Hes·so
Haute Ecole Spécialisée
de Suisse occidentale
Fachhochschule Westschweiz
University of Applied Sciences and Arts
Western Switzerland

UNI FR
UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

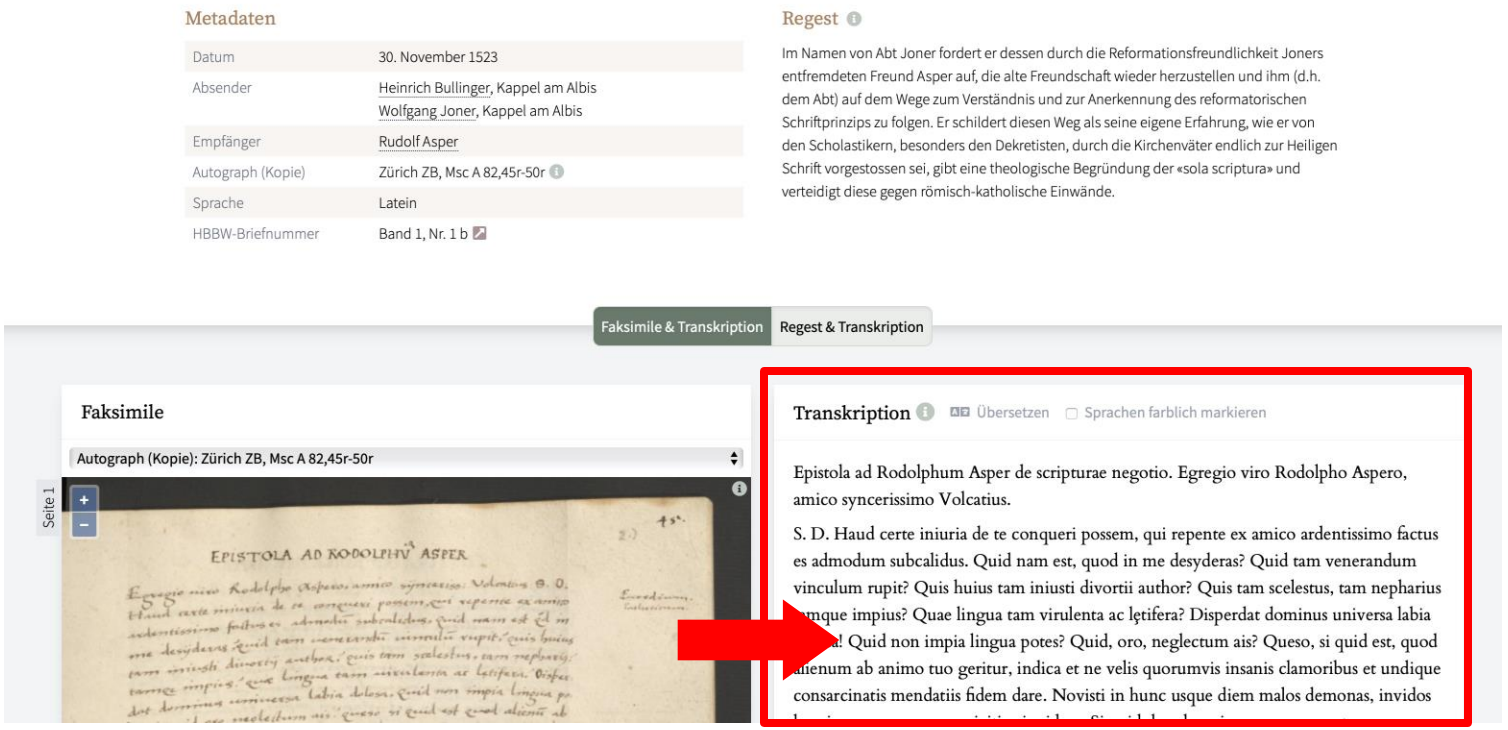
Vielschreiber und Wenigschreiber: Herausforderungen für die automatische Handschrifterkennung

