Search methods for documents in non-standard spelling

Andrea Ernst-Gerlach

Thomas Pilz

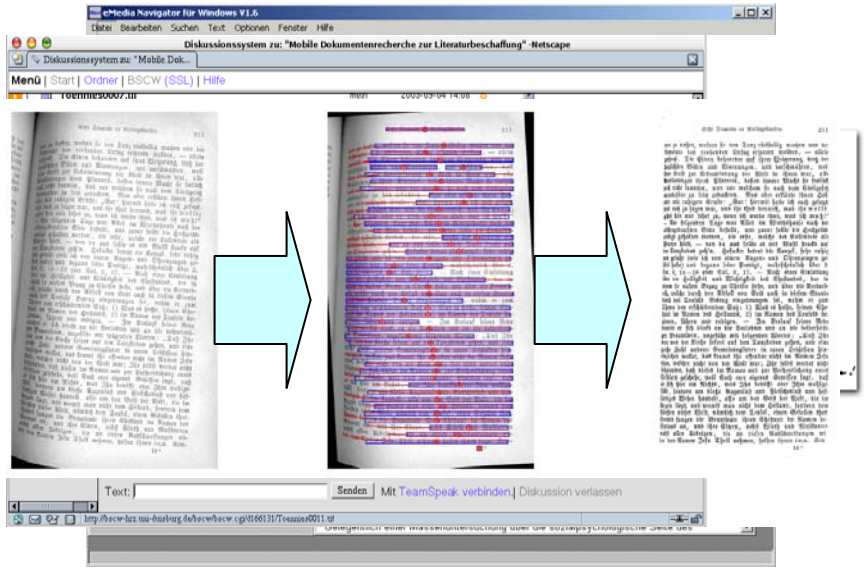
University of Duisburg-Essen

Overview

- Project „RSNSR“
- German spelling variation
- Our approaches in detail
- Conclusion
- Outlook

Origins of RSNSR

- „Projekt Nietzsche-CD“
- „A web-based system for assisted literature research“
- Partial text recognition of German Fraktur documents; document image de-warping



What is RSNSR about?

- Development of a search-engine for historical documents where
  - users are not required to be language experts
  - specific rule-sets and metrics ensure reliable instant retrieval results

Am abgewichenen Sonnabend haben Jhro Königl. Hoheit, die Prinzeßin Amalia, an das Königliche Haus, und viele hohen Standespersohnen (...) und die Printzeßin Amalia

