Training ‚smarter' Modelle im MultiHTR-Projekt - Ergebnisse und Herausforderungen

Milanka Matić-Chalkitis

Albert-Ludwigs-Universität Freiburg

Gefördert durch

Baden-Württemberg

MINISTERIUM FÜR WISSENSCHAFT, FORSCHUNG UND KUNST

Inhalt

- 1. Das MultiHTR-Projekt
- 2. ‚Smarte' Modelle
- 3. Ergebnisse
- 4. Herausforderungen
- 5. Ausblick

1. MultiHTR

- Laufzeit:

Juni 2020 – Mai 2022

Ort:

Slavisches Seminar

Albert-Ludwigs Universität Freiburg

- Team:
- Projektleiter Prof. Dr. Achim Rabus
- MA Lesley Loew / Dr. Aleksej Tikhonov
- MA Milanka Matić-Chalkitis
- BA Martin Meindl

ProjektpartnerInnen:

Prof. Dr. Veronika Lipphardt

(Science and Technology Studies)

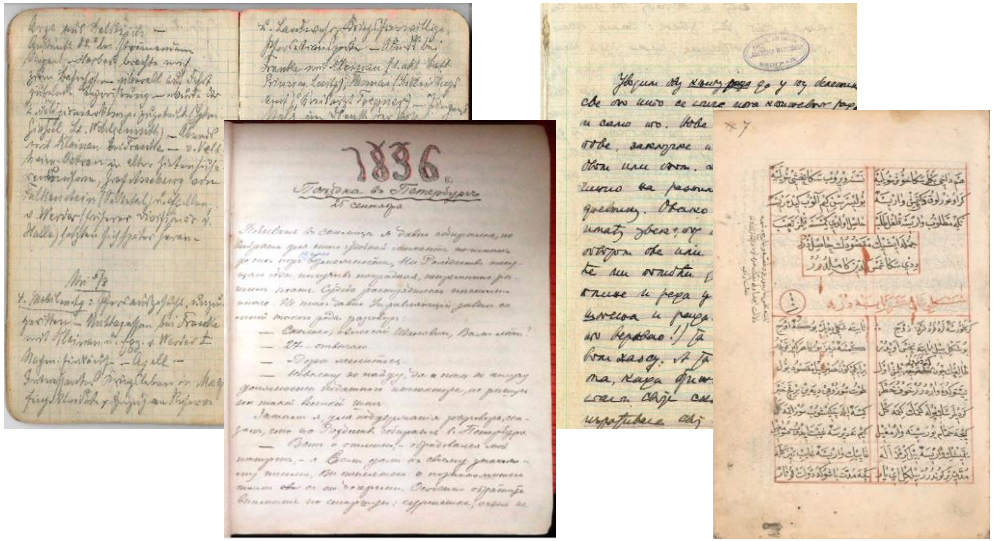
Prof. Dr. Johanna Pink

(Islamwissenschaft)

1. MultiHTR

Ziel

Erstellung und öffentliche Zurverfügungstellung mehrerer für unterschiedliche Handschriften geeigneter HTR-Modelle (Transkribus) für



Deutsch (Kurrent),

Russisch (Kyrillisch)

Serbisch (Kyrillisch)

Osmanisch (Arabisch)

1. MultiHTR

Um

1. Zugang zu multilingualem persönlichen bzw. kulturellen Erbe (für entsprechende Communities),
2. Zugang zu HTR-Werkzeugen zum ‚einfachen' Entschlüsseln von Handschriften,
3. Transformation „analogen" Wissens ins digitale Zeitalter

zu ermöglichen

1. MultiHTR: Ausgangslage

- Heterogenes handschriftliches Material
- in verschiedenen Sprachen
- sowie verschiedenen Schriften/arten (Russisch-Kyrillisch, Serbisch-Kyrillisch, Osmanisch-Arabisch, Deutsch-Kurrent)
- mit jeweiligen spracheigenen Besonderheiten

- Bedarf an verschiedenartigen Modellen,

u.a. Modellen mit besonderen Eigenschaften

‚smarte' Modelle

1.MultiHTR: HTR-Modelle

- Erstellung und Training verschiedenartiger HTR-Modelle

  - Spezifische Modelle (i.d.R. kleinere Modelle mit einem spezifischen Merkmal: z.B. ein/e Schreiberin, ein bestimmter Zeitraum, Schreibstil, usw.)