Bullinger Digital
Zürich, 24.2.2023

Prof. Dr. Andreas Fischer
andreas.fischer@hefr.ch

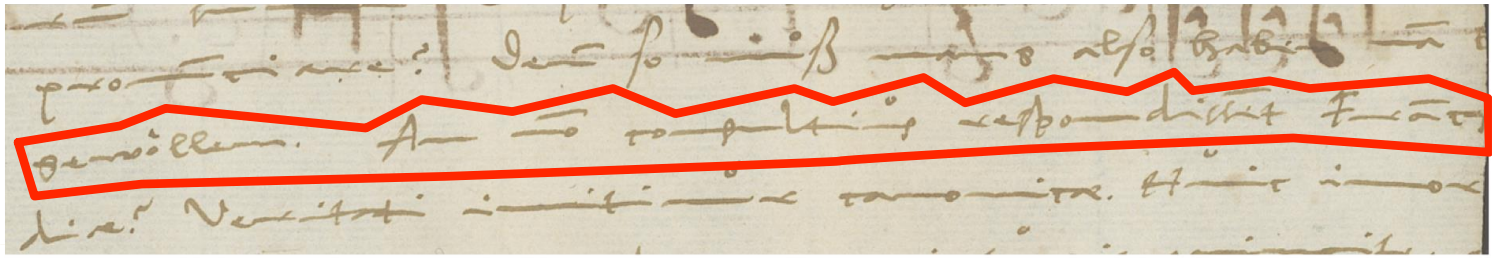
Ziel der Handschrifterkennung

- Ein Schritt weiter als das Scannen der Briefe und Erfassen von Metadaten: Automatischer **Zugang zum textuellen Inhalt** der Briefe in maschinen-lesbarer Form.



Aufgaben: Layout-Analyse und Texterkennung

1. **Layout-Analyse:** Finden von Textregionen, d.h. Titel, Spalten, Paragraphen, Zeilen und Wörter; typischerweise Zeilen.

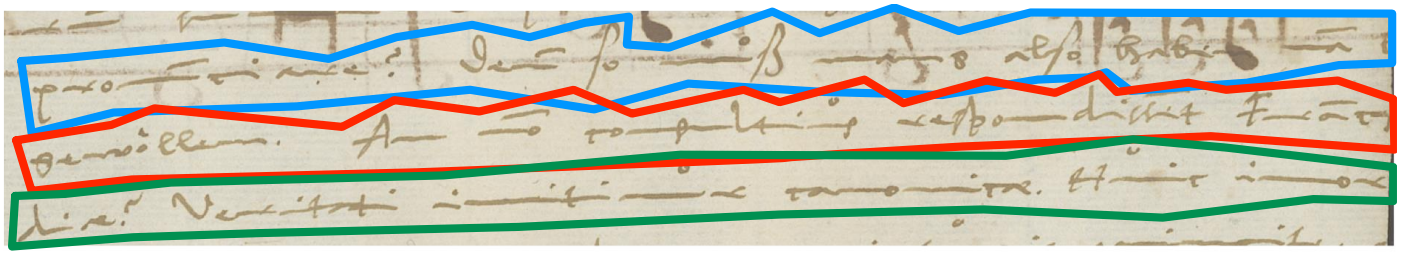


2. **Texterkennung:** Erkennen der Buchstabensequenz im Bild in maschinen-lesbarer Form.

pronunciaret denn sparnuß michs also habe man
gewöllen. An non consultius respondissent Franc
veritati invitimur canonicae. Hinc immor

Methode: Lernen anhand von Beispielen

- Maschinelles Lernen erfordert **Lernbeispiele** in Form von (manuellen oder halb-automatischen) Annotationen:
  - Für die Layout-Analyse sind dies Bereiche auf der gescannten Seite, welche Textzeilen enthalten.