Berlinische Privilegirte Zeitung, 23.01.1748

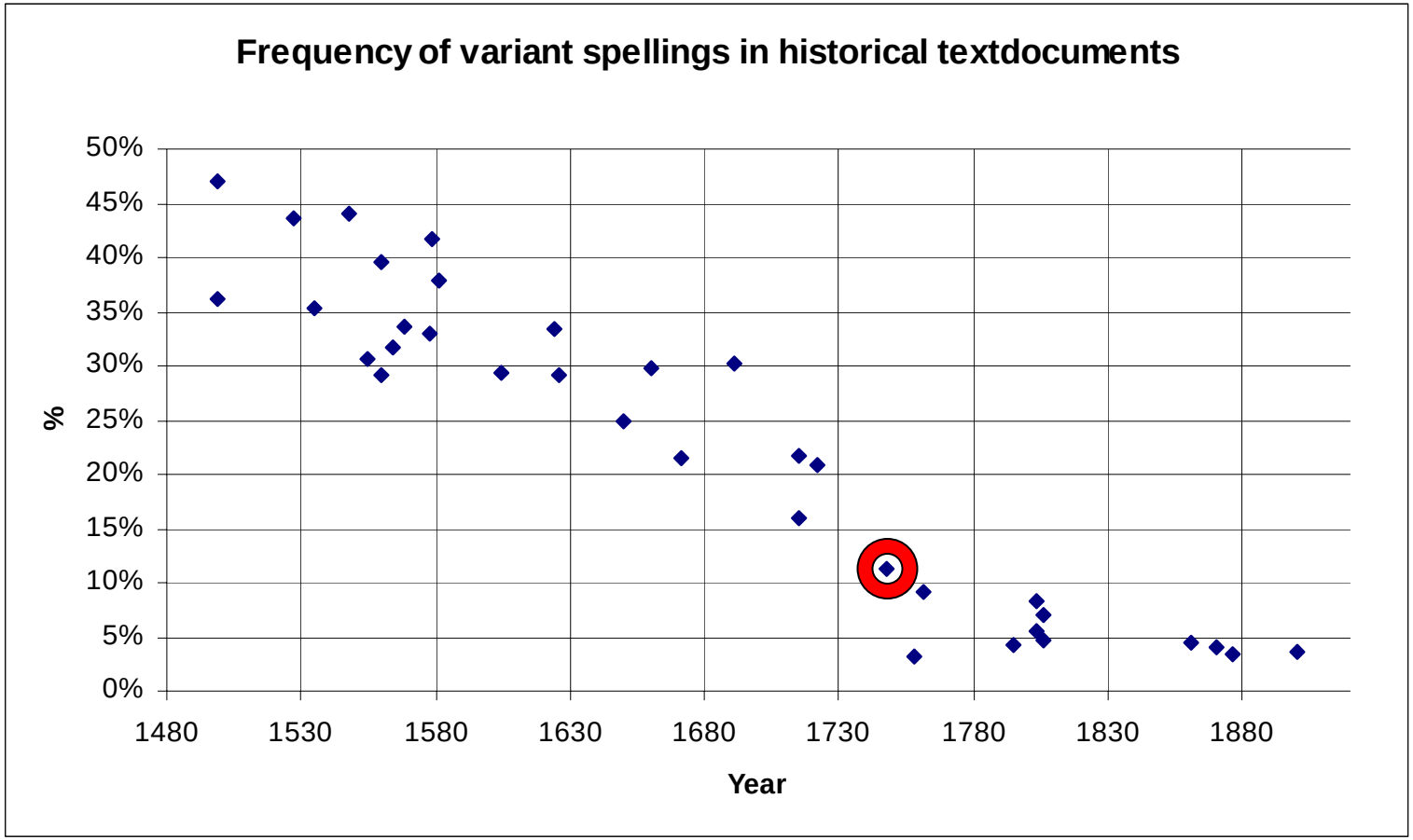
German documents prior to 1901 are not orthographically standardized.

OCR of German Fraktur-font is error-prone

Nach uns zugegangenen Pnvatmeldungen waien die veiheeienden Wirkungen dei Epidemie in diesem Jahre bedeutend sclrwachei als im Voi-jahie Sie \\uide ^on den MeLLapilgein nach Palastina eingeschleppt

Altneuland, Heft 1, January 1904

Variant spellings



What are we dealing with?