  - Generische Modelle (i.d.R. große Modelle mit mehreren Merkmalen: z.B. mehrere SchreiberInnen, mehrere Sprachen und/oder Schriften, verschiedene Zeiträume usw.)

  - Modelle mit ‚smarten' Eigenschaften (können sowohl spezifische als auch generische Modelle sein)

2.‚Smarte‘ Modelle

- Modelle mit Eigenschaften, die über reine eins-zu-eins Transkription einzelner Zeichen hinausgehen
- erweiterte/zusätzliche Eigenschaften:

  - Transkription in eine andere Schrift (z.B. arabisches -> lateinisches Alphabet)
  - Auflösung von Abbreviaturen, Ligaturen
  - Ergänzung von nicht ausgeschriebenen Zeichen

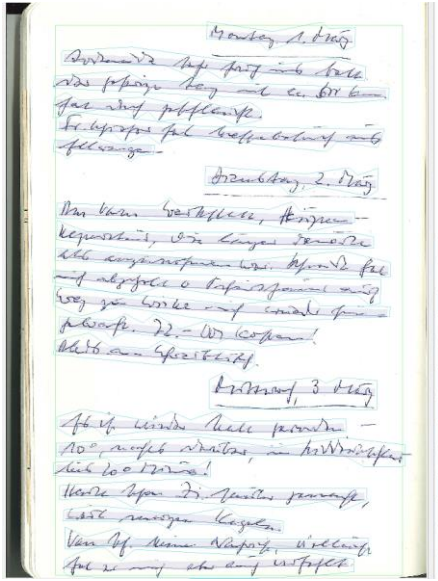
  usw.

3. Ergebnisse: **Deutsch**

- Beispiel: Spezifisches HTR-Modell

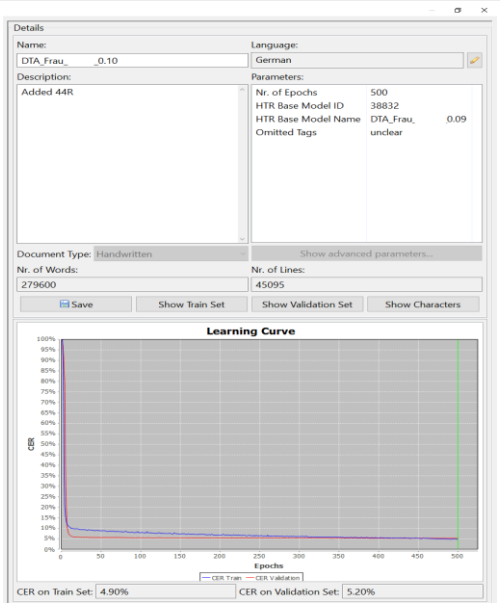
- Trainingsdaten aus dem Deutschen Tagebucharchiv Emmendingen (**DTA**)
- Tagebücher einer besonders produktiven Schreiberin (Frau B.)

3. Ergebnisse: Deutsch



1-1 Montag, 1. März.
1-2 Todmüde sehr früh ins Bett
1-3 Der gestrige Tag mit ca 500 km.
1-4 hat doch geschlaucht.
1-5 Fr. Schiefer hat Bettelbrief aus
1-6 Ellwangen.
1-7 Dienstag, 2. März
1-8 Am Vorm. Werkstatt, Huper
1-9 Reparatur, die länger dauerte
1-10 als anzunehmen war. Schmitt hat
1-11 mich abgeholt + Scheitsames auf
1-12 Weg zum Wicke mich wieder hin-
1-13 gebracht. 72.- DM Kosten!
1-14 Kleis im Schreibtisch
1-15 Mittwoch, 3. März
1-16 Es ist wieder kalt geworden
1-17 10 Grad, nachts darüber im Kirdonnschland
1-18 bis 20 Minus!
1-19 Habe schon Zi. sauber gemacht,
1-20 weil morgen Kegeln.
1-21 Von Sch. keine Nachricht, vielleicht
1-22 hat er mich aber auch verfehlt.