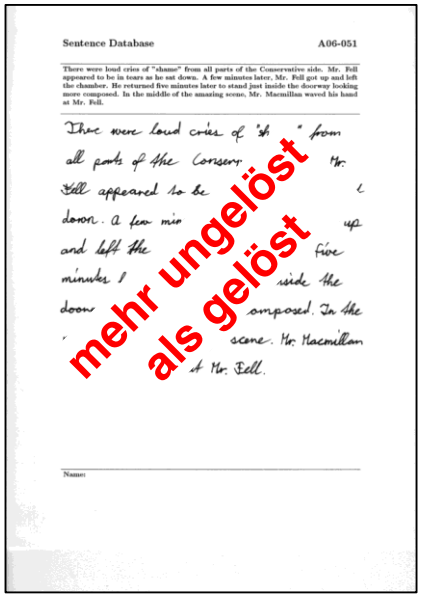


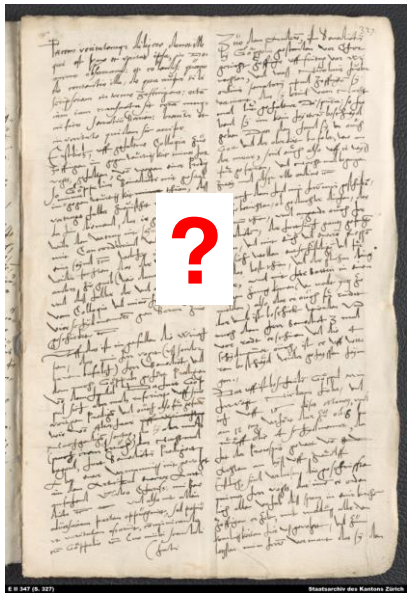
  - Für die Texterkennung sind dies Buchstaben-getreue, maschinen-lesbare Transkriptionen von Textzeilen.

Bullingere charissime, nihil literarum ad te dedi,

Stand der Technik

- Für historische Handschriften kann man keine allgemein gültige Aussage zur Fehlerrate machen. Sie ist abhängig von der Scan-Qualität, Layout-Komplexität, Lesbarkeit, Sprache, Anzahl Lernbeispielen, etc.





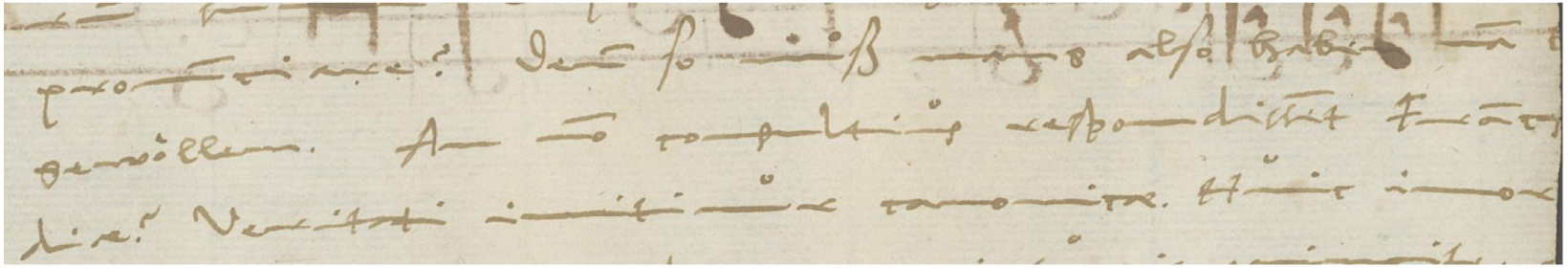
Bullinger Digital

- **Unterstützend:**
  - Gute Scan-Qualität
  - Relativ übersichtliches Seiten-Layout
  - Buchstabenschrift
  - Viele Briefe wurden bereits manuell transkribiert