- Different variants for one word

  Erkenntniss Erkendtnüß

  Erkenntnis : Erkenntniß Erkändtnuß

  Erkandtnüß

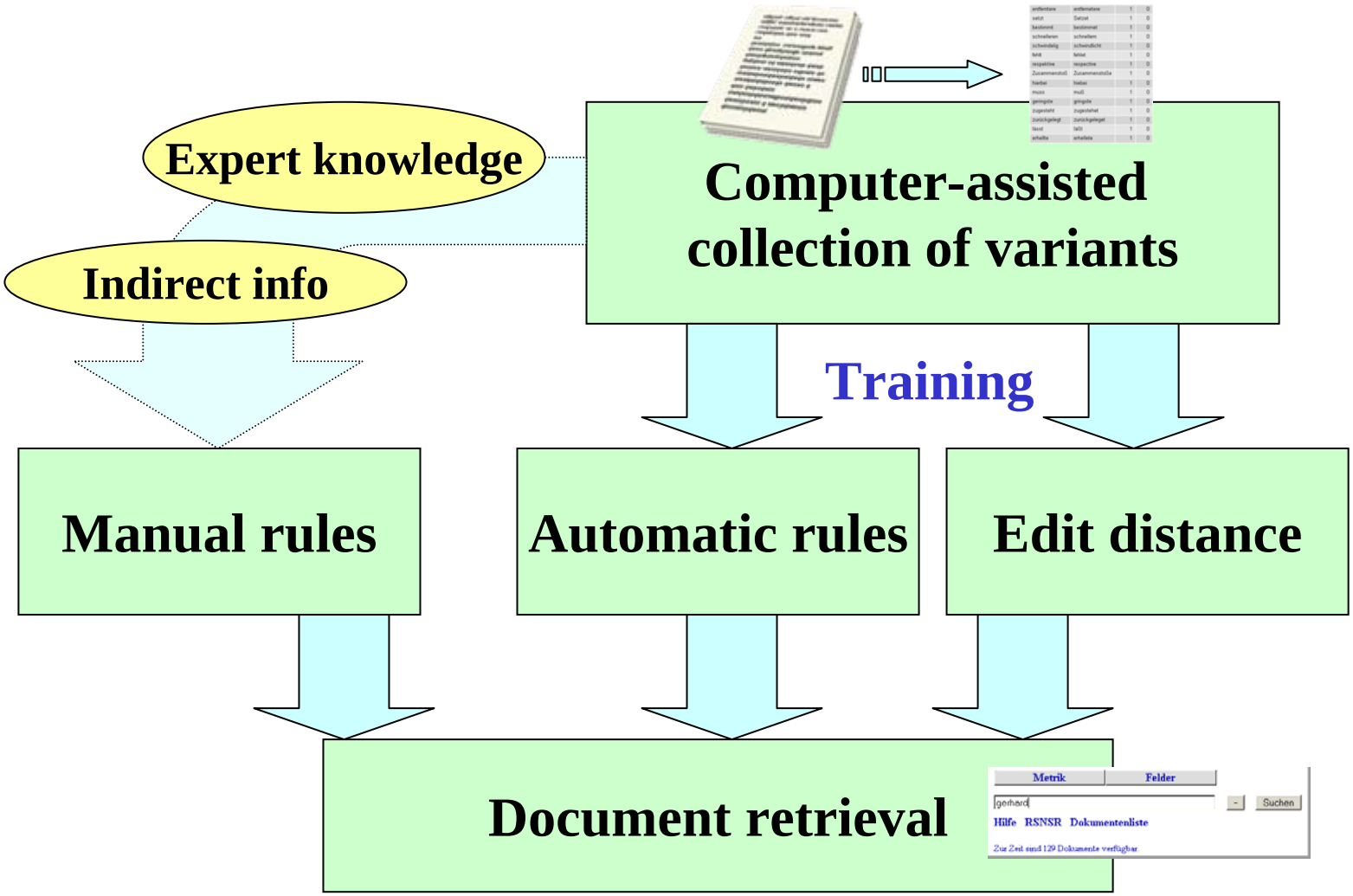
- Highly variable words

  Dompropst : Thuembbröbst

- Phonetic and graphematic variation (time-dependend)

  allzu : allzvo

How do we proceed?



