EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Universitätsbibliothek Tübingen

OCR-BW

Kompetenzzentrum OCR der Universitätsbibliotheken Mannheim und Tübingen

Automatische Texterkennung von Handschriften im Projekt OCR-BW

22.6.2022, Dorothee Huff

Agenda

- Evaluierung von Transkribus anhand eigener Textkorpora
- Anwendungsbeispiele aus der wissenschaftlichen Praxis
- Serviceangebote
- Fazit

Projektteam Tübingen

- Projektleitung: Kristina Stöbener
- Dr. Regina Keyler
- Olaf Brandt
- Dorothee Huff

1. Projektphase (2019-2021)

- Bearbeitung und Texterkennung für verschiedene Textkorpora
  - Tagebücher des Geologen u. Paläontologen Edwin Hennig (1897-1973)
  - lateinische Tagebücher (1573-1605) und griechische Predigtnachschriften (1563-1604) des Altphilologen Martin Crusius
  - ausgewählte Bände der juristischen Konsilien (1602-1879) und Senatsprotokolle (1524-1912)
  - Handschriften und Drucke in Malayalam
- Unterstützung von Projekten und Wissenschaftler/innen beim Umgang mit Transkribus (Schulungen, Workshops)

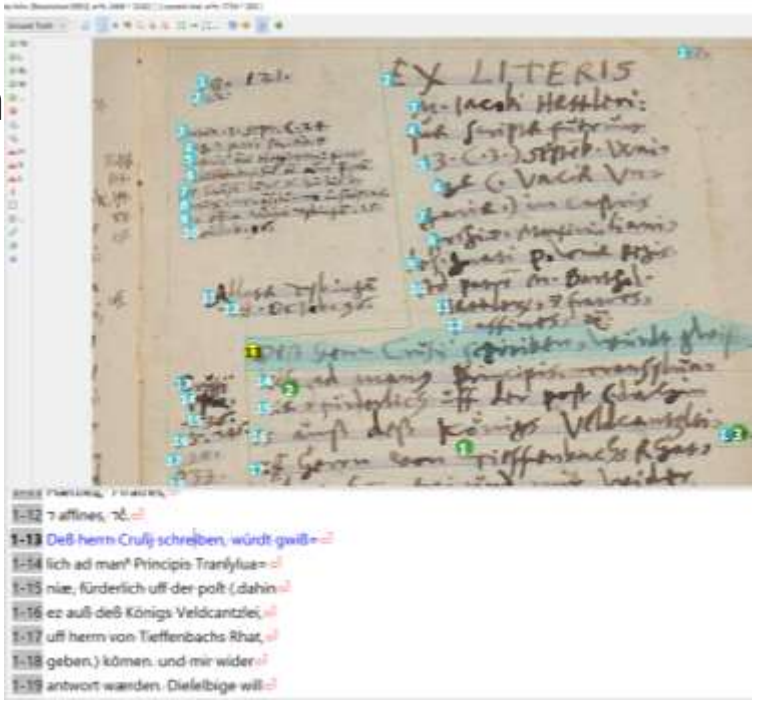
2. Projektphase (2021-2022)

- Bearbeitung von kleineren, heterogenen Textkorpora (werksspezifisches Training auf Grundlage von generischen Modellen)
  - mittelalterliche Handschriften
  - vermischtes, loses Schriftgut (auch Maschinenschrift, 19./20. Jh.)
  - Inkunabeln
- Evaluierung von eScriptorium zusammen mit der UB Mannheim
- Unterstützungsangebote durch Schulungsmaterialien, persönliche Beratung sowie Schulungen und Workshops

Bearbeitung von Textkorpora aus der Handschriften-abteilung und dem Universitätsarchiv der UB Tübingen

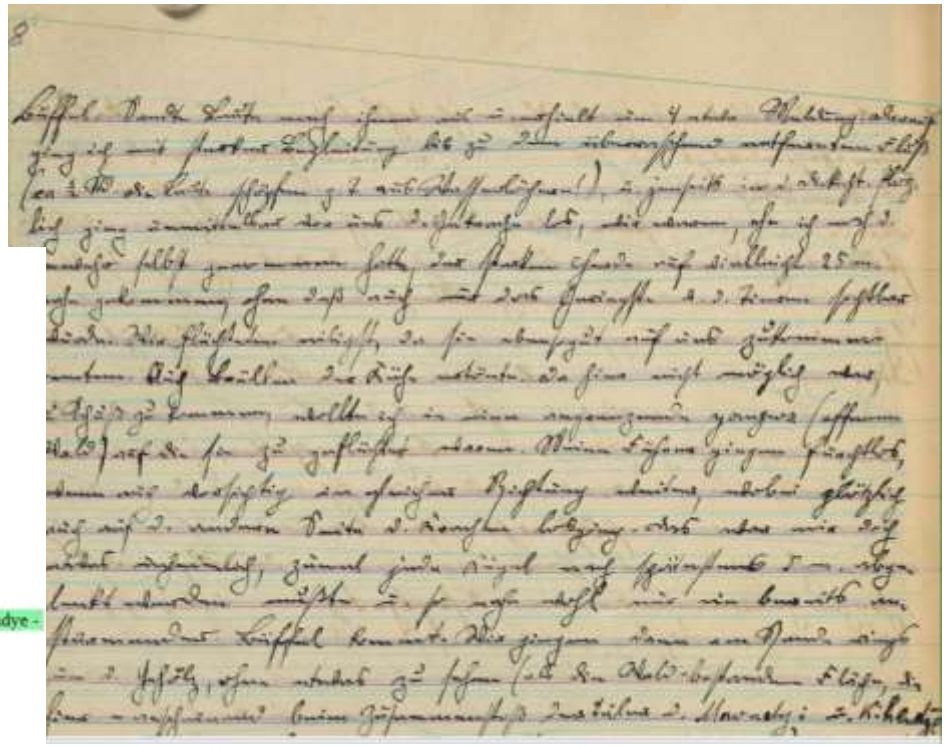
Vorhaben: Evaluierung von Transkribus anhand verschiedener Textkorpora aus dem Bestand der UB Tübingen

