Möglichkeiten und Grenzen ‚smarter' HTR für unterschiedliche Sprachen und Schriftsysteme

Achim Rabus

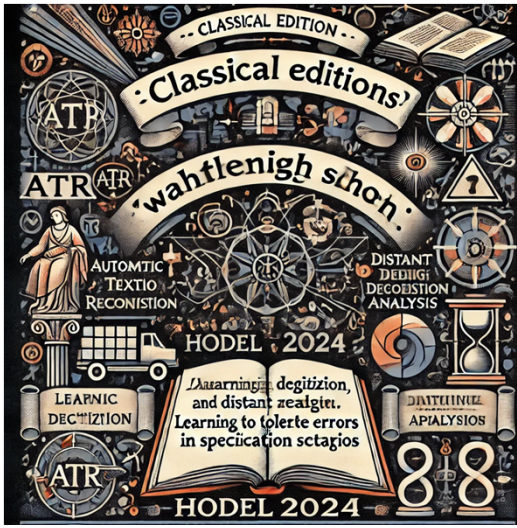
achim.rabus@slavistik.uni-freiburg.de

‚Normale' HTR und mögliche Konsequenzen

- ‚Normal': Diplomatische Transkription inkl. aller Besonderheiten bspw. im Bereich von Orthographie, Worttrennung, Interpunktion
- Sehr gute Modelle (vor allem für westliche Schriften/Sprachen) für unterschiedliche Plattformen
- Enorme Vereinfachung für traditionelle (digitale) Editionswissenschaft

HTR-Standardmodelle und Konsequenzen

- Aber: Hat sich im digitalen Zeitalter die klassische Edition überlebt? „[W]ahrscheinlich schon.“ (Hodel 2024)
- HTR als methodologischer Gamechanger → Massendigitalisierung, Distant Reading, quantitative Ansätze
- Lernen, Fehler auszuhalten! (zumindest für bestimmte Anwendungsszenarien)



Gamechanger, Paradigmenwechsel, Konsequenzen

- Im quantitativen Zeitalter (und im Zeitalter knapper Kassen): höherer Rechtfertigungsdruck für kleinteilige, aufwändige, qualitativ orientierte (Editions-)Arbeit
- Impact/Reichweite der Forschung rückt verstärkt in den Fokus
- HTR/KI zur Demokratisierung des Wissens? Zum Abbau von Herrschaftswissen?
- Dazu ‚smarte' Modelle nötig (Rabus 2024, Rabus in press)
- Vor allem „Recycling-Ansatz" zur GT-Erstellung

- Dank an Mitarbeiter*innen der Projekte MultiHTR 1, 2, Contínslav, Quantislav: Lesley Loew, Milanka Matić-Chalkitis, Martin Meindl Elena Renje, Aleksej Tikhonov

Transkribus

universität freiburg

digital@bw

Bundesministerium für Bildung und Forschung

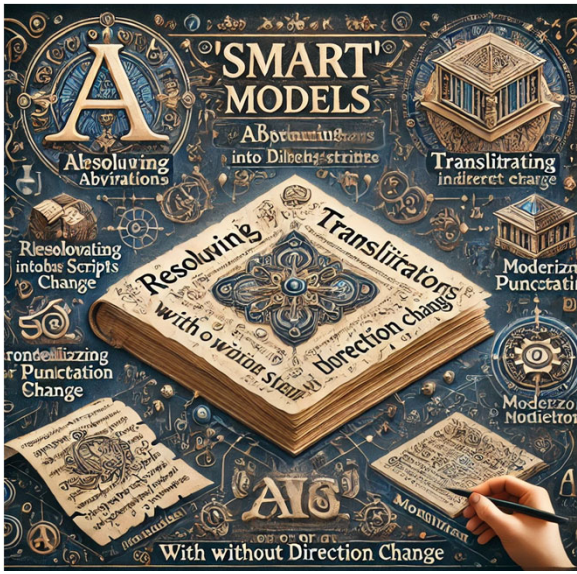
BAdW Bayerische Akademie der Wissenschaften