Distance metrics

- Commonly used in *dialectrometry*
- Many different methods

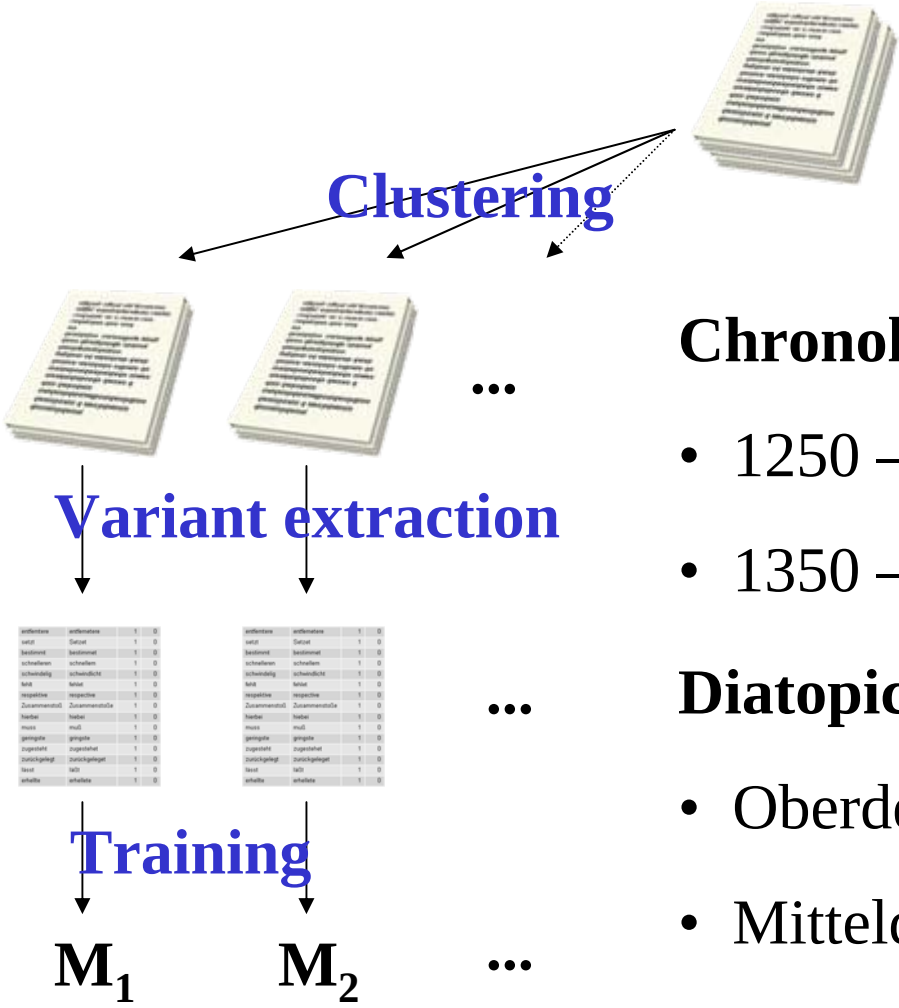
in %	lemma - variant						variant - lemma					
	P	R_1	R_2	R_3	R_4	R_5	P	R_1	R_2	R_3	R_4	R_5
Ristad	74	60	82	85	88	90	87	83	90	92	92	93
Editex	68	56	69	75	81	85	80	71	80	83	86	90
Meyer-Wilde	60	47	60	72	76	80	71	61	74	78	83	84
Jaro	45	30	44	57	60	63	67	56	69	75	78	80
Bigram	50	35	47	55	61	70	78	71	79	85	87	88
Levenshtein	53	37	52	64	71	76	74	65	75	78	83	85

String Edit Distance

- Ristad & Yianilos 1996
- Learning string edit distance
- Uses EM-Algorithm
- Table of edit costs: |alphabet| x |alphabet|

h	_	0,01414282
d	t	0,01379537
c	_	0,01245433
_	h	0,00863217
_	n	0,00853391
n	_	0,00729049
_	t	0,00702836
k	c	0,00691858
e	i	0,0055387
f	v	0,00553486
_	s	0,00427657
b	p	0,00415115
g	c	0,00415115

Metric classification



Chronological classification (T)