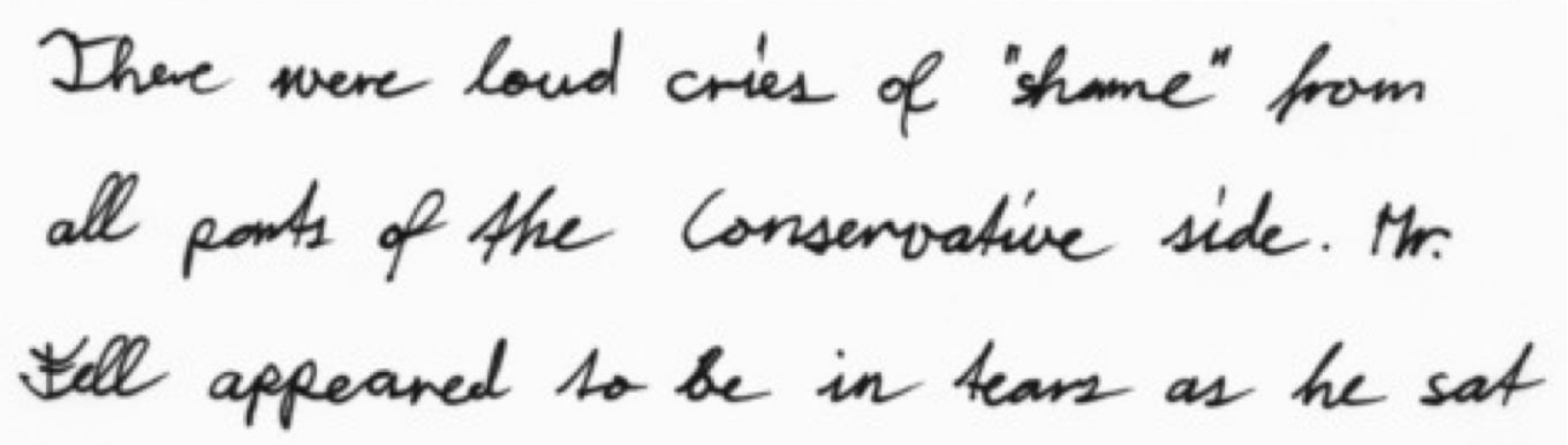
- **Erschwerend:**
  - Schlechte Lesbarkeit, z.B. für Bullinger selbst
  - Viele Wenigschreiber mit wenigen Briefen
  - Manuelle Transkription nicht Zeilen-genau
  - Manuelle Transkription nicht Buchstaben-genau
  - Latein und Deutsch gemischt
  - Viele Abkürzungen
  - Viele Worttrennungen beim Zeilenumbruch

Lesbarkeit

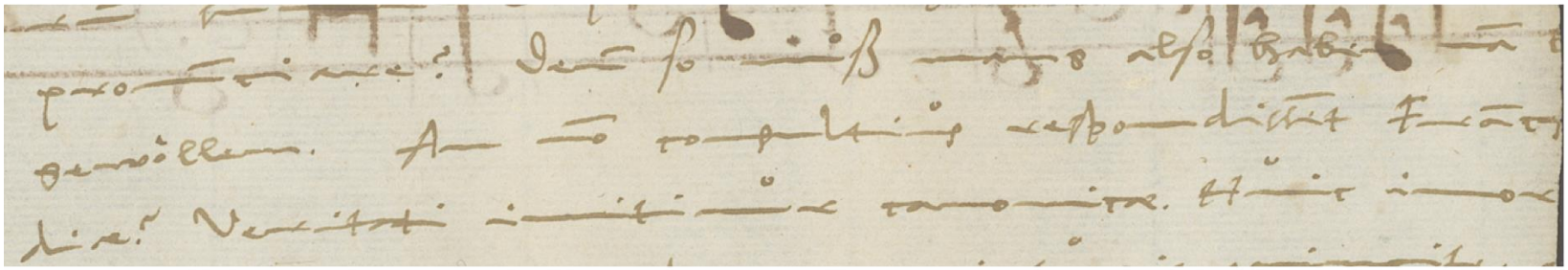
There were loud cries of "shame" from all parts of the Conservative side. Mr. Fell appeared to be in tears as he sat