- Erstellung von Ground-Truth-Daten
- Modelltraining (Ziel: CER < 5%)



Tagebücher Edwin Hennig (1897-1973)

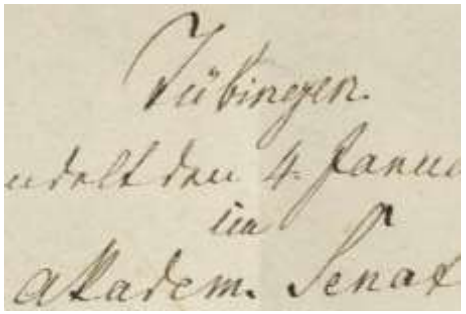
- ein Schreiber
- 165 S. GT (1897-1962)
- Deutsch, Suaheli
- 4,05% CER (3,61% mit LM)

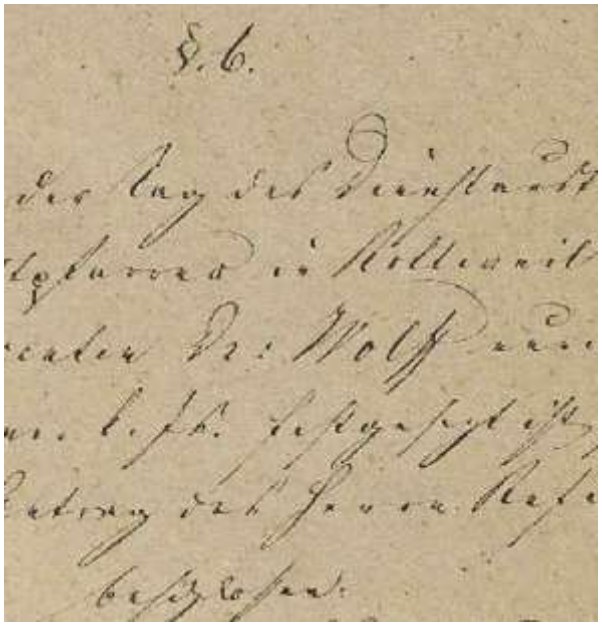


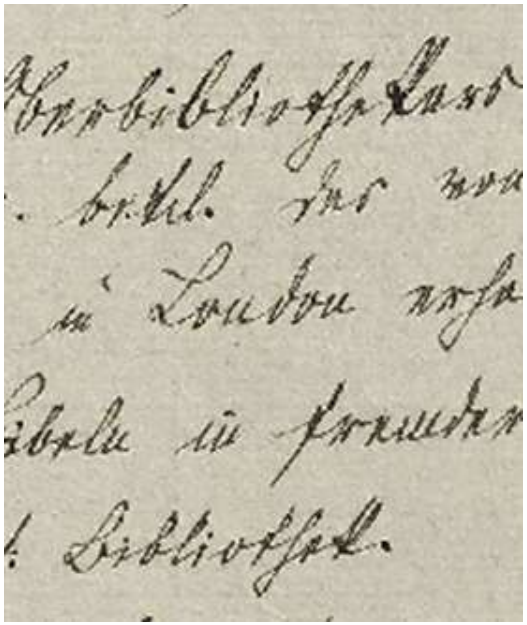
8
Büffel. Sandte Leute nach ihnen aus u. erhielt um 4 etwa Meldung. Darauf
ging ich mit starker Begleitung bis zu dem überraschend entfernten Fluß
(ca ½ Std. Die Leute schöpfen z. T. aus Wasserlöchern!), u. jenseits ~~im~~ in d. Dickicht. Plötz,,
lich ging unmittelbar vor uns d. Gekrache los, wir ~~waren~~ waren, ehe ich noch d.
Gewehr selbst genommen hatte, der starken Herde auf vielleicht 25 m.
nahe gekommen, ohne daß auch nur das Geringste v. d. Tieren sichtbar
wurde. Wir flüchteten ~~nächst~~ eiligst, da sie ebensogut auf uns zukommen
konnten. Auch ~~Brullen~~ Brüllen der Kühe ertönte. Da hier nicht möglich war,
zu Schuß zu kommen, wollte ich in eine angrenzende yangwa (offenen
Wald) auf die sie zu geflüchtet waren. Meine Führer gingen furchtlos,
wenn auch vorsichtig in gleicher Richtung ~~weiter~~ weiter, wobei plötzlich
auch auf d. andern Seite d. Krachen losging. Das war mir doch
etwas unheimlich, zumal jede Kugel nach spätestens 5 m. abge,,
lenkt werden mußte u. so nahe wohl nur ein bereits an,,
stürmender Büffel kommt. Wir gingen dann am Rande rings
um d. Gehölz, ohne etwas zu sehen (als die Wald-bestandene Fläche, die
hier - ~~anscheinend~~ anscheinend beim Zusammenstoß der Täler d. Mavudyi u. ~~Kihendee~~ Kihendye -
in den Busch u. d. Hügelland unvermittelt eingestreut ist.) Auf d. Weg
zurückgekehrt trafen wir den ~~Jumben~~ Jumben, einen alten freundlichen hübschen
~~Greis~~ Greis mit 2 kleinen ~~Kindern~~ Kindere ruhig auf d. Nachhausewege, er berichtete
daß die Spuren nahe am Fluß üb. d. Weg führten. In d. Tat war die
ganze Herde im Bogen rückwärts durchgebrochen u. zwar im übrigen
völlig lautlos. Es war überflüssig in dem Dickicht zu folgen, ich kehrte wieder
unverrichteter Sache ~~heim~~ heim, aber doch um einige seltsame Erfahrungen
reicher. Die Leutchen hier in diesem verlornen Talkessel ~~kannen~~ kamen
mir sehr viel freundlicher vor als alle andern Kilwa-Bewohner.
4. Aug. Rückkehr nach Makangaga. Morgens schöne Wanderung

