Vom Bleiletterndruck zum Forschungskorpus

Neue Dienstleistungen jenseits der OCR

Elisabeth Klein | Dr. Klaus T. Weber
BID-Kongress 2016
17. April 2016

UNIVERSITÄTS BIBLIOTHEK MAINZ

JG|U



Überblick

1 | Faktoren der Zusammenarbeit

2 | Volltextsammlungen vs. Korpora

3 | Methoden und Test-Design

4 | Ausblick: Post-Processing

Volltextsammlung

- Menge von Volltexten in Text oder XML-Format
- meist unstrukturiert
- variable Erkennungsrate
- Metadaten optional vorhanden
- keine Annotationen enthalten

eingeschränkt durchsuchbar,

nur für qualitative Forschung geeignet (wenn überhaupt...)

Korpus

- kuratierte Sammlung strukturierter Volltexte in Text oder XML-Format
- Metadaten für alle Texte
- stabile Erkennungsrate durch Wörterbuchabgleich
- annotiert

Negativsuchen

für qualitative & quantitative Forschung geeignet

Volltextsammlung

- Menge von Volltexten in Text oder XML-Format
- meist unstrukturiert
- variable Erkennungsrate
- Metadaten optional vorhanden
- keine Annotationen enthalten

eingeschränkt durchsuchbar,

nur für qualitative Forschung geeignet (wenn überhaupt...)

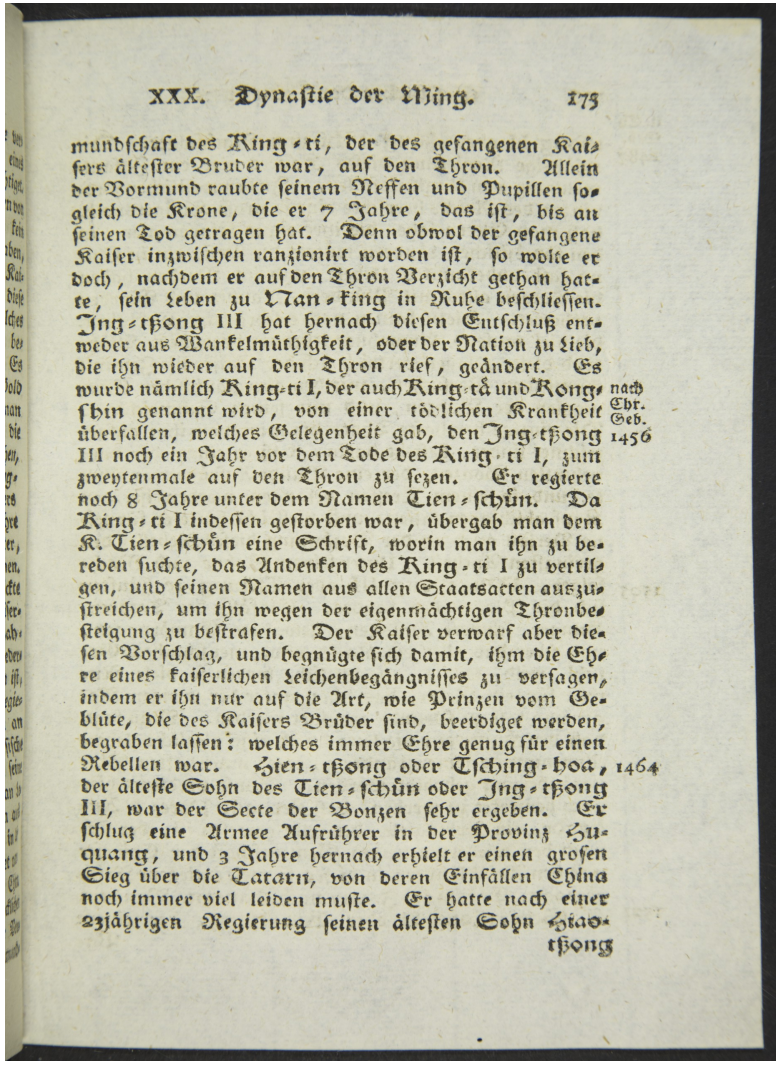
Korpus

- kuratierte Sammlung strukturierter Volltexte in Text oder XML-Format
- Metadaten für alle Texte
- stabile Erkennungsrate durch Wörterbuchabgleich
- annotiert