Narodowe Centrum Nauki

DFG

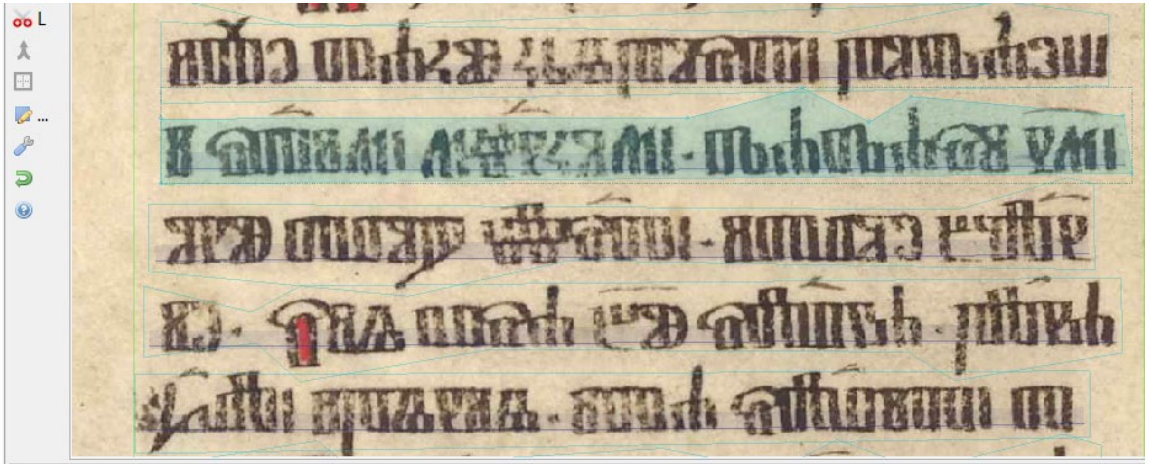
Was sind ‚smarte' Modelle?

- Jenseits der reinen Transkription zusätzliche Eigenschaften, z.B.
  - Auflösung von Abbreviaturen
  - Transliteration in ein anderes Schriftsystem (mit/ohne Wechsel der Direktionalität)
  - Modernisierung von Orthographie, Interpunktion, Zahlensystem etc.



Glagolitisch (Kroatien, 15. Jh.)

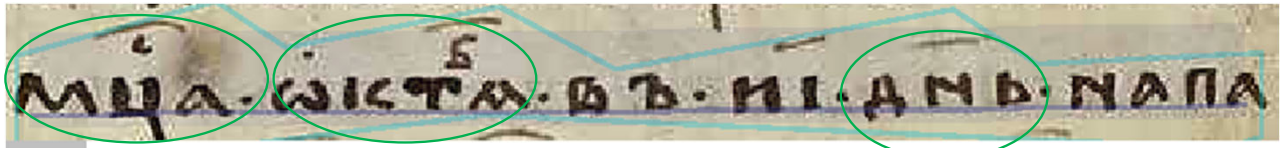
- Transkription von Glagolica in Latinica
- Auflösung von Abbreviaturen
- Lesehilfe



2-18 iže taku krêpostь podaeš
2-19 i s(ve)timь m(u)č(e)n(i)komь . da da si n(a)mь
2-20 onu tvoû m(i)l(o)stь . i tvoe bl(agoslovle)n
2-21 ie . siê vsa b(og)u sl(a)vna . plna
2-22 g(lago)lь i pêniê . i ta sl(a)vitь t

- Dank an Sanja Zubčić, Staroslavenski institut Zagreb; Guido und Jagoda Kappel, Wien