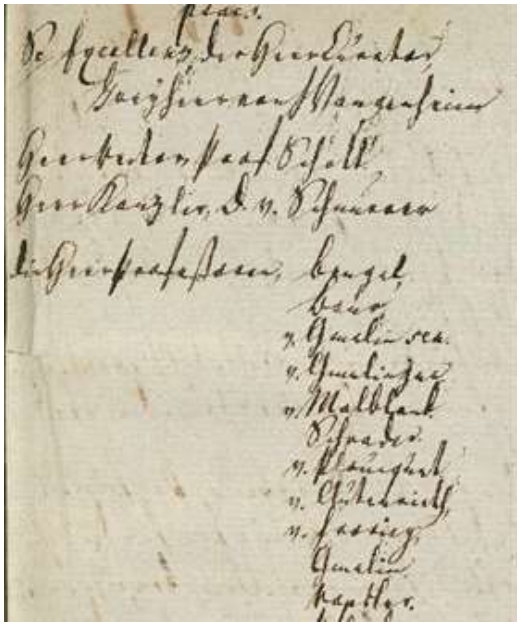
Senatsprotokolle (1524-1912)

- mehrere Schreiber
- 214 S. GT (1799-1847)
- Deutsch
- 4,60% CER (4,24% mit LM)



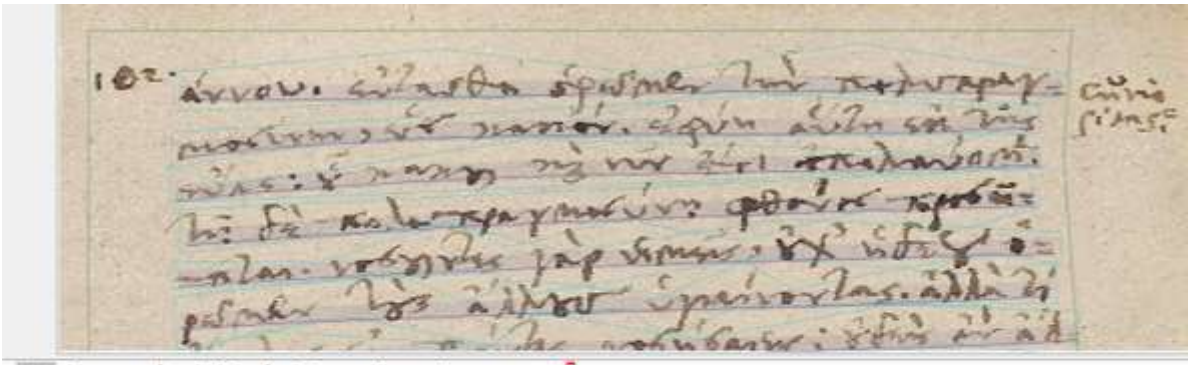






Griechische Predigtnachschriften Martin Crusius (1563-1604)

- ein Schreiber
- 180 S. GT (1563-1604)
- Griechisch
- 3,54% CER (3,41% mit LM)