Negativsuchen,

für qualitative & quantitative Methoden geeignet

Textsammlung



XXX. Dynastie der Ming. 173

mundschaft des King-ti, der des gefangenen Kaisers ältester Bruder war, auf den Thron. Allein der Vormund raubte seinem Neffen und Pupillen sogleich die Krone, die er 7 Jahre, das ist, bis an seinen Tod getragen hat. Denn obwol der gefangene Kaiser inzwischen ranzionirt worden ist, so wolte er doch, nachdem er auf den Thron Verzicht gethan hatte, sein Leben zu Nan-king in Ruhe beschliessen. Jng-tßong III hat hernach diesen Entschluß entweder aus Wankelmüthigkeit, oder der Nation zu Lieb, die ihn wieder auf den Thron rief, geändert. Es wurde nämlich King-ti I, der auch King-tå und Kongshin genannt wird, von einer tödlichen Krankheit nach Chr. Geb. 1456 überfallen, welches Gelegenheit gab, den Jng-tßong III noch ein Jahr vor dem Tode des King-ti I, zum zweytenmale auf den Thron zu sezen. Er regierte noch 8 Jahre unter dem Namen Tien-schün. Da King-ti I indessen gestorben war, übergab man dem K. Tien-schün eine Schrift, worin man ihn zu bereden suchte, das Andenken des King-ti I zu vertilgen, und seinen Namen aus allen Staatsacten auszustreichen, um ihn wegen der eigenmächtigen Thronbesteigung zu bestrafen. Der Kaiser verwarf aber diesen Vorschlag, und begnügte sich damit, ihm die Ehre eines kaiserlichen Leichenbegängnisses zu versagen, indem er ihn nur auf die Art, wie Prinzen vom Geblüte, die des Kaisers Brüder sind, beerdiget werden, begraben lassen: welches immer Ehre genug für einen Rebellen war. Hien-tßong oder Tsching-hoa, 1464 der älteste Sohn des Tien-schün oder Jng-tßong III, war der Secte der Bonzen sehr ergeben. Er schlug eine Armee Aufrührer in der Provinz Huquang, und 3 Jahre hernach erhielt er einen grosen Sieg über die Tatarn, von deren Einfällen China noch immer viel leiden muste. Er hatte nach einer 23jährigen Regierung seinen ältesten Sohn Hiaotßong

- Projekt „Vor der Kulturgeschichte“ (FSP „Historische Kulturwissenschaften“ der JGU)
- 11 Monographien à 200-600 Seiten
- Druckdatum vor 1850: manufakturierte Frakturtypen
- ABBYY Finereader 11