Resultate



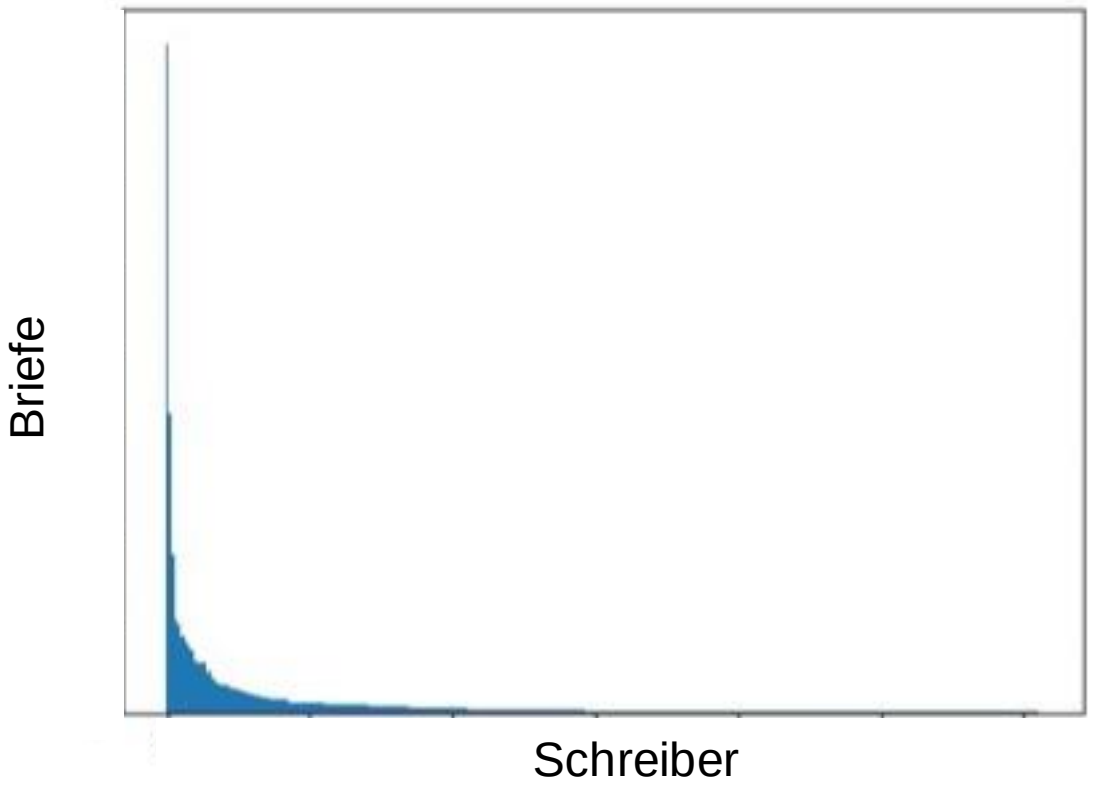
Buchstaben-Fehlerrate IAMDB: weniger als 5%



Buchstaben-Fehlerrate Bullinger: **weniger als 10%**
(typische Resultate zwischen 6.1% und 9.3%)

