall'
67	uader	vader	x	'father'
14	echter	echtes		'again' (diff. morphology)
15	ghecomen	komen		'come' (diff. inflection)
15	iiij	iii		Roman numerals
15	schare	hare		'cohort'; 'hair'
16	loue	louen		'believe, praise, promise' (diff. inflection)
22	jk	jodoch		'I';'but'
61	ioden	boden		'jews'; 'messengers'

Examples of Conflations from Johannes

Freq.	Type	Conflation	is spelling variant	translation
11	hochtijt	hochtid	x	'celebration'
12	hijr	hir	x	'here'
12	scole	schole	x	'shall'
13	scrift	schrifft	x	'writing'
17	efte	eft	x	'or'
17	sic	sick	x	'herself,himself...'
18	neman	nemande	x	'nobody'
21	uadere	vadere	x	'father'
27	comen	komen	x	'come'
30	lef	leff	x	'beloved'
43	scolen	scholen	x	'shall'
67	uader	vader	x	'father'
14	echter	echtes		'again' (diff. morphology)
15	ghecomen	komen		'come' (diff. inflection)
15	iiij	iii		Roman numerals
15	schare	hare		'cohort'; 'hair'
16	loue	louen		'believe, praise, promise' (diff. inflection)
22	jk	jodoch		'I';'but'
61	ioden	boden		'jews'; 'messengers'

Examples of Conflations from Johannes

Freq.	Type	Conflation	is spelling variant	translation
11	hochtijt	hochtid	x	'celebration'
12	hijr	hir	x	'here'
12	scole	schole	x	'shall'
13	scrift	schrifft	x	'writing'
17	efte	eft	x	'or'
17	sic	sick	x	'herself,himself...'
18	neman	nemande	x	'nobody'
21	uadere	vadere	x	'father'
27	comen	komen	x	'come'
30	lef	leff	x	'beloved'
43	scolen	scholen	x	'shall'
67	uader	vader	x	'father'
14	echter	echtes		'again' (diff. morphology)
15	ghecomen	komen		'come' (diff. inflection)
15	iiij	iii		Roman numerals
15	schare	hare		'cohort'; 'hair'
16	loue	louen		'believe, praise, promise' (diff. inflection)
22	jk	jodoch		'I';'but'
61	ioden	boden		'jews'; 'messengers'

Examples of Conflations from Johannes

Freq.	Type	Conflation	is spelling variant	translation
11	hochtijt	hochtid	x	'celebration'
12	hijr	hir	x	'here'
12	scole	schole	x	'shall'
13	scrift	schrifft	x	'writing'
17	efte	eft	x	'or'
17	sic	sick	x	'herself,himself...'
18	neman	nemande	x	'nobody'
21	uadere	vadere	x	'father'
27	comen	komen	x	'come'
30	lef	leff	x	'beloved'
43	scolen	scholen	x	'shall'
67	uader	vader	x	'father'
14	echter	echtes		'again' (diff. morphology)
15	ghecomen	komen		'come' (diff. inflection)
15	iiij	iii		Roman numerals
15	schare	hare		'cohort'; 'hair'
16	loue	louen		'believe, praise, promise' (diff. inflection)
22	jk	jodoch		'I';'but'
61	ioden	boden		'jews'; 'messengers'

Examples of Conflations from Johannes

Freq.	Type	Conflation	is spelling variant	translation
11	hochtijt	hochtid	x	'celebration'
12	hijr	hir	x	'here'
12	scole	schole	x	'shall'
13	scrift	schrifft	x	'writing'
17	efte	eft	x	'or'
17	sic	sick	x	'herself,himself...'
18	neman	nemande	x	'nobody'
21	uadere	vadere	x	'father'
27	comen	komen	x	'come'
30	lef	leff	x	'beloved'
43	scolen	scholen	x	'shall'
67	uader	vader	x	'father'
14	echter	echtes		'again' (diff. morphology)
15	ghecomen	komen		'come' (diff. inflection)
15	iiij	iii		Roman numerals
15	schare	hare		'cohort'; 'hair'
16	loue	louen		'believe, praise, promise' (diff. inflection)
22	jk	jodoch		'I';'but'
61	ioden	boden		'jews'; 'messengers'

Examples of Conflations from Johannes

Freq.	Type	Conflation	is spelling variant	translation
11	hochtijt	hochtid	x	'celebration'
12	hijr	hir	x	'here'
12	scole	schole	x	'shall'
13	scrift	schrifft	x	'writing'
17	efte	eft	x	'or'
17	sic	sick	x	'herself,himself...'
18	neman	nemande	x	'nobody'
21	uadere	vadere	x	'father'
27	comen	komen	x	'come'
30	lef	leff	x	'beloved'
43	scolen	scholen	x	'shall'
67	uader	vader	x	'father'
14	echter	echtes		'again' (diff. morphology)
15	ghecomen	komen		'come' (diff. inflection)
15	iiij	iii		Roman numerals
15	schare	hare		'cohort'; 'hair'
16	loue	louen		'believe, praise, promise' (diff. inflection)
22	jk	jodoch		'I';'but'
61	ioden	boden		'jews'; 'messengers'