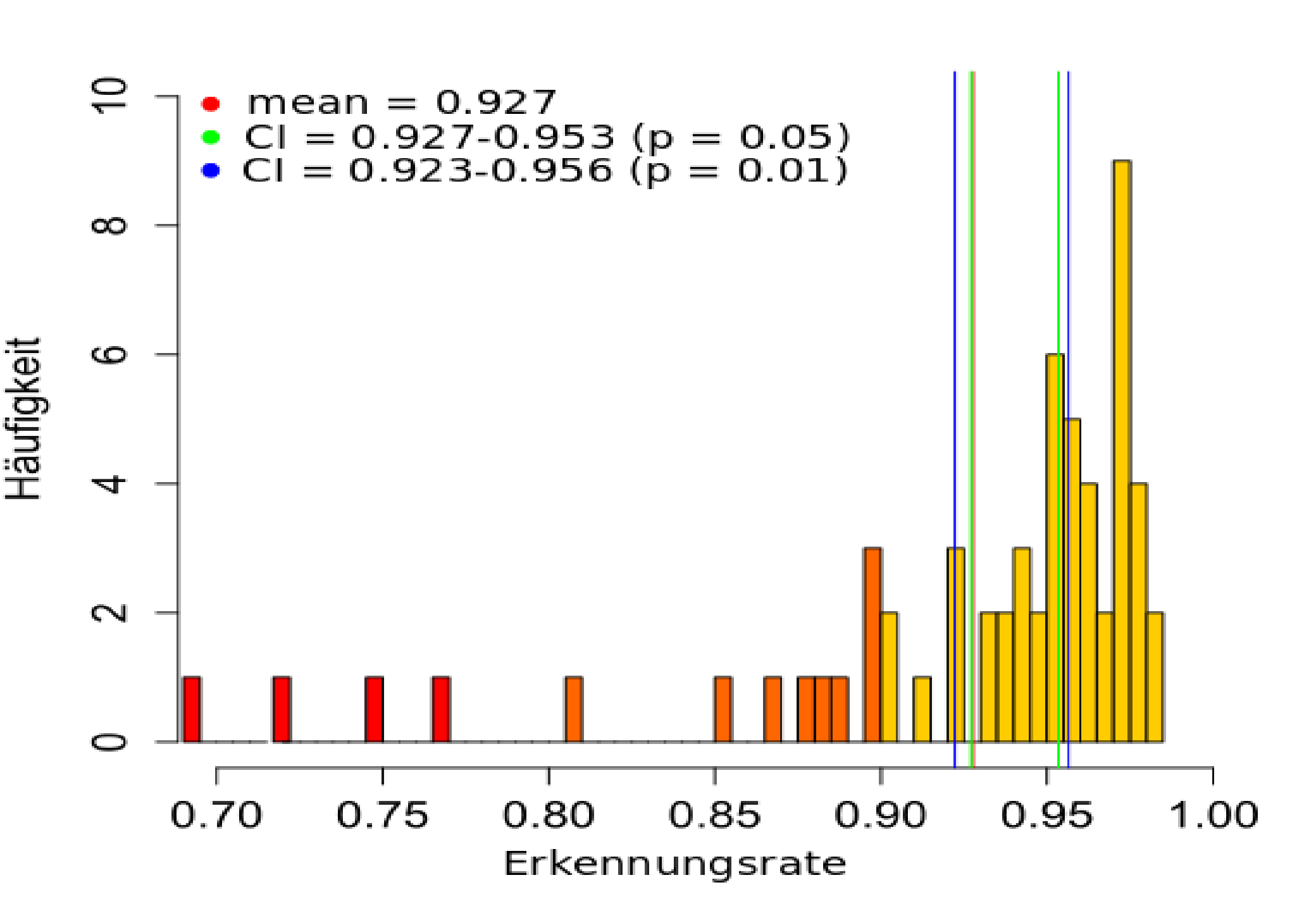
Stichprobe

- Verfahren orientiert an Stäcker (2013)

- Anpassung des Stichprobenverfahrens
  - randomisierte Stichprobe von 60 Seiten aus dem Gesamtkonvolut (83.542 Zeichen)
  - **keine** Berücksichtigung von Verschmutzungen, Durchstreichungen, Annotationen, Marginalien auf den Testseiten
  - Test unter realistischen Bedingungen

Ground Truth

- 60 Seiten manuelle Transkription durch wissenschaftliche Mitarbeiter des Projekts als Vergleichsprobe
- Herausforderung: Software zum Abgleich Stichprobe/Ground Truth finden
- Abgleich per Levenshtein-Algorithmus


mean = 0.927
CI = 0.927-0.953 (p = 0.05)
CI = 0.923-0.956 (p = 0.01)
Häufigkeit
0
2
4
6
8
10
0.70
0.75
0.80
0.85
0.90
0.95
1.00
Erkennungsrate

Ergebnisse

veränderter Stichprobenansatz:

- Verschmutzungen, Widerdrucke, Annotationen etc. führen NICHT zu schlechteren Erkennungsraten
- fremdsprachige Abkürzungen in Kombination mit Typenwechsel und wenig bekannte Lexik/Orthographien erzeugen Erkennungsfehler
- zeigt, ob Software alltagstaugliche Ergebnisse liefert

Ergebnisse

Auswirkungen für Forschungsprojekte:

- keine phraseologischen oder Negativsuchen möglich
- Methodenmix zur wissenschaftlichen Bearbeitung des Textes nötig
  - Fehleranalyse nach der OCR zur Abschätzung kritischer Textbereiche
  - Verbindung von close & distant reading
- entstandene Volltexte sind nicht zur Bearbeitung mit quantitativen Methoden geeignet

Chancen des Post-Processing

- historische Wörterbücher
- language models
- tokenizer
- lemmatizer

} Korpuserstellung

Vielen Dank !

UNIVERSITÄTS
BIBLIOTHEK
MAINZ

JG|U