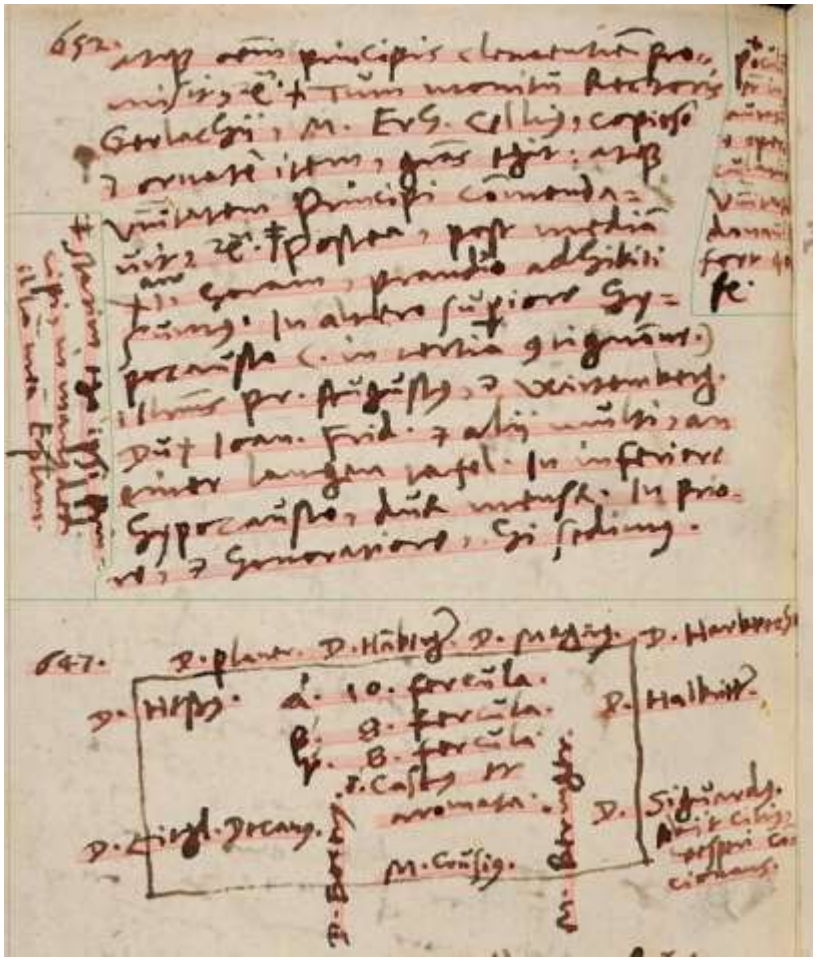
1-1 ἄννου. ἐνταῦθα ὁρῶμεν τὴν πολυπραγ=
1-2 μοσύνην, ὡς κακόν. ἐφύη αὕτη ἐκ τῆς
1-3 εὔας : οὗ κακοῦ χ᾽ νῦν ἔτι ἀπολαύομεν.
1-4 τῇ δὲ πολυπραγμοσύνῃ φθόνος προσῆ=
1-5 πται. νοσοῦντες γὰρ ἡμεῖς, οὐχ᾽ ἡδέως ὁ=
1-6 ρῶμεν τοὺς ἄλλους ὑγιαίνοντας. ἀλλὰ τί
1-7 ὄφελος, εἰ πάντες νοσήσαιεν; οὐδεὶς ἂν ἄλ=

	M4d
Mb 19-1, S. 58 (1587)	
Mb 19-2, S. 145 (1566)	2,56
Mb 19-3, S. 118 (1564)	3,66
Mb 19-4, S. 107 (1564)	2,56
Mb 19-7, S. 484 (1575)	3,52
Mb 19-8, S. 14 (1575)	3,64
Mb 19-9, S. 188 (1577)	6,4
Mb 19-10, S. 12 (1579)	1,93
Mb 19-11, S. 70 (1581)	3,66
Mb 19-12, S. 24 (1582)	3,87
Mb 19-13, S. 12 (1583)	3,08
Mb 19-14, S. 34 (1587)	1,93
Mb 19-15, S. 192 (1589)	3,23
Mb 19-16, S. 612 (1594)	3,54
Mb 19-17, S. 204 (1595)	3,08
Mb 19-18, S. 116 (1597)	3,53
Mb 19-19, S. 55 (1600)	3,76
Mb 19-20, S. 69 (1602)	3,56
Mb 12,	
Mb 17, S. 14 (1594)	2,44
CER im Durchschnitt	3,41
GT Training	162
GT Validation	18
GT insgesamt	180

Lateinische Tagebücher Martin Crusius (1573-1605)

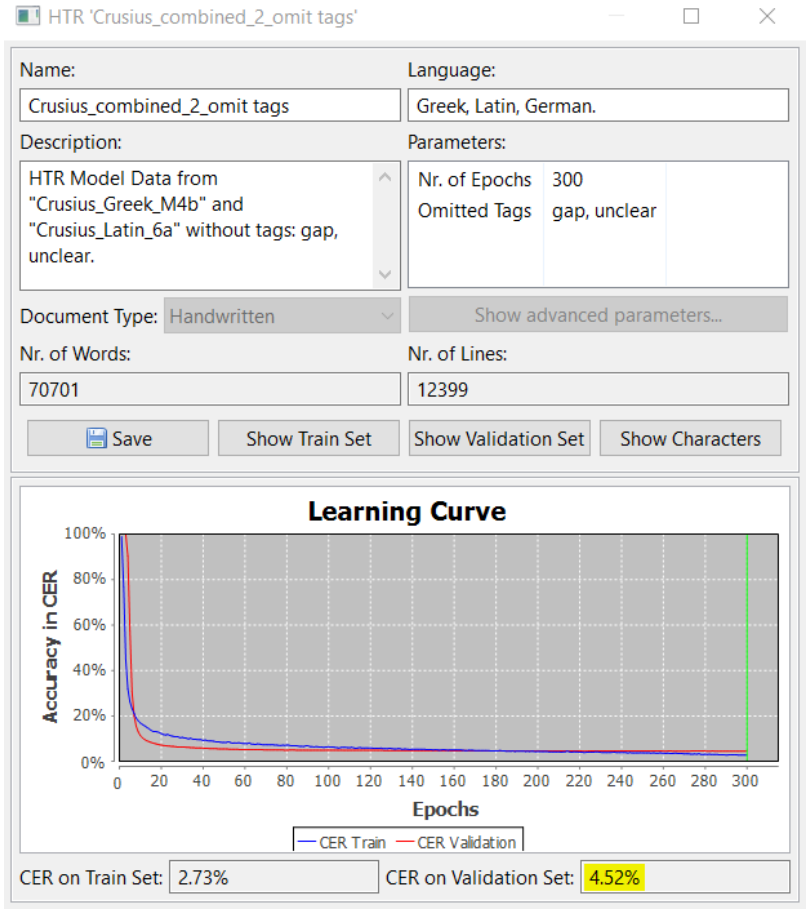
- ein Schreiber
- 136 S. GT (1573-1605)
- Latein, Deutsch
- 5,13% CER (4,66% mit LM)