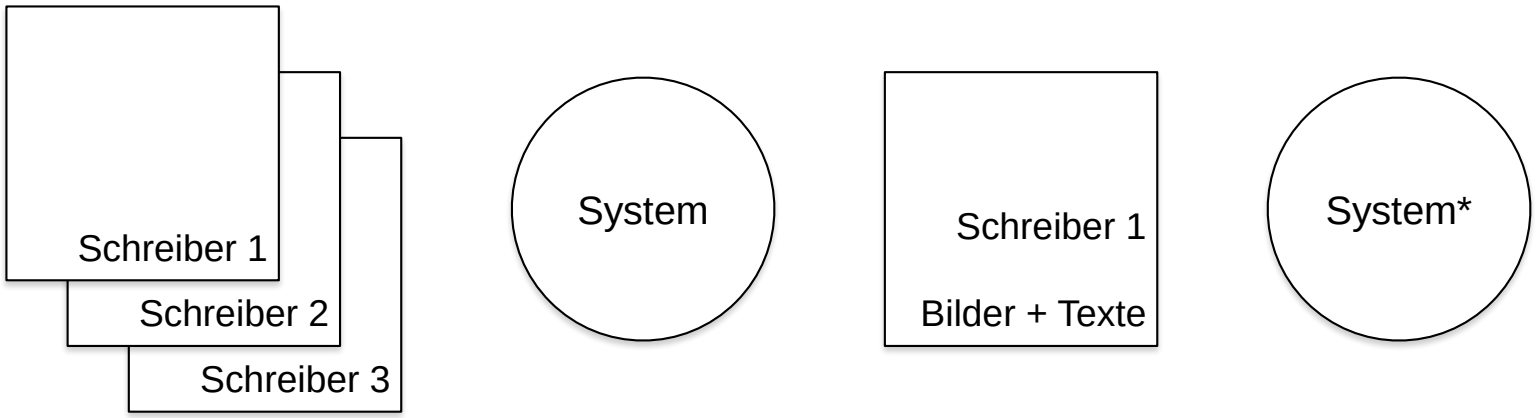
Vielschreiber und Wenigschreiber

- Von ca. 10’000 Briefen sind ca. 2’000 von Bullinger geschrieben.
- Problem für das Lernen von Schreibstilen:
  - Wenige Vielschreiber
  - Viele Wenigschreiber



Anpassung an Vielschreiber

- Zuerst wird ein System auf allen transkribierten Briefen trainiert.
- Danach wird es auf den transkribierten Briefen des Vielschreibers nachtrainiert.



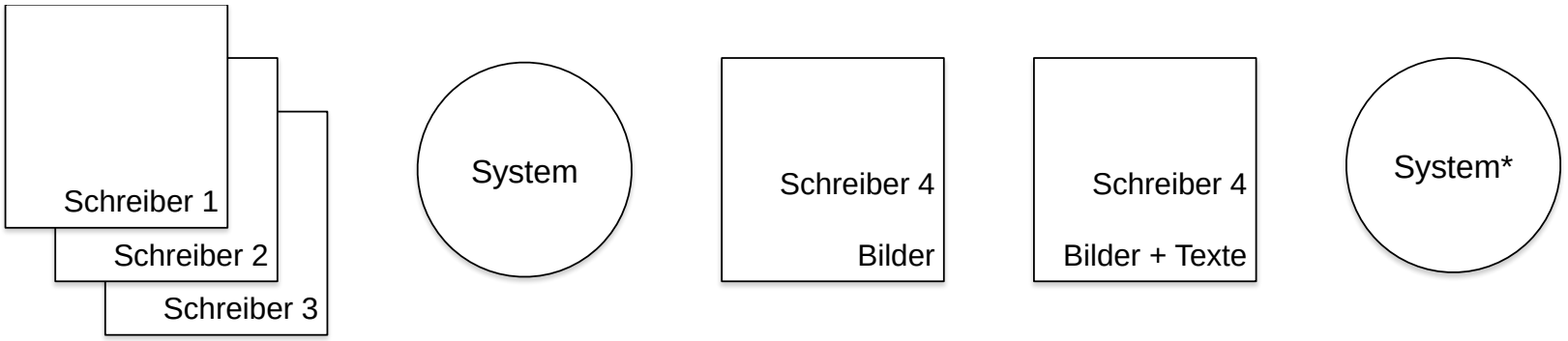
Anpassung an Vielschreiber: Resultate

- Verbesserung: Relative Reduktion der Buchstaben-Fehlerrate.
- Deutliche Verbesserung der Vielschreiber-Resultate.