Examples of Conflations from Johannes

Freq.	Type	Conflation	is spelling variant	translation
11	hochtijt	hochtid	x	'celebration'
12	hijr	hir	x	'here'
12	scole	schole	x	'shall'
13	scrift	schrifft	x	'writing'
17	efte	eft	x	'or'
17	sic	sick	x	'herself,himself...'
18	neman	nemande	x	'nobody'
21	uadere	vadere	x	'father'
27	comen	komen	x	'come'
30	lef	leff	x	'beloved'
43	scolen	scholen	x	'shall'
67	uader	vader	x	'father'
14	echter	echtes		'again' (diff. morphology)
15	ghecomen	komen		'come' (diff. inflection)
15	iiij	iii		Roman numerals
15	schare	hare		'cohort'; 'hair'
16	loue	louen		'believe, praise, promise' (diff. inflection)
22	jk	jodoch		'I';'but'
61	ioden	boden		'jews'; 'messengers'

Overview

- Spelling Variation
- Modernization
- Regularization
- Unsupervised, Language-independent Regularization
  - The Approach
  - Evaluation
- Conclusion

Conclusion

- Regularization as an alternative to Modernization
  (to reduce spelling variation in non-standard texts)

Conclusion

- Regularization as an alternative to Modernization
  (to reduce spelling variation in non-standard texts)
- Regularization approach that works on unlabeled text

Conclusion

- Regularization as an alternative to Modernization
  (to reduce spelling variation in non-standard texts)
- Regularization approach that works on unlabeled text
- Based on two simple assumptions
  (and therefore language-independent)
  - Formal similarity
  - Contextual similarity

Conclusion

- Regularization as an alternative to Modernization
  (to reduce spelling variation in non-standard texts)
- Regularization approach that works on unlabeled text
- Based on two simple assumptions
  (and therefore language-independent)
  - Formal similarity
  - Contextual similarity
- Evaluation using POS tagging accuracy

Future Directions

- Evaluation

Future Directions

- Evaluation
  - Evaluate with different annotations/tasks

Future Directions

- Evaluation
  - Evaluate with different annotations/tasks
  - Evaluate the whole conflation set

Future Directions

- Evaluation
  - Evaluate with different annotations/tasks
  - Evaluate the whole conflation set
- Regularization approach

Future Directions

- Evaluation
  - Evaluate with different annotations/tasks
  - Evaluate the whole conflation set
- Regularization approach
  - Explore alternative ways to model string and context similarity

Future Directions

- Evaluation
  - Evaluate with different annotations/tasks
  - Evaluate the whole conflation set
- Regularization approach
  - Explore alternative ways to model string and context similarity
  - Exploit systematicity of spelling variation

Thank iu for iwe attention.

Thank iu for iwe attention.

Fabian Barteld
fabian.barteld@uni-hamburg.de

References I

Baldwin, T., de Marneffe, M. C., Han, B., Kim, Y.-B., Ritter, A., and Xu, W. (2015).
Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition.
In Proceedings of the ACL 2015 Workshop on Noisy User-generated Text (W-NUT), Beijing, China, pages 126–135.

Baron, A., Rayson, P., and Archer, D. (2009).
Word frequency and key word statistics in corpus linguistics.
Anglistik, 20(1):41–67.

References II


Bollmann, M., Dipper, S., Krasselt, J., and Petran, F. (2012).
Manual and Semi-automatic Normalization of Historical Spelling—Case Studies from Early New High German.
In Proceedings of the 11th Conference on Natural Language Processing (KONVENS 2012), LThist 2012 workshop, pages 342–350.


Brown, P. F., deSouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992).
Class-based N-gram Models of Natural Language.
Computational linguistics, 18(4):467–479.


Hathout, N. (2014).
Phonotactics in morphological similarity metrics.
Language Sciences, 46, Part A:71–83.

References III

Jurish, B. (2013).
Canonicalizing the Deutsches Textarchiv.
In Hafemann, I., editor, Perspektiven einer corpusbasierten historischen Linguistik und Philologie. Internationale Tagung des Akademienvorhabens „Altägyptisches Wörterbuch“ an der Berlin-Brandenburgischen Akademie der Wissenschaften, 12.–13. Dezember 2011, pages 235–244.

Kestemont, M., Daelemans, W., and Pauw, G. D. (2010).
Weigh your words—memory-based lemmatization for Middle Dutch.
Literary and Linguistic Computing, 25(3):287–301.

References IV

Ljubešić, N., Erjavec, T., and Fišer, D. (2014).
Standardizing Tweets with Character-Level Machine Translation.
In Gelbukh, A., editor, 15th International Conference CICLing 2014, Kathmandu, Nepal, April 6-12, 2014, Proceedings, Part II, number 8404 in Lecture Notes in Computer Science, pages 164–175. Springer, Berlin, Heidelberg.

Logačev, P., Goldschmidt, K., and Demske, U. (2014).
POS-Tagging Historical Corpora: The Case of Early New High German.
In Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13), pages 103–112.

References V


Saloot, M. A., Idris, N., Shuib, L., Raj, R. G., and Aw, A. (2015).
Toward Tweets Normalization Using Maximum Entropy.
Proceedings of the ACL 2015 Workshop on Noisy User-generated Text (W-NUT), Beijing, China, pages 19–27.


Sanders, W. (1982).
Sachsensprache, Hansesprache, Plattdeutsch. Sprachgeschichtliche Grundzüge des Niederdeutschen.
Vandenhoeck + Ruprecht, Göttingen.


Scherrer, Y. and Erjavec, T. (2016).
Modernising historical Slovene words.
Natural Language Engineering, First View (available on CJO2015):1–25.

References VI

Schmid, H. and Laws, F. (2008).
Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging.
In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 777–784.