Kyrillisch: 11.–16. Jh.: Modelle für unterschiedliche Zielgruppen



1-1 месяца октября въ 18 день, на па

1-1 м҃ца. ѡктѧ. въ. ӥ҃. дн҃ь. напа-

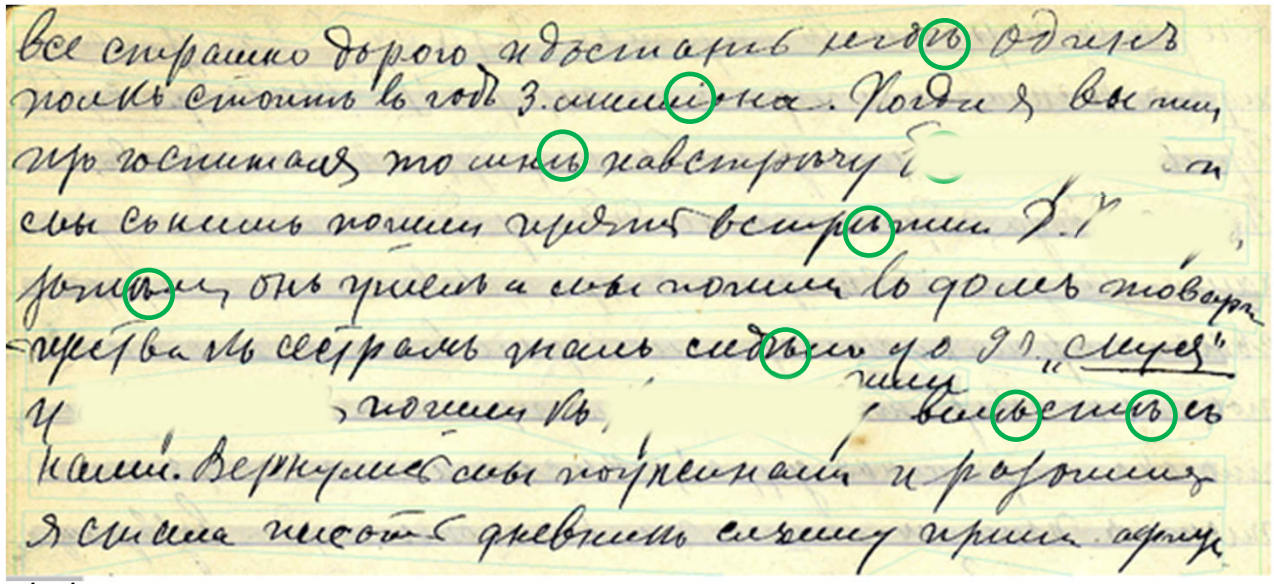
Beibehaltung der Ligaturen und Abkürzungen

Für modernisierende Editionen und Nicht-Philolog*innen

Russisch (20. Jh.)

- Orthographiereform 1918
- Abschaffung verschiedener Grapheme
- Modell mit modernisierenden Fähigkeiten

Dokumente aus dem Prozhito-Projekt:
Tagebücher, insb. Anfang 20. Jh.

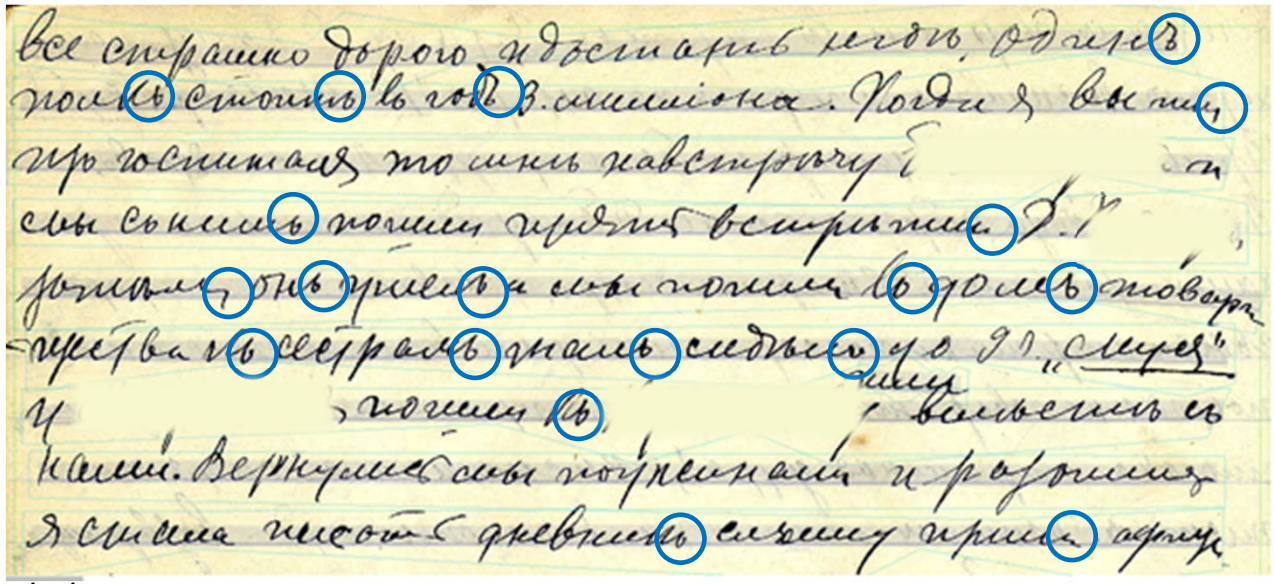


1-1 все страшно дорого и достать негде один
1-2 полк стоит в год 3. миллиона. Когда я вышил
1-3 из госпиталя то мне навстречу <name> и
1-4 мы с ним пошли гулять встретил Доктор <name>
1-5 затем он ушел и мы пошли в дом товари
1-6 щества к сестрам тал сидел ио 8 ч. муся»
1-7 шлу
1-8 у <name> пошли к <name> вмлесте с
1-9 нами. Вернулись мы поужинали и разошлис
1-10 Я стала писать дневник слышу пришл офица

Automatische Modernisierung:
ѣ > е
і > и

Dank an das Prozhito-Projekt Sankt Petersburg

Dokumente aus dem Prozhito-Projekt:
Tagebücher, insb. Anfang 20. Jh.