Beispielseite mit automatisch generierter Transkription aus einem Tagebuch der Frau B. (DTA)



Modell: DTA_Frau_B_0.10
CER (VS): 5.20%

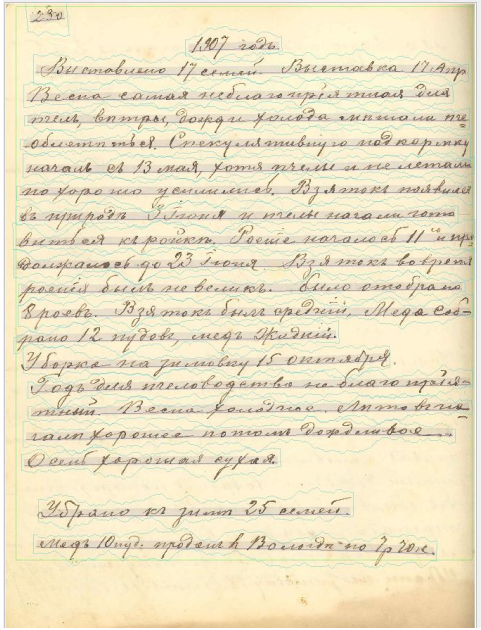
3. Ergebnisse: **Russisch**

- Beispiel: generisches Modell

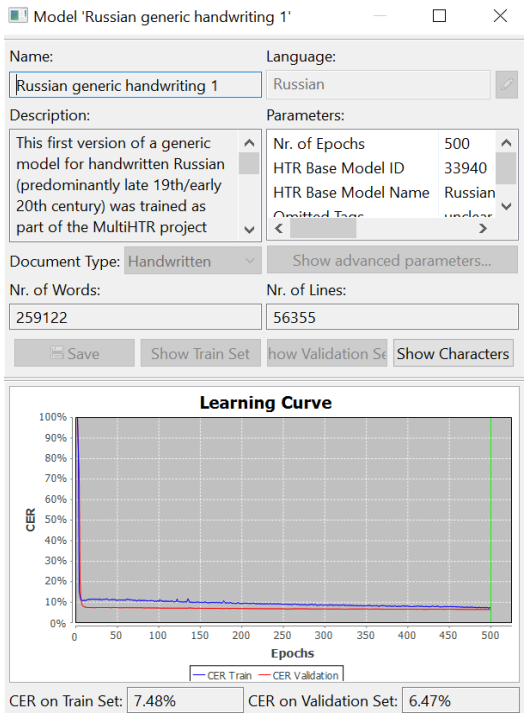
- Trainingsdaten aus dem Projekt ‚Прожито‘ an der unabhängigen/nicht-staatlichen Universität St. Petersburg

- Tagebücher mehrerer SchreiberInnen aus verschiedenen Jahrzehnten

3. Ergebnisse: Russisch



1-1 230
1-2 1907 год
1-3 Выставлено 17 семей. Выставка 17 Апр
1-4 Весна самая неблагоприятная для
1-5 плеч, ветры, дожди холода мешали пчелам
1-6 облететься. Спекулятивную подкормку
1-7 начал с 13 мая, хотя пчелы и не летали
1-8 по хорошо усилились. Взяток появился
1-9 в природе 3 Июня и пчелы начали гото¬
1-10 виться к ройке. Роение началось 11го про¬
1-11 должалось до 23 Июня. Взяток во время
1-12 роения был не велик. было отобрано
1-13 8 роев. Взяток был средний. Меда соб¬
1-14 рано 12 пудов, мед Жидкий.
1-15 Уборка на зимовку 15 октября.
1-16 Год для пчеловодства не благоприя¬
1-17 тный. Весна холодное. Лето вна¬
1-18 чале хорошее потом дождливое.
1-19 Осень хорошая сухая.
1-20 Убрано к зиме 25 семей.
1-21 Мед 10 пуд. продан в Вологде по 7р 70к.



Beispielseite aus einem Tagebuch mit automatisch generierter Transkription, 1907 (Прожито)

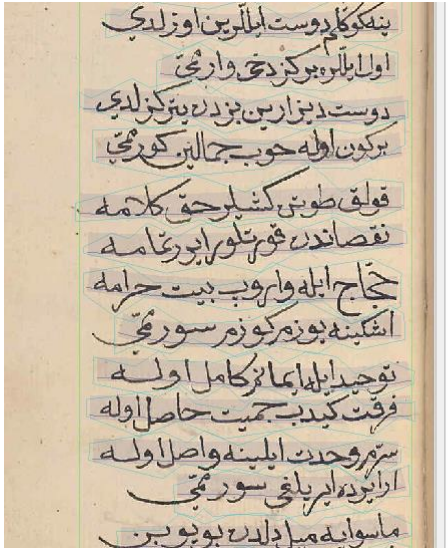
Modell: Russian generic handwriting 1
CER (VS): 6.47%

