

Messprotokoll: Maschinelle Transkription gegen Referenztexte des Deutschen Textarchivs

Messprotokoll MP-2026-001 · Version 1.1 · Stand: 13. August 2026 · Messung durchgeführt: 11. August 2026 · Verantwortlich: Alaa Mroue, alaa.mroue@hexworld.eu, +49 351 25065907

Ergebnis: 17 Edits auf 5.129 Referenzzeichen über drei Seiten (1591, 1615, 1701), Zeichenübereinstimmung 99,67 % (Abschnitt 3). Vollständige Abweichungsliste in Abschnitt 4, Belegausschnitte in Anhang A.

1. Gegenstand

Drei Seiten aus dem AEDit-Bestand des Deutschen Textarchivs wurden maschinell allein anhand des Faksimiles transkribiert und anschließend mit der Textfassung des DTA kollationiert. Werk- und Seitenauswahl, Erfassungskonventionen und Metrik wurden vor Beginn schriftlich festgelegt (Vorregistrierung).

Seite	Vorlage (DTA-Nr., Bild)	Jahr
E1	Wenzel: Von rechter Prob vnd merckzeichen Christgläubiger menschen (Frankfurt/Oder); DTA 511528, Bild 20	1591
E2	Pia ad Mortem Praeparatio. Einfaltige Predigt von Christlicher Vorbereitung; DTA 509377, Bild 12	1615
E3	Das unschuldig vergoßne Blut; DTA 421583, Bild 12	1701

Werkauswahl regelbasiert vorab, dokumentiert in der Vorregistrierung (je Zeitstratum: niedrigste DTA-Nummer des Stratum bzw. für E3 das dem Jahr 1700 nächstliegende Werk); Seitenauswahl je Werk: 40 % der Bildzahl, aufgerundet auf die nächste vollständige Prosaseite. E1 ist stark abgenutzt und liegt nur bitonal digitalisiert vor; E3 enthält einen lateinischen Antiqua-Einschub und gemischte r-Formen.

Referenztexte: Deutsches Textarchiv (BBAW), Lizenz CC BY-SA; Erfassungsart laut AEDit- und DTA-Korpusdokumentation: Double Keying. Wiedergabe einzelner Lesarten im Rahmen des Zitatrechts. Digitalisat-URLs in Abschnitt 6.

2. Vorgehen (Blindprinzip)

1. Transkription ausschließlich anhand des Faksimiles, ohne Kenntnis der DTA-Textfassung, nach den DTA-Richtlinien zur Texterfassung: Schaft-s (ſ, U+017F), u/v und I/J-Majuskel nach Regelwerk, r rotunda (ꝛ, U+A75B), aufgelöste Konsonantenligaturen, übergeschriebenes e (ē, U+0364), Kürzungsstriche (U+0303, U+0315), Silbentrennstrich (U+002D), Zeilenfall erhalten.
2. System: HexRead-Transkriptionspipeline (VLM-basierte Volltexterkennung, Stand 08/2026), mehrstufig: Ersterkennung, anschließend automatische Nachprüfung markierter unsicherer Stellen an vergrößerten Bildausschnitten. Die festgeschriebene Fassung ist das Ergebnis dieses maschinellen Ablaufs; eine manuelle Nachkorrektur erfolgte nicht. Die Unsicherheitsvermerke im Dateikopf stammen aus demselben Ablauf.
3. Festschreiben der drei Transkriptionsdateien; SHA-256-Prüfsummen dokumentiert vor dem ersten Abruf der Referenz (11.08.2026, 17:57:11 UTC): E1 d0e79cc0224c15bf..., E2 02e8d2b4e0ab856f..., E3 3c97541c02537321... (vollständige Werte im Reproduktionspaket).
4. Erst danach: Abruf der TEI-Fassungen, Seitenschnitt an den pb-Marken, Extraktion der jeweiligen Referenzseite, NFC-Normalisierung beider Fassungen, Vergleich per Levenshtein-

Distanz auf Zeichenebene, Zeilenumbrüche mitgezählt. Strukturelemente (Marginalien, Kolumnentitel, Kustoden) wurden separat kollationiert; sie sind nicht Teil der 5.129 Zeichen.

3. Ergebnis

Übereinstimmung = $1 - d/n$ (Levenshtein-Distanz d , Referenzzeichenzahl n); dies entspricht einer Zeichenfehlerrate (CER) von 0,33 % über alle drei Seiten.

Seite	Edits	Zeichen (Referenz)	Übereinstimmung
E1 (1591)	1	1.312	99,92 %
E2 (1615)	0	1.580	100,00 %
E3 (1701)	16	2.237	99,28 %
Gesamt	17	5.129	99,67 %

Strukturelemente: keine Abweichungen im Textbestand; einzige Differenz ist die gedruckte Seitenzahl "14" (E3), die die Referenz nicht kodiert (Konventionsebene).

95%-Konfidenzintervall (Wilson): 99,47 % bis 99,79 %, unter der Annahme unabhängiger Zeichenentscheidungen; da 16 der 17 Edits auf E3 liegen und nur drei Seiten gemessen wurden, ist das Intervall als Größenordnung zu lesen, nicht als Aussage über die Streuung im Bestand. Nach der Zählweise von Haaf/Wiegand/Geyken (JTEI, Issue 4, 2013; kombinierende Diakritika nicht als eigene Einheit): 17 Edits auf 5.082 Einheiten, Übereinstimmung 99,67 %.

4. Vollständige Abweichungsliste (17 Edits an 15 Wortstellen)

Jede Abweichung wurde nach der Kollation am Digitalisat (Vergrößerung bis 6-fach) geprüft und klassifiziert; die Zählung blieb davon unberührt. Beleg = Bildausschnitt in Anhang A.

Nr.	Stelle	Maschinelle Fassung	Referenz	Klassifikation	Beleg
1	E1, Z. 25	Glüchenden	Glühenden	Fehllesung maschinenseitig (c statt e im abgenutzten Schriftbild)	A1
2	E3, Z. 1	Natu2	Natur	r-Form ($r/2$), maschinenseitig	
3	E3, Z. 2	Jfraeliten	Ifraeliten	Konventionsebene: I/J-Majuskel-Regel unterschiedlich angewandt (im Druck nur eine gemeinsame Type