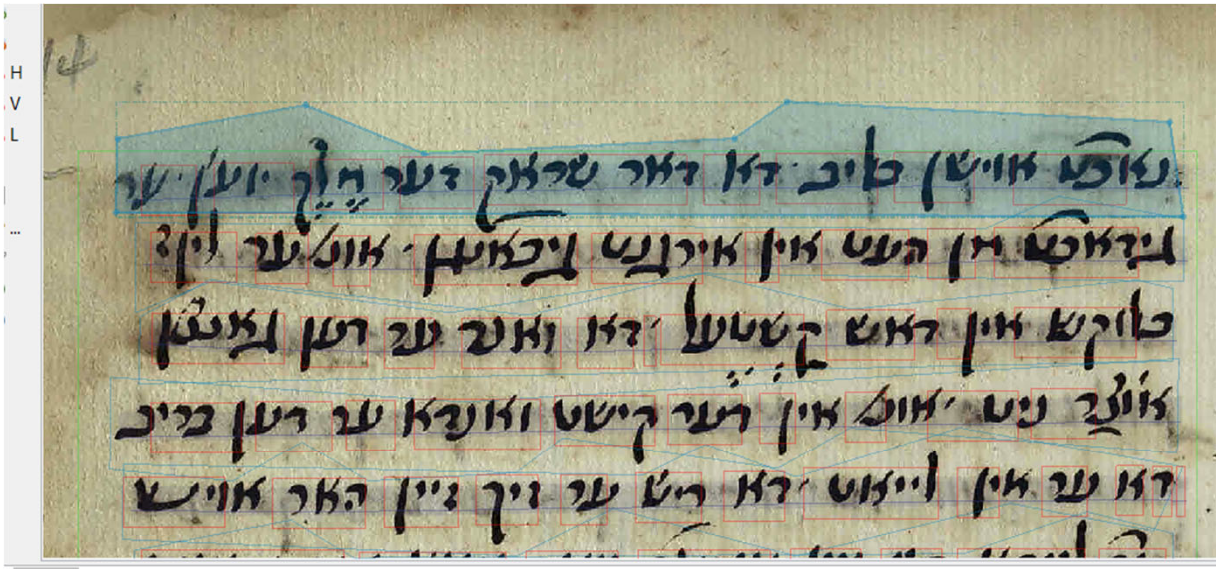
1-1 все страшно дорого и достать негде один_
1-2 полк_ стоит_ в год_ 3. миллиона. Когда я вышил_
1-3 из госпиталя то мне навстречу <name> и
1-4 мы с ним_ пошли гулять встретил_ Доктор <name>
1-5 затем_ он_ ушел_ и мы пошли в дом_ товари
1-6 щества к_ сестрам_ тал_ сидел_ ио 8 ч. муся»
1-7 шлу
1-8 у <name> пошли к_ <name> вмлесте с
1-9 нами. Вернулись мы поужинали и разошлис
1-10 Я стала писать дневник_ слышу пришл_ офица

Automatische Modernisierung:
ѣ > е
і > и

Auslassung von ъ am Ende der Wortformen

Jiddisch (16. Jh.)

- Transliteration
- Direktionalität
- Diakritika
- Zugang auch für Nicht-Spezialist*innen
- Zwei öffentliche Modelle
- Ähnlich auch für Osmanisch



1-1 nacht oušėn blib, do dar-schrakh der melech joel er
1-2 gėdocht mȧn het in irgėnt gėbangėn. un` er lif-
1-3 blukś in daś kaštęl , do vand er den ganzen
1-4 ozer nit. un` in der kiśt vand᾽ er den brib
1-5 er d o er in lai᾽t, do diś er sich seiėn hor ouś

- Dank an Astrid Lembke (Mannheim)