Foto: Valentin Marquardt/Universität Tübingen



Kombinationsmodell Crusius

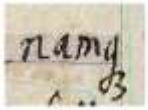
- Schreiber: hauptsächlich Martin Crusius
- 346 S. GT (1563-1605)
- Griechisch, Latein, Deutsch
- 4,52% CER (4,22% mit LM)



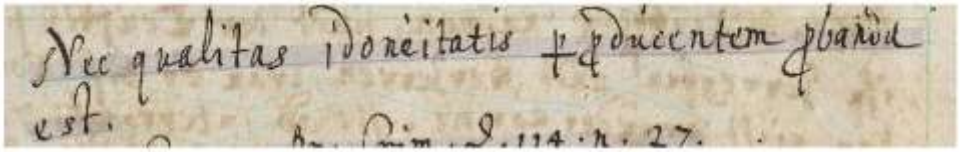
Juristische Konsilien (1602-1879)

- mehrere Schreiber
- 223 S. GT (1659-1665)
- Deutsch, Latein
- 2,09% CER (1,95% mit LM)

Foto: Valentin Marquardt/Universität Tübingen



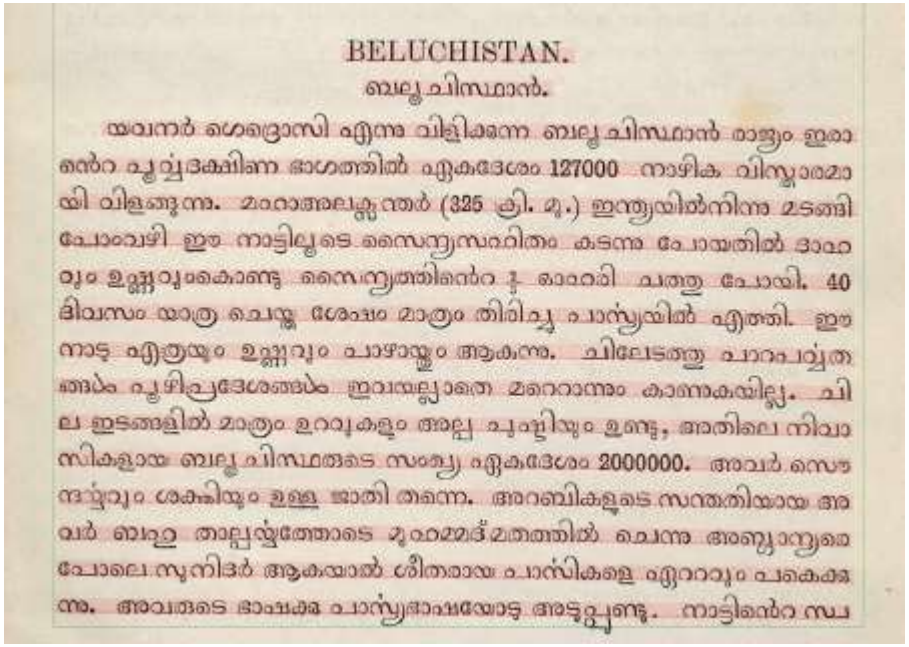
- **Jur_Kons_Tue:** namq3
- **Acta_17:** namque



- **Jur_Kons_Tue:** Nec qualitas idoneitatis p pducentem pbandu
- **Acta_17:** Nec qualitas idoneitatis per producentem pbanda

Malayalam

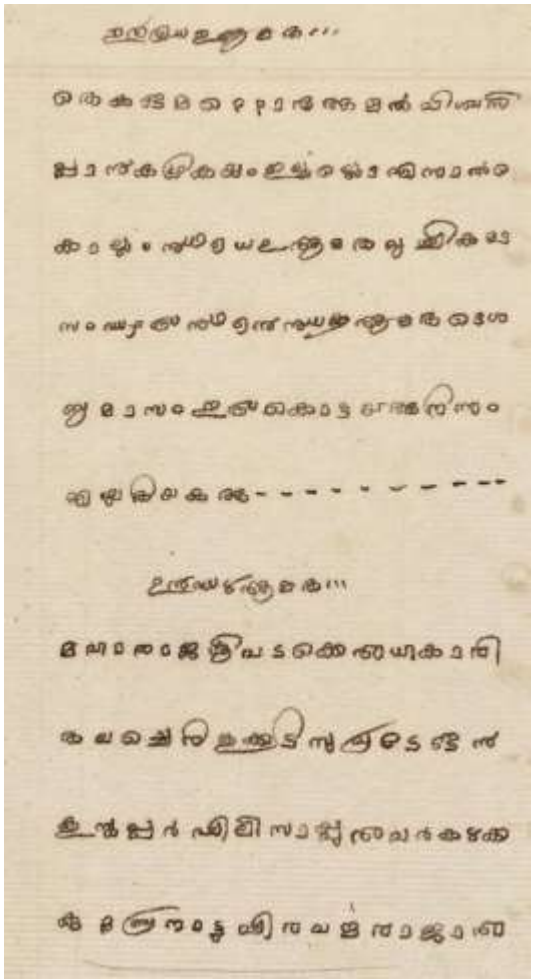
- mehrere Drucke
- 48 S. GT (19. Jh.)
- Malayalam
- 2,29% CER (2,53% mit LM)



