ÖAW AUSTRIAN ACADEMY OF SCIENCES

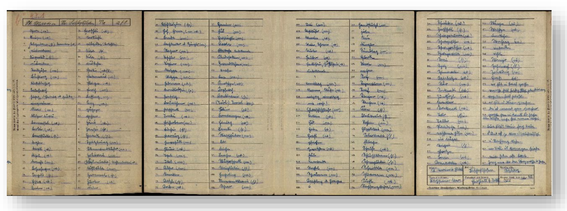
austrian centre for digital humanities & cultural heritage

Vom Digitalisat zur Ressource: Der Workflow zu „DWA Österreich Pilotstudie“

WORKFLOW

VORARBEITEN

- Scans (TIF) konvertieren: TIF → JPG
- Qualitätskontrolle: Farbkontrast verstärken

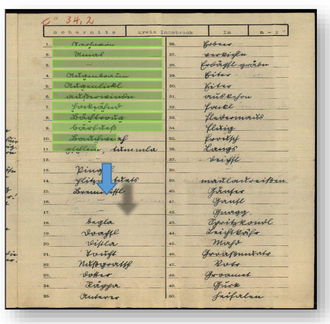


UPLOAD

- Upload der Scans nach Transkribus

LAYOUT-ANALYSE

- Zuweisen von Textregionen
- Zuweisen von Zeilen (Baselines)

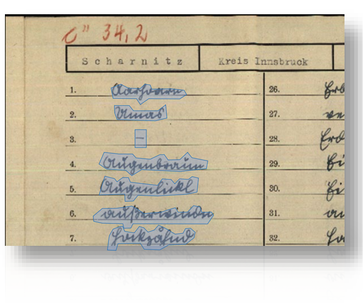


HTR

- automatische Transkription durch HTR-Modell auf der Basis korrekter DWA-Bögen

KORREKTUR

- manuelle Korrektur der transkr. Bögen

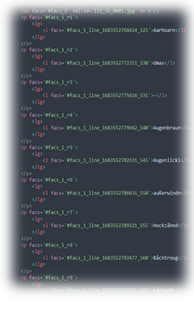


1-1 Aarhoarn
2-1 Umas
3-1
4-1 Augenbraun
5-1 Augenlickl
6-1 außerwinden.

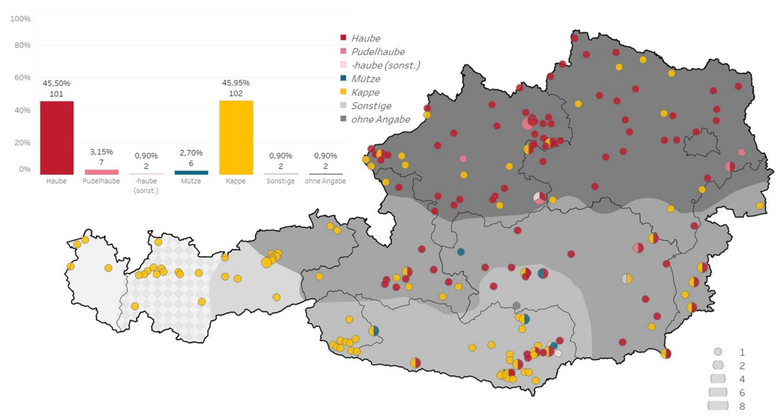
EXPORT

- Export über API

1-1 Aarhoarn
2-1 Umas
3-1
4-1 Augenbraun
5-1 Augenlickl
6-1 außerwinden.



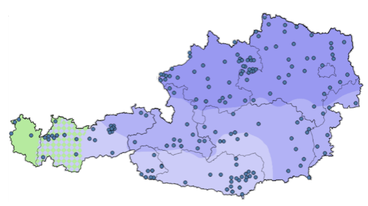
ANALYSE



PILOTSTUDIE

DWA Österreich Pilotstudie

Projektteam: (10/2022–11/2023)
Alexandra N. Lenz
Markus Kunzmann
Amelie Dorn
Veronika Höbart
Paulina Huber



Kooperationspartner:
- ÖAW: Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH)
- Universität Wien: SFB „Deutsch in Österreich“
- Philipps-Universität Marburg: Forschungszentrum Deutscher Sprachatlas

Ziel:
- Aufbau eines Workflows
- HTR-Modell zur Volltextdigitalisierung von ca. 3.700 Erhebungsbögen
- Testdatensample: 200 Bögen
- Evaluierung

Wissenschaftliche Einordnung

(+)
- flächendeckende lexikalische Erhebungen für Österreich
- umfangreiche Quelle Beginn 20. Jh.
- gleichbleibende Erhebungsmethode
- lexikalische und geostatistische Analysen der lexikalischen Dialektlandschaft
- Analysen zu lexikalischen Dialektlandschaft in Österreich
- diachroner Vergleich mit aktuellen Sprachdaten lexikalischer → lexikalischer Dialektwandel in den letzten 80 Jahren

(-)
- indirekte Erhebung Laien
- beschränktes phonetisches Inventar (Standard-Alphabet)
- Priming durch Stimuli

Deutscher Wortatlas (DWA)

Mitzka, Walther & Ludwig Erich Schmitt. 1951 – 1980. Deutscher Wortatlas. Gießen: Schmitz.