3. Ergebnisse: **Osmanisch**

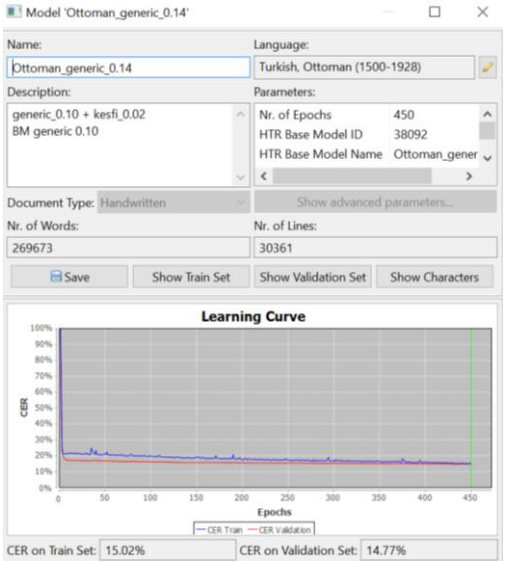
- Beispiel: generisches Modell mit ‚smarten' Eigenschaften

- Trainingsdaten akquiriert nach dem ‚data recycling' Prinzip aus verschiedenen Quellen
- ‚smarte' Eigenschaften:
  - Transkription aus der arabisch-persischen in die lateinische Schrift
  - Anpassung der Schreib- und Leserichtung (RTL-Schreib- und Leserichtung des Ausgangstextes und LTR-Richtung der Transkriptionen)
  - Ergänzung der Vokale sowie Anpassung der Vokalharmonie
  - Auflösung von Ligaturen
  - usw.

3. Ergebnisse: Osmanisch



1-1 Yinegöñle dost illerin uzildı
1-2 Ol eillere bir gez daḫı varnaḫı
1-3 Dost dīzarın bizden yetr gizldi
1-4 Bir gün ola ḫüb cemālin görümci
1-5 Ḳuluḳ ṭutan kişiler ḥaḳḳ kelāma
1-6 Noḳṣāndan ḳurtulur ėrrʿāma
1-7 Ḥücāc ile varub beyt i ḥarāme
1-8 Eşkine yüzüm gözüm sürmi
1-9 Tevcīd ile īʿānı kāmil ola
1-10 Fürḳat gidüb camīyt ḥāṣıl ola
1-11 Serrr vūḥdet ėline vāṣıl ola
1-12 Ar yėrde ayrıblağı sürümi
1-13 Māsevāya mel dilden buyünın
1-14 Baḥr i ʿālem Muṣṭafāsına uyubın



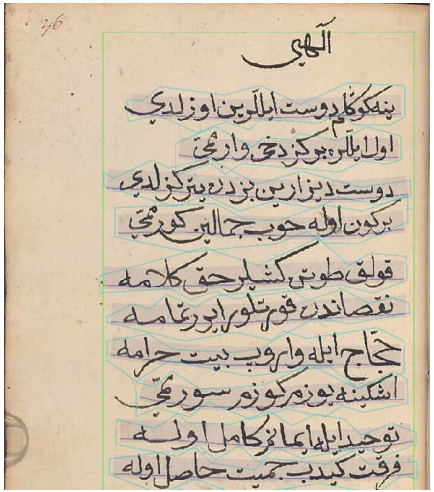
Beispielseite einer osmanischen Handschrift
mit automatisch generierter Transkription
aus der ‚Mecmua' Kollektion (Österreichische Nationalbibliothek)

Modell: Ottoman_generic_0.14
CER (VS): 14.77%

4. Herausforderungen

- Osmanisch als Sonderfall -> Vielzahl an Herausforderungen
  - Sprachkomplexität (lexikalisch, phonologisch, orthografisch, grammatisch)
    - z.B. das Fehlen der Vokale

vokalisiert



nicht vokalisiert

4. Herausforderungen

- Paläografische Besonderheiten

  z.B. 6 (Grund)Schriftarten

  Neskhî (Nesih) = Druckschrift

  Ruq‘a (Rıqa‘)

  Thuluth (Sülüs)