System	Vielschreiber
PyLaia	**3.7%**
HTR-Flor	**9.8%**

Anpassung an Wenigschreiber

- Zuerst wird ein System auf allen transkribierten Briefen trainiert.
- Danach werden die unbekannten Briefe des Wenigschreibers automatisch transkribiert.
- Schliesslich wird das System auf den automatischen Transkriptionen des Wenigschreibers nachtrainiert.



Anpassung an Wenigschreiber: Resultate

- Verbesserung: Relative Reduktion der Buchstaben-Fehlerrate.
- Verbesserung der Wenigschreiber-Resultate, aber in deutlich geringerem Ausmass als für die Vielschreiber.

System	Vielschreiber	Wenigschreiber
PyLaia	**3.7%**	**0.7%**
HTR-Flor	**9.8%**	**2.0%**

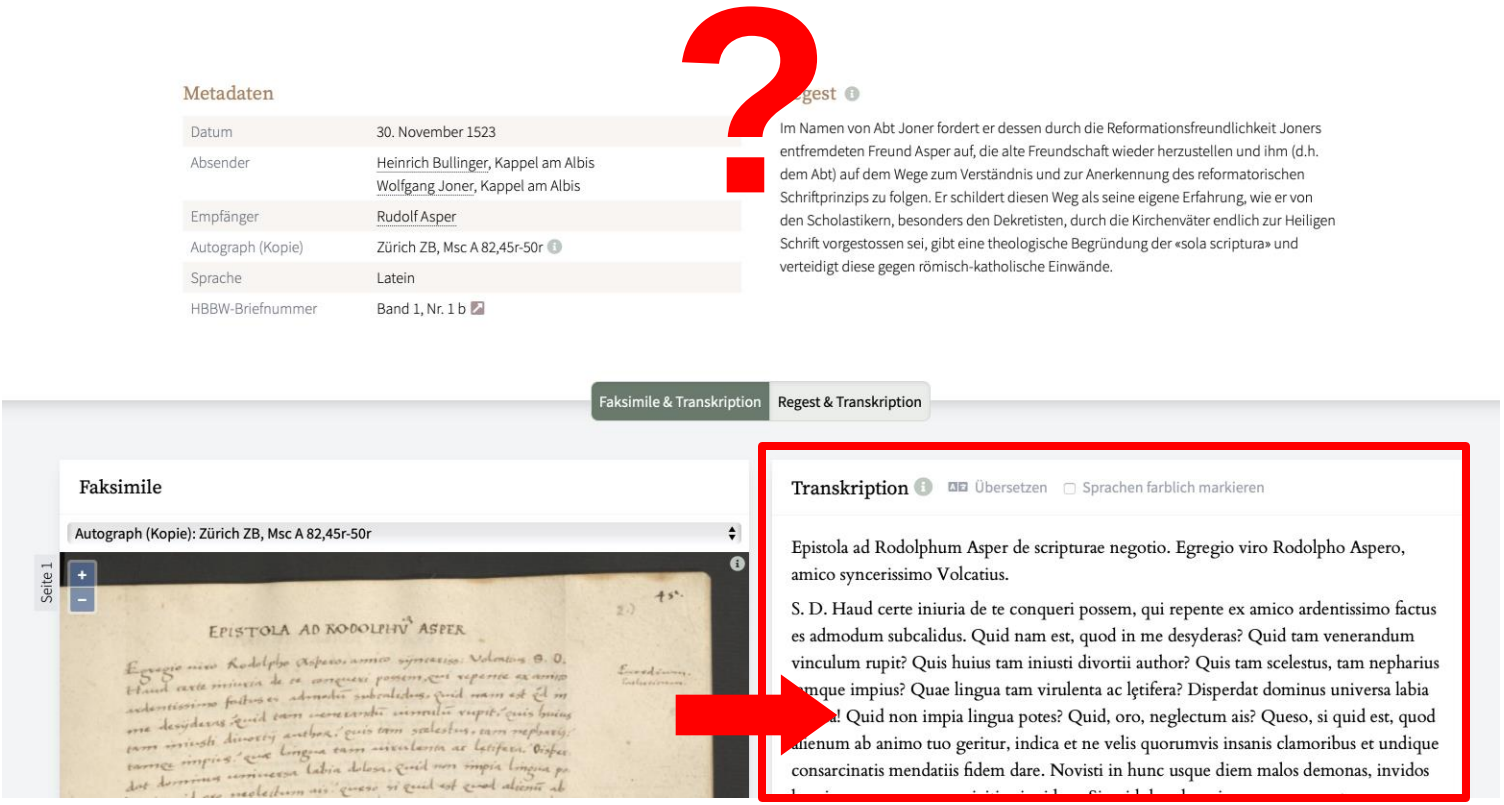
Aktuelle Forschung

- Automatische Gruppierung von Schreibstilen unter Verwendung von strukturellen Methoden, welche sich auf die Geometrie konzentrieren.
- Idee: Stilspezifische Systeme trainieren anstatt eines generischen Systems, welches alle Schreibstile kennt.

Schlussfolgerungen

- Trotz erschwerenden Bedingungen (schlechte Lesbarkeit, viele Wenigschreiber, Latein und Deutsch gemischt, …) ist es mit vereinten Kräften gelungen, eine **Buchstaben-Fehlerrate von weniger als 10%** zu erreichen. Dies erlaubt einen relativ genauen automatischen **Zugang zum textuellen Inhalt** der bisher nicht transkribierten Briefe.
- Um die Buchstaben-Fehlerrate weiter zu reduzieren, gibt es noch zahlreiche Ideen für weiterführende Arbeiten:
  - Gezielte manuelle Transkriptionen
  - Verbesserung der Bild-Text Übereinstimmung
  - Gruppieren von Schreibstilen
  - Synthetisieren von Schreibstilen
  - Kombination von Erkennungssystemen
  - Einbindung neuster Erkennungssysteme
  - u. v. m.

Fragen


Metadaten