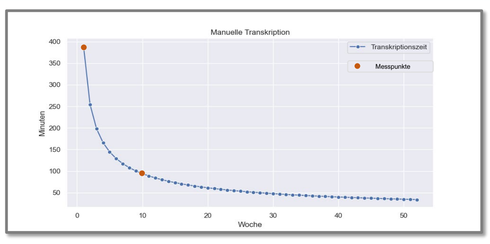
- 22 Bände | 240+ Karten
- Von Walther Mitzka und [ab Band 5] Ludwig Erich Schmitt, [ab Band 18] redigiert von Reiner Hildebrandt
- Als Ergänzung zum „Deutschen Sprachatlas“ (1927 – 1956) (DSA)

Deutscher Wortatlas (DWA)

- 188 Einzelwörter, 12 Sätze
- Erhebungszeitraum: 1939 – 1942
- Methode: indirekt mittels ausgesendeter Fragebögen (Laienverschriftung)
- ca. 50.000 Ortspunkte (orientiert am Ortsnetz des DSA), davon 3.700 in Österreich

BEWERTUNG

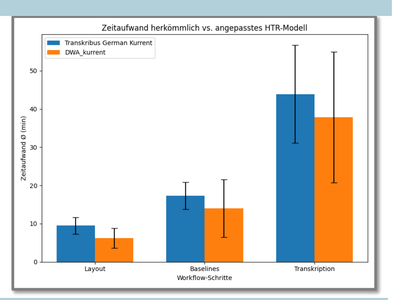
Erstellung Trainingsset – Zeitlicher Rahmen



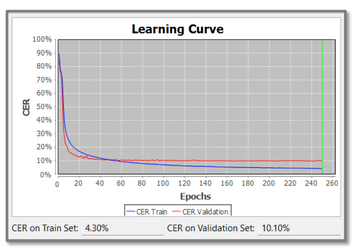
Für die Erstellung eines Ground-Truth-Datensatzes von 100 Bögen (Training: 75; Validierung: 25) werden ca. zehn Wochen (40 h/Woche) benötigt. Betrifft die Arbeitsschritte Layout, Baselines und manuelle Transkription.

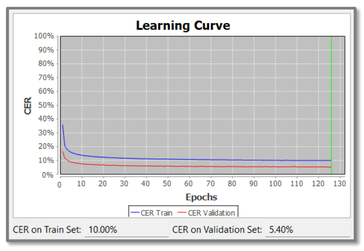
Dauer Arbeitsschritte nach Modell

Mittelwerte der benötigten Zeit für die Arbeitsschritte zur Anpassung von Layout und Baselines sowie der Korrektur der automatischen Transkriptionen je Bogen. (Stichproben *Transkribus German Kurrent*: n=5; *DWA_Kurrent*: n=10)



Modellkarte



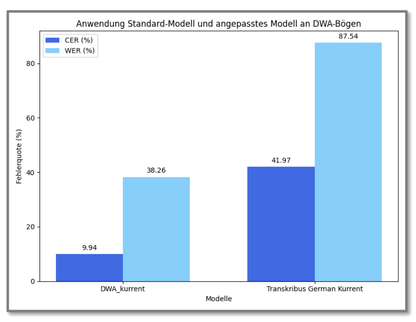


HTR-Modell (public): *Transkribus German Kurrent*

HTR-Modell (privat): *DWA_Kurrent*

Modell	nrOfWords	CER (Val.set)	CER (DWA)
Transkribus German Kurrent	3.209.689	5,4	41,97
DWA_Kurrent	28.274	10,1	9,94

Modellkarte



CER (Character Error Rate): Lässt nur eingeschränkt eine Eignung eines HTR-Modells im Bezug auf die spezielle Anwendung zu.

POSTER CREATOR Markus Kunzmann
markus.kunzmann@oeaw.ac.at

WBÖ

Lexikalisches Informationssystem Österreich

Deutsch in Österreich
Variation – Kontakt – Perzeption