  Diw(v)ani

4. Herausforderungen

- keine eins-zu-eins Korrespondenz einzelner Zeichen (Original - Transkription)
  - Vokale u,o,ü,ö und ev werden durch او wiedergegeben
  - z.B. mehrere arabische Konsonanten س ص ث

    entsprechen nur einem türkischen Konsonanten, s

- kein einheitliches Transkriptionssystem
  - 4 verschiedene Transkriptionssysteme : **IJMES** (International Journal for Middle Estern Studies); **EI²** (Encyclopaedia of Islam, 2nd Edition); **İA** (İslam Ansiklopedisi) und **DMG** (Deutsche Morgenländische Gesellschaft)

4. Herausforderungen

- Quantität sowie Qualität der Trainingsdaten
- Heterogenität des handschriftlichen Materials -> aufwendiges preprocessing und manuelle (Vor)Korrekturen
- unterschiedliche Erwartungen bezüglich der Qualität an automatisch generierten Transkriptionen (für philologische Analysen in der Wissenschaft vs. leichtere Lesbarkeit für die breitere Öffentlichkeit)

  -> verschiedenartige Modellen

5. Ausblick

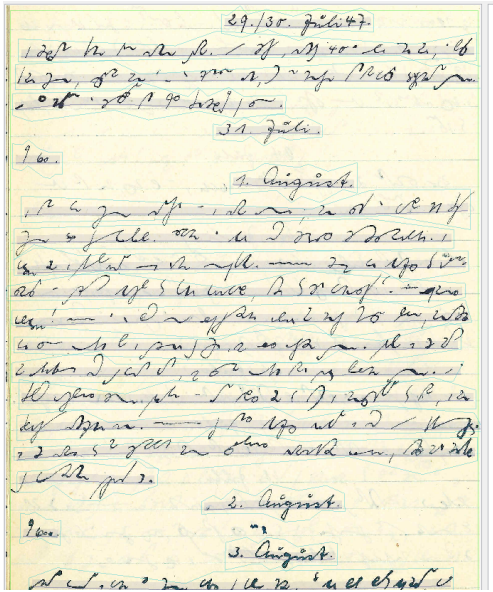
- MultiHTR2 Juni 2022 – Mai 2024

- Ziel:

- Optimierung sowie Erweiterung bestehender HTR-Modelle für Osmanisch, Deutsch und Russisch

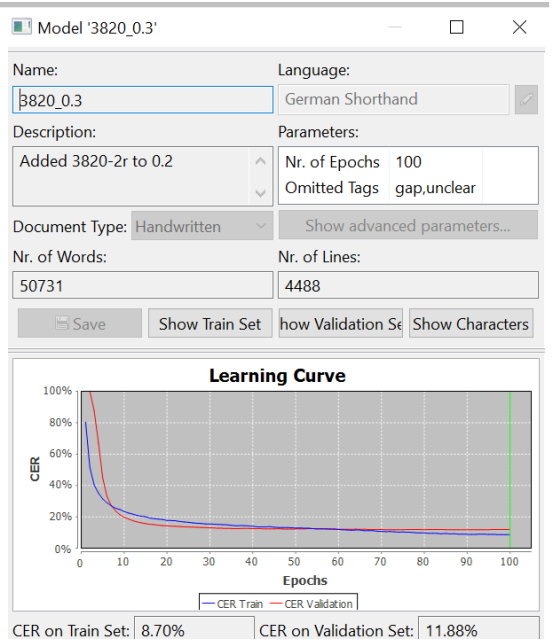
- Erstellung neuer HTR-Modelle für weitere Sprachen wie Ukrainisch, Jiddisch sowie deutsche stenografische Handschriften