Datum
30. November 1523
Absender
Heinrich Bullinger, Kappel am Albis
Wolfgang Joner, Kappel am Albis
Empfänger
Rudolf Asper
Autograph (Kopie)
Zürich ZB, Msc A 82,45r-50r
Sprache
Latein
HBBW-Briefnummer
Band 1, Nr. 1 b
Im Namen von Abt Joner fordert er dessen durch die Reformationsfreundlichkeit Joners entfremdeten Freund Asper auf, die alte Freundschaft wieder herzustellen und ihm (d.h. dem Abt) auf dem Wege zum Verständnis und zur Anerkennung des reformatorischen Schriftprinzips zu folgen. Er schildert diesen Weg als seine eigene Erfahrung, wie er von den Scholastikern, besonders den Dekretisten, durch die Kirchenväter endlich zur Heiligen Schrift vorgestossen sei, gibt eine theologische Begründung der «sola scriptura» und verteidigt diese gegen römisch-katholische Einwände.
Faksimile & Transkription
Regest & Transkription
Faksimile
Autograph (Kopie): Zürich ZB, Msc A 82,45r-50r
Seite 1
EPISTOLA AD RODOLPHV ASPER
Transkription
Übersetzen
Sprachen farblich markieren
Epistola ad Rodolphum Asper de scripturae negotio. Egregio viro Rodolpho Aspero, amico syncerissimo Volcatius.
S. D. Haud certe iniuria de te conqueri possem, qui repente ex amico ardentissimo factus es admodum subcalidus. Quid nam est, quod in me desyderas? Quid tam venerandum vinculum rupit? Quis huius tam iniusti divortii author? Quis tam scelestus, tam nepharius
mque impius? Quae lingua tam virulenta ac lętifera? Disperdat dominus universa labia
! Quid non impia lingua potes? Quid, oro, neglectum ais? Queso, si quid est, quod
lienum ab animo tuo geritur, indica et ne velis quorumvis insanis clamoribus et undique consarcinatis mendatiis fidem dare. Novisti in hunc usque diem malos demonas, invidos