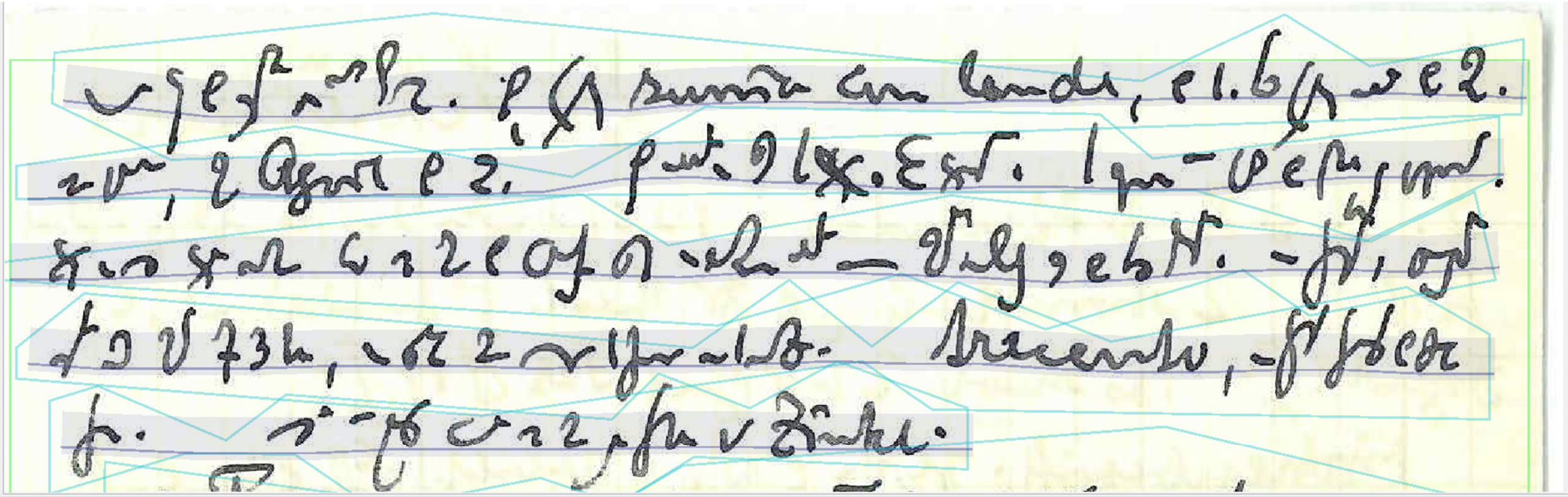
Stenographie

- Endgegner von HTR: keine Eins-zu-eins-Relation, teilweise extrem verkürzt
- Optisches Signal tw. weniger aussagekräftig → starkes Sprachmodell (und damit große Menge an Trainingsdaten) nötig
- Drei Systeme: Stolze-Schrey (eher Schweiz/eher protestantisch), Gabelsberger (eher Deutschland/eher katholisch), Deutsche Einheitskurzschrift
- Vier öffentliche Modelle (2x Gabelsberger)

- Tw. synthetische Trainingsdaten
- Verbessert CER, aber nicht Real-World-Performance