- 1250 – 1350
- 1350 – 1450
- 1450 – 1650
- 1650 – 1900

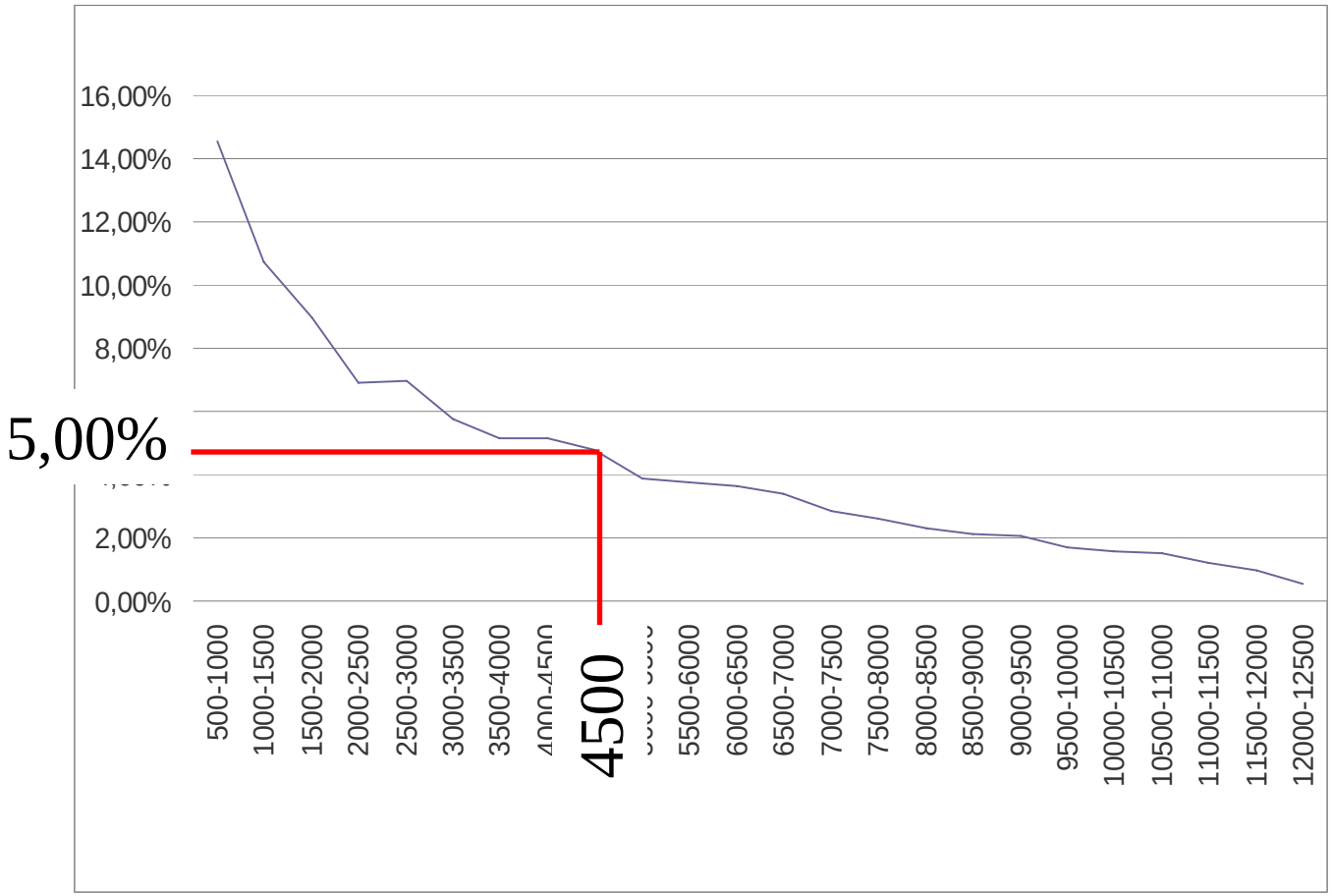
Diatopical classification (L)

- Oberdeutsch
- Niederdeutsch
- Mitteldeutsch

Special class (OCR)

And...why?

- Saturation of variation



And...why?

- Specific metrics can increase efficiency

Application of different metrics to historical English documents

Std. Levenshtein	German metric			English metric
	15th-16th	13th-16th	13th-15th	
89%	91%	93%	94%	100%

Generate rule core

- Training set of triplets
  - Contemporary word form
  - Historic word form
  - Collection frequency
- Find necessary transformations e. g.

  unnütz → unnuts

  rule cores: ^unn(ü,u)t t(z,s)$

Generate rule candidates

- Successively adding context to rule cores
  - e. g. unnütz → unnuts rule core: ^unn(ü,u)t