5. Ausblick: Stenografie



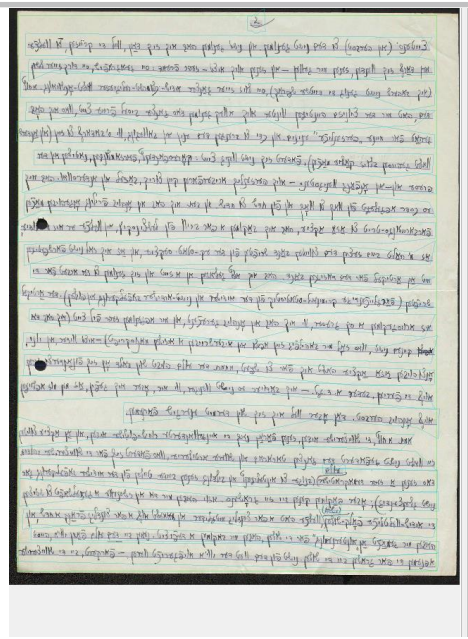
1-1 29./30. Juli47.
1-2 Die Hundstage tragen ihren namen zurecht. Eine Hitze, nacheuz 40 ° oder sogar mehr, ist fasst
1-3 nicht mehr schön, außerdem man wird in der glücklichen Lage, sich den ganzen Tag im Wesser aufzuthalten zu göngen.
1-4 Es soll gestern der heißte Tag dieses Jahrhundert zu sein.
1-5 31. Juli.
1-6 Nichts besonderes.
1-7 1. August.
1-8 Die Tage werden schone Kirzer und die Nacht langer; man sagt der Wind geht jetzt
1-9 Schon bu Stobelfeld. Desmegen ist aber noch herrliches Hochsommerwetter. Die
1-10 Bauern haben die getreif erante ohne Regen eingebracht. Heute war Abschlus vom Millke
1-11 Semmester und zugleich Abschied von Patter Bernwand, welcher von hier wegversezt wirde . ausgemeines
1-12 Bedauern ! Ist er doch ein ausgezeichneter Redner mit ganz großen ien, manchmal
1-13 war sein Vortrag für die Zuhörer zu schwer, um alles

Beispielseite aus einem Tagebuch mit automatisch generierter Transkription, 1947 (DTA)

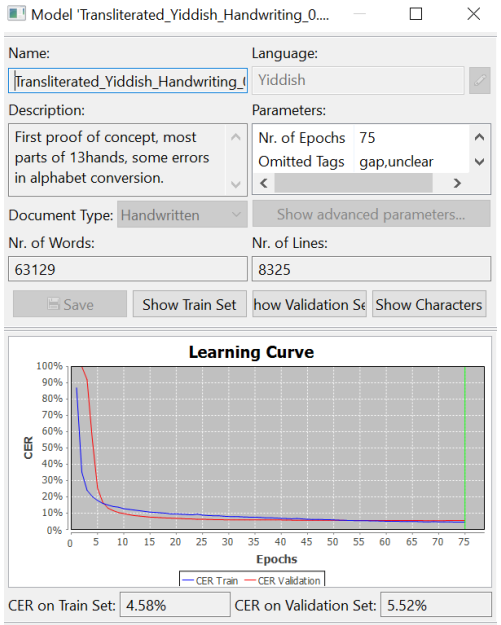


Modell: 3820_0.3
CER(VS): 11.88%

5. Ausblick: Jiddisch



1-1 2
1-2 tsaytens (in herbst) tsu dem nisht genumen. un nisht genumen hab ikh zikh dan, vayl di krazen, tsu velkhe
1-3 men darf zikh venden, zenen mir geven - un zenen oykh itst - zehr ɔremd! say geagrafish, say durkh zayer lshun
1-4 (ikh bahersh niyht genug di daytshe shfrakh), say luiz zayer gantser idish- ksmrsht-keligyezer velt -onshuung. a khut
1-5 dem, hat mir der tsiunizm hayntigen vinter oykh uuek genumen das gantse bisel fraye tsayt, vas ikh hob
1-6 gehat far mayne, ferzenlikhe eninim, un khdi tsu brengen dem enin in buuegung, vi s badarf tsu zayn (un andersh
1-7 valt gehaysen bluiz kalye makhen), fadert zikh nisht venig tsayt. koresfandents, farzamlungen, natitsen in der
1-8 frese un -in onɔang venigsten. - oykh ferzenlik ariberfaren kayn tsurik, bakhel un andersvu. hab ikh
1-9 es khsdr afgelegt ɔun vakh tsu vok un fun khdsh tsu khdsh. un bes ikh hab in onhuib ɔriling ongehuiben makhen
1-10 farbaraytungs-trit tsu aza aktsye, hab ikh bakumen a far briv fun leshtsinskin, in velkhe er iz amudye
1-11 az mi halt baym zetsen dem tsvayten band shriften ɔun der ek-stat. sektye, un az ikh zal nisht farshfetigen
1-12 mit on artikel far dem dazigen band. hab ikh alts gelasen an a zayt un zikh genumen tsu der arbet far di



Beispielseite einer jiddischen Handschrift mit automatisch generierter Transkription

Modell: Transliterated_Yiddish_Handwriting_0.1
CER(VS): 5.52%

Danke

Kontakt:

milanka.matic@slavistik.uni-freiburg.de

@MultiHTR

www.multihtr.uni-freiburg.de