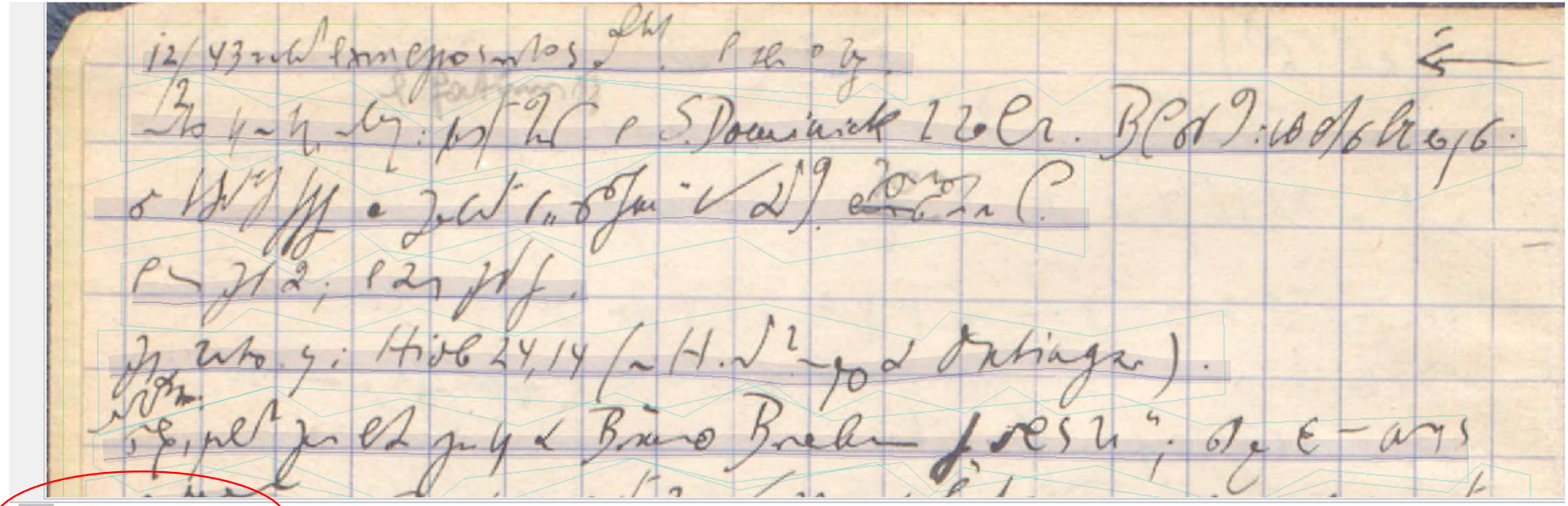
Name Search	Wörter	CER	Sprache
Gabelsberger_natural	429 663	13.38%	GER
Faulhaber	259 537	12.17%	GER
Stolze-Schrey_combined	23 023	9.50%	GER
DEK_German_combined	144 709	9.50%	GER

Erfolgreiche HTR (Faulhaber Gabelsberger)



1-1 Ohne auf die heutige Verhältnisse zikommen. Die Trüftung summar cm lande, die ersten Staatsprüfung als das zwei.
1-2 in Bayern im Aprrt die zweite. Dafür sollte er sich nicht dürüber a wieder überabiten. Nicht krank sei bis der eigentliche Befuf beginnt
1-3 würder es überalnehmen wenn ich ihm die wirtschaftlichen Sorge erlichtern wollte - mit der Entscheidung haben das nichts zu tun. Er zegt, sein Vater
1-4 arbeite noch mit 73 Jahren, er selber habe keine Geschenke angenommen. trecento, er ziht zuerst die Hand
1-5 zurück. Es war eine Pause weil ich im Vorzimmer bei Zinkl.

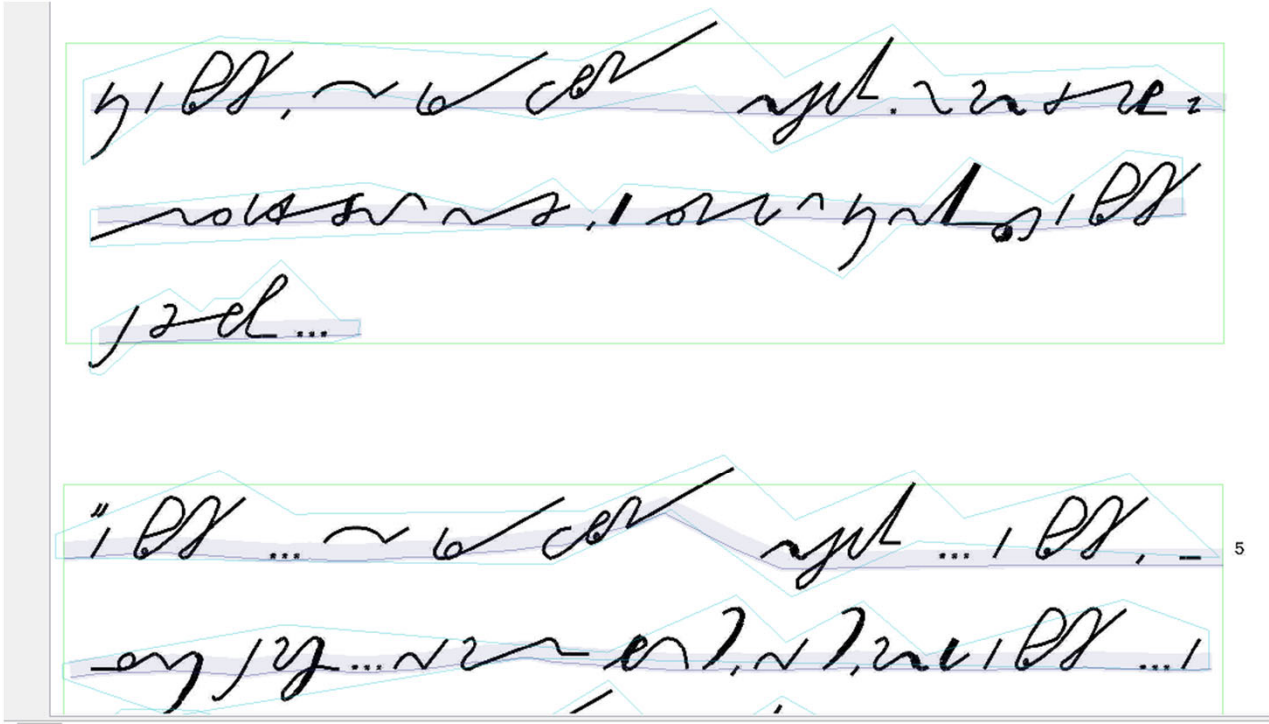
Nicht erfolgreiche HTR (Generisch Gabelsberger)