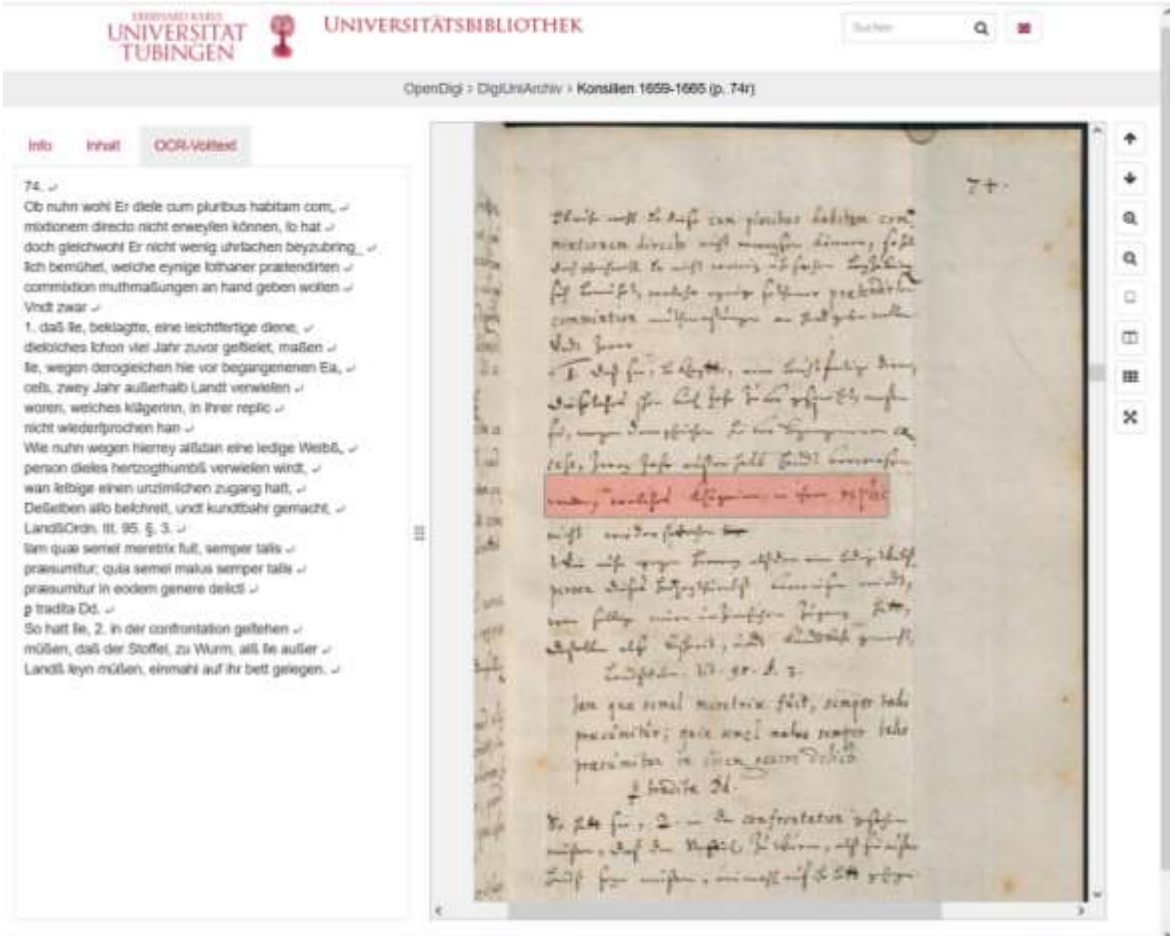
BELUCHISTAN.

ബലൂചിസ്ഥാൻ.

യവനർ ശൈറ്റോസി എന്നു വിളിക്കുന്ന ബലൂചിസ്ഥാൻ രാജ്യം ഇരാന്റെ പൂർവ്വദക്ഷിണ ഭാഗത്തിൽ ഏകദേശം 127000 നാഴിക വിസ്താരമായി വിളങ്ങുന്നു. മഹാഅലക്സന്തർ (325 ക്രി. മു.) ഇന്ത്യയിൽനിന്നു മടങ്ങിപോരുവഴി ഈ നാട്ടിലൂടെ സൈന്യസഹിതം കടന്നു പോയതിൽ ദാഹവും ഉണ്ണുവുംകൊണ്ടു സൈന്യത്തിന്റെ ½ ഓരോരി ചത്തു പോയി. 40 ദിവസം യാത്ര ചെയ്ത ശേഷം മാത്രം തിരിച്ചു പാസ്യയിൽ എത്തി. ഈ നാട്ടു എന്തുയും ഉണ്ണുവും പാഴായ്തും ആകുന്നു. ചിലേടത്തു പാറപ്പുറ്റങ്ങളും പൂഴിപ്രദേശങ്ങളും അല്ലാതെ മറ്റൊന്നും കാണുകയില്ല. ചില ഇടങ്ങളിൽ മാത്രം ഉറവുകളും അല്പ പുഴുവിയും ഉണ്ടു, അതിലെ നിവാസികളായ ബലൂചിസ്ഥരുടെ സംഖ്യ ഏകദേശം 2000000. അവർ സൌന്ദര്യവും ശക്തിയും ഉള്ള ജാതി തന്നെ. അറബികളുടെ സന്തതിയായ അവർ ബഹു താല്പര്യത്തോടെ മുഹമ്മദ് മതത്തിൽ ചെന്നു അബ്യൂന്യരെ പോലെ സുനിദർ ആകയാൽ ശീതരായ പാർസികളെ എപ്പോഴും പകെക്കുന്നു. അവരുടെ ഭാഷക്കു പാസ്യഭാഷയോടു അടുപ്പുണ്ടു. നാട്ടിന്റെ സ്ഥ