    ü → u nü → nu üt →ut nüt→ nut
- Abstracting of context
  - Consonant sounds (C) / Vowel sounds (V)

    e. g. Cü → Cu
  - Word beginning (^) / ending ($)

    e. g. z$ →s$

Rule Pruning: PRISM Algorithm

- PRISM
  - Classifies set of instances into set of classes
  - Instances are fixed sets of attributes
  - Tries to generate high precision values for each class C identifying instances belonging to C
- Extension necessary
  - Perfect rules on this data set do not generalise unseen words
  - Generalisation / specialisation relationships between rule antecedents

Rule Pruning: PRISM Extension

- Generate negative examples
  - applying rule candidates on evidences
  - e. g.

evidence:	ab – aab
rule candidate:	a – az
→ negative example:	ab – azb

- Sort instances by rules
- Calculate
  - occurence frequency q_i
  - precision p_i
- Remove all instances where $p_i < p_{min} \vee q_i < q_{min}$

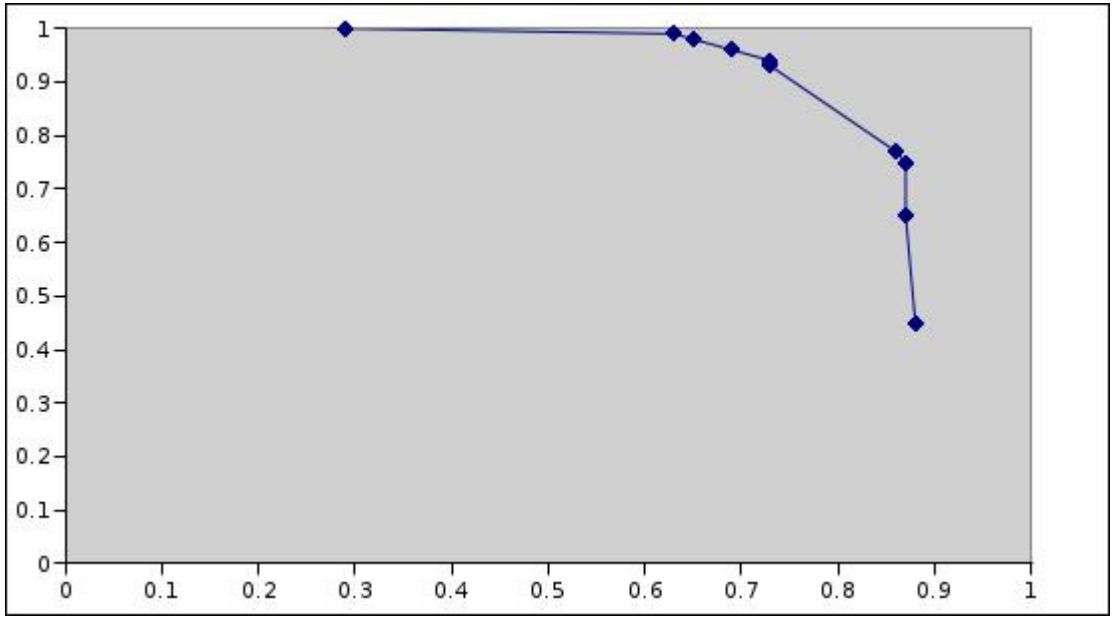
Evaluation results

- Recall 0.88
- Precision 0.45
- Most Frequently used rules:

Rules	Frequency	Precision	Examples
t → th	113	0.99	Einteilung – Eintheilung
ä → ae	42	0.98	Ämter - Aemter
s → ß	32	0.94	aus – auß
k → c	24	0.96	Kollegien – Collegien
ü → ue	19	0.86	Übertragung - Uebertragung
ä→ ai	18	1	souverän - souverain

Evaluation results (2)

- Recall/Precision for different values p_{min}



Conclusion

- Generate historic variants for search terms
- Three approaches:
  - Distance measures
  - Automatic rule generation
  - Manual rule generation
- Good retrieval results
- Working application “Nietzsche-Archiv”

Outlook

- Enhancement of automatic rule generation and distance metric algorithm
- Evaluation of rule generation in contrast to distance-measures
- Application to other languages and cross-language evaluation
- Automatic variant extraction
- Automatic text classification with trained metrics

Thank you for your interest!

Any questions?