1-1 e de me e e e
1-2 nachs Brief an Junger Enweruf. bleghte Sittation die aSominik im Malstom. I Sta sagt sich ab bisse durch sie stinme als für sie.
1-3 selche behrtischen Ppaize in Schulwerte wie in Esseziell war bei von asn
1-4 die Urschlag an; die Henk schligt zu.
1-5 ch Megs. auf: Hoo 2414 (antat. Unter dem Anfluss von Ontinge).
1-6 sonderbar befedete Schnecke durch den schnen Brief von Buro Brekem zu Land und mehr wirt; sah plötzlich wieder eine Werk und

Erfolgreiche, aber synthetische HTR (Stolze-Schrey)

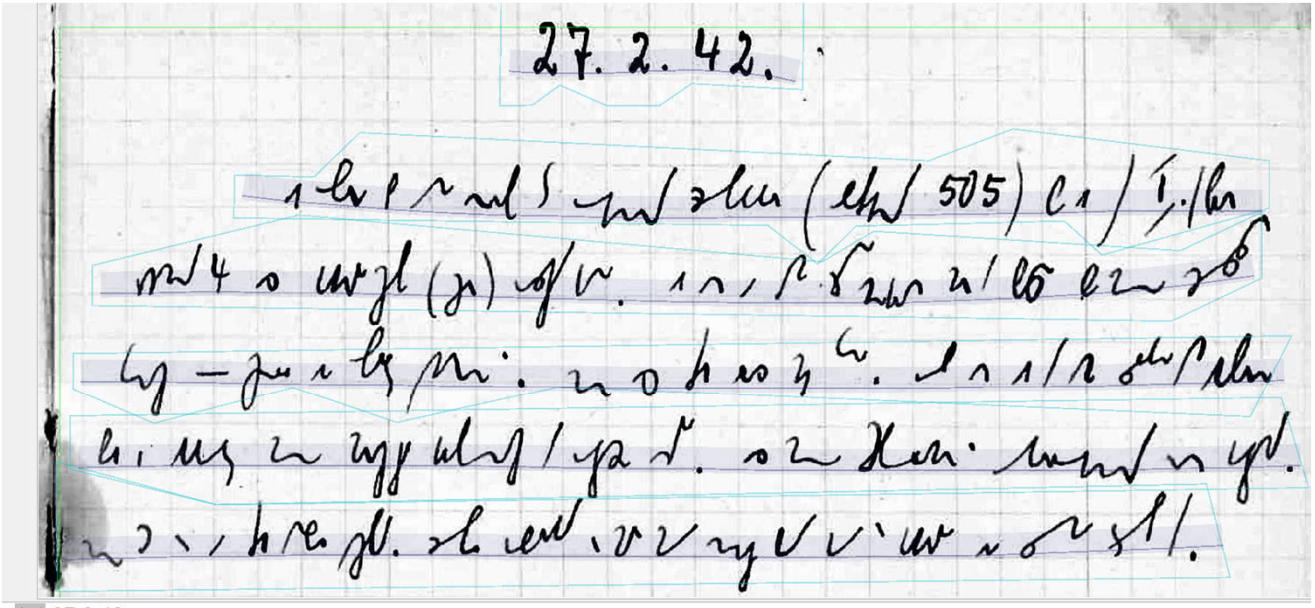
- Modell leider zu klein für vernünftige Real-World-Performance