Ansicht in den digitalen Sammlungen (UB Tübingen)



Anwendungsfälle aus der wissenschaftlichen Praxis

Vorhaben: Erzeugung von möglichst guten automatischen Transkripten mit möglichst geringem Aufwand zur Weiterverarbeitung der Daten in anderen Kontexten

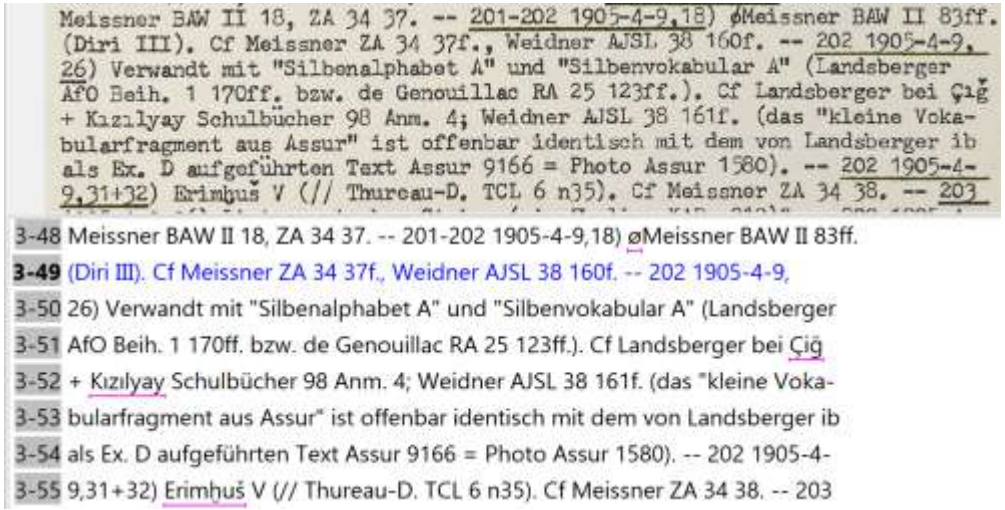
Wer: Anglistik, Altorientalische Philologie, Biologie, Germanistik, Geschichte, Judaistik, Kulturanthropologie, Orient- und Islamwissenschaft, Paläontologie, Rechtswissenschaft, Skandinavistik, Theologie

Wozu: Lesehilfe, maschinelle Auswertung und Weiterverarbeitung, Vorbereitung von Editionen, Aufbereitung von Druckeditionen, Durchsuchung automatischer Transkriptionen mit Keyword-Spotting

1. Werksspezifisches Training

Beispiel: Handbuch für Keilschriftliteratur (Ziel: Aufbereitung für Datenbank)

→ automatische Transkription mit dem Modell „Transkribus Typewriter 0.1“, anschließend Korrektur und Hinzufügung sprachspezifischer Sonderzeichen


Meissner BAW II 18, ZA 34 37. -- 201-202 1905-4-9,18) øMeissner BAW II 83ff. (Diri III). Cf Meissner ZA 34 37f., Weidner AJSL 38 160f. -- 202 1905-4-9, 26) Verwandt mit "Silbenalphabet A" und "Silbenvokabular A" (Landsberger AfO Beih. 1 170ff. bzw. de Genouillac RA 25 123ff.). Cf Landsberger bei Çig + Kızılyay Schulbücher 98 Anm. 4; Weidner AJSL 38 161f. (das "kleine Vokabularfragment aus Assur" ist offenbar identisch mit dem von Landsberger ib als Ex. D aufgeführten Text Assur 9166 = Photo Assur 1580). -- 202 1905-4-9,31+32) Erimḫuš V (// Thureau-D. TCL 6 n35). Cf Meissner ZA 34 38. -- 203

3-48 Meissner BAW II 18, ZA 34 37. -- 201-202 1905-4-9,18) øMeissner BAW II 83ff.
3-49 (Diri III). Cf Meissner ZA 34 37f., Weidner AJSL 38 160f. -- 202 1905-4-9,
3-50 26) Verwandt mit "Silbenalphabet A" und "Silbenvokabular A" (Landsberger
3-51 AfO Beih. 1 170ff. bzw. de Genouillac RA 25 123ff.). Cf Landsberger bei Çig
3-52 + Kızılyay Schulbücher 98 Anm. 4; Weidner AJSL 38 161f. (das "kleine Voka-
3-53 bularfragment aus Assur" ist offenbar identisch mit dem von Landsberger ib
3-54 als Ex. D aufgeführten Text Assur 9166 = Photo Assur 1580). -- 202 1905-4-
3-55 9,31+32) Erimḫuš V (// Thureau-D. TCL 6 n35). Cf Meissner ZA 34 38. -- 203

Borger, Rykle: Handbuch der Keilschriftliteratur. Bd. 1: Repertorium der sumerischen und akkadischen Texte. Berlin, 1967, S. 337.

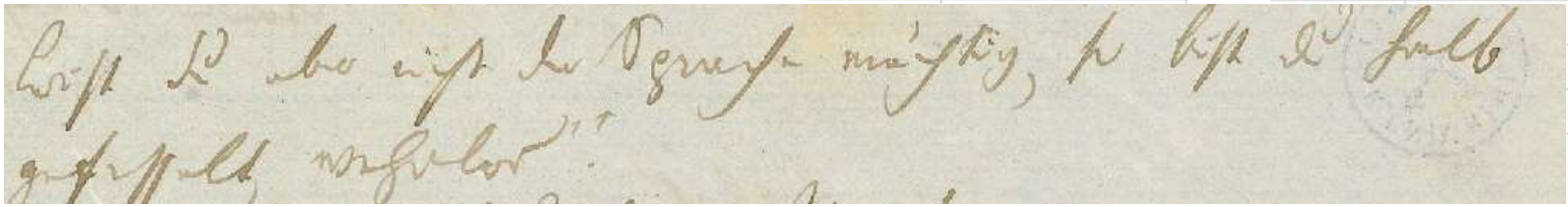

Ç LATIN CAPITAL LETTER C WITH CEDILLA
É LATIN CAPITAL LETTER E WITH ACUTE
Ú LATIN CAPITAL LETTER U WITH ACUTE
à LATIN SMALL LETTER A WITH GRAVE
á LATIN SMALL LETTER A WITH ACUTE
ä LATIN SMALL LETTER A WITH DIAERESIS
ç LATIN SMALL LETTER C WITH CEDILLA
è LATIN SMALL LETTER E WITH GRAVE
é LATIN SMALL LETTER E WITH ACUTE
ê LATIN SMALL LETTER E WITH CIRCUMFLEX
í LATIN SMALL LETTER I WITH CIRCUMFLEX
ó LATIN SMALL LETTER O WITH ACUTE
ö LATIN SMALL LETTER O WITH DIAERESIS
ø LATIN SMALL LETTER O WITH STROKE
ú LATIN SMALL LETTER U WITH ACUTE
û LATIN SMALL LETTER U WITH CIRCUMFLEX
ü LATIN SMALL LETTER U WITH DIAERESIS
ý LATIN SMALL LETTER Y WITH ACUTE
ā LATIN SMALL LETTER A WITH MACRON
č LATIN SMALL LETTER C WITH CARON
ē LATIN SMALL LETTER E WITH MACRON
ğ LATIN SMALL LETTER G WITH BREVE
ī LATIN SMALL LETTER I WITH MACRON
ı LATIN SMALL LETTER DOTLESS I
ś LATIN SMALL LETTER S WITH ACUTE
Š LATIN CAPITAL LETTER S WITH CARON
š LATIN SMALL LETTER S WITH CARON
ū LATIN SMALL LETTER U WITH MACRON
ˈ MODIFIER LETTER SMALL D
Ḫ LATIN CAPITAL LETTER H WITH BREVE BELOW
ḫ LATIN SMALL LETTER H WITH BREVE BELOW
Ṣ LATIN CAPITAL LETTER S WITH DOT BELOW
ṣ LATIN SMALL LETTER S WITH DOT BELOW
₆ SUBSCRIPT SIX
← LEFTWARDS ARROW
◦ SQUARE LOZENGE

2. Nutzung generischer Modelle als Transkriptions-grundlage

Beispiel: Wissenschaftlicher Nachlass Friedrich Theodor Vischer
(Ziel: Edition)
→ automatische Transkription mit dem Modell „German_ Kurrent_XIX-XX_M6-1“, anschließend Korrektur nach den gewünschten Transkriptionsrichtlinien und Modelltraining

Pages	CER (Vischer_5)
Overall	4,54
Page 1	5,78
Page 2	3,38
Page 3	26,9
Page 4	5,56
Page 5	5,82
Page 6	5,57
Page 7	2,54
Page 8	3,55
Page 9	1,79
Page 10	4,15
Page 11	2,5
Page 12	3,61

Pages	CER (German_Kurrent _XIX-XX_M6-1)
Overall	9,82
Page 1	16,33
Page 2	9,69
Page 3	14,52
Page 4	20
Page 5	11,82
Page 6	12,2
Page 7	6,17
Page 8	6,03
Page 9	6,55
Page 10	7,85
Page 11	9,41
Page 12	9,23



3. Nachnutzung vorhandener Transkriptionen

Beispiel: Projekt „Narrative Vermittlung religiösen Wissens: Edition und Kommentierung geistlicher Vers- und Prosatexte des 13. bis 16. Jahrhunderts“ der Universitäten Köln und Tübingen (Ziel: Edition)

→ Nachnutzung von bereits im Projekt angefertigten Transkriptionen für ein Modelltraining mit sukzessivem Nachtraining
→ Transkriptionsrichtlinien: Auflösung von Abkürzungen in Klammern

Serviceangebote der UB Tübingen

- Beratung allgemein
- Einführungsveranstaltungen für Mitarbeiter/innen von Bibliotheken und Archiven, Wissenschaftler/innen und Studierende
- Bearbeitungen von Probeseiten (Empfehlungen)
- Unterstützung von Projekten und bei Projektanträgen

Fazit

- Ergebnisse der automatischen Texterkennung besser als erwartet (auch heterogenes Material wie z.B. unterschiedliche Schreiber und/oder lange Schreibzeiträume beeinträchtigt das Ergebnis nicht wesentlich und verlangt bei entsprechender Planung nicht unbedingt einen höheren Ressourcenaufwand)

- unterschiedliche Sprachen und Schriftsysteme sind kein Problem

- für ein bestmögliches Ergebnis ist ein Modelltraining notwendig, wobei sich oftmals auch schon mit generischen Modellen gute Ergebnisse erzielen lassen (für ein fehlerfreies Ergebnis bedarf es manueller Nachkorrektur)

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Danke.

Kontakt:
Universitätsbibliothek Tübingen
Wilhelmstraße 32, 72074 Tübingen

Dorothee Huff
Telefon: +49 7071 29-72852
Dorothee.huff@uni-tuebingen.de

https://uni-tuebingen.de/de/179298
https://uni-tuebingen.de/de/230831