1-1 doch die Pflicht seine Beretwilligkeite anzubitten wenn man jemanden in
1-2 einer Bedrängnis sehe, da ergebe sich doch natürlich die Pflicht
1-3 zu helfen ...
2-1 die Pflicht ... seine bereitwilligkeit anzubitten die Pflicht, den
2-2 versuch zu machen ... Sie meinen Also auch, Sie auch, man habe die Pflicht ... die

Erfolgreiche natürliche HTR (DEK)

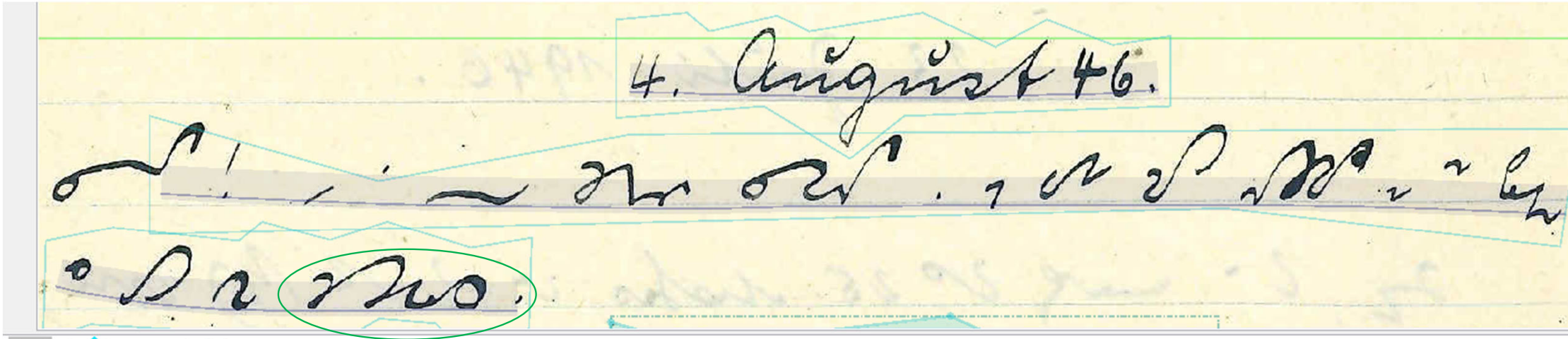
- Booster durch Kombination von natürlicher und synthetischer GT



1-1 27.3.42.
1-2 Ich erfahre durch einen Anruf von Leutnant Hofweber (adjutant 505) dass ich zur I,/Flak
1-3 Regiment 4 als Batterie schaff (Schwer) versetzt bin. Ich kann es im erste Augenblick gar nicht fassen dass Mein heißest
1-4 Wunasch so schnell in Erfühung gegangen ist. Nun muss ja alles gut warde. Leider kann ich nicht am seelben Tag abfahren
1-5 da die Abteilung mein Marschziltelefönisch nicht verstehen konnte. Als mein Eechfolger ist Oberleutnant Reck bestimmt.
1-6 nun hate er es ja endlich geschaft. Hoffentlich verdilebt er mir meine kurz aArbeit bei der Batterie in Seinem überort nicht.

Qualitativ tw. besser als Expert*innenwissen

- In GT hier < Hoch >, da Transkriptionsexpertin dies nicht entziffern konnte
- Hochablass: Stauwehr am Lech in Augsburg



1-1 4. August 46
1-2 litg ! Es wirdt ein herrlicher Sommertag. Ich biege mich nachmittags in den Fluten
1-3 des Lech am Hochblas.

Wie weiter?

- Smarte und komplexe HTR mit beachtlichen Ergebnissen, aber dennoch teilweise an der Grenze der Brauchbarkeit für manche Einsatzzwecke (Steno-Editions-Community…)
- LLMs zur Post-Korrektur? Unseren Experimenten zufolge nur bei modernen Resourced Languages hilfreich, bei anderen nicht hilfreich bis katastrophal schädlich
- Bessere Engines? Transformer (TrOCR) hilfreich, erste ermutigende Ergebnisse für latein-basierte Schriftsysteme (Romein et al. 2024)
  - Benutzerfreundlichkeit? Skalierung? Hardware?
- Andere Engines/Systeme? Wie sinnvoll Transkribus-GT nachtrainieren?
  - eScriptorium Trans/Conformer, PERO Transformer, OCR4all?
- Reflexion über Aufgaben/Zielgruppen/Outreach/GT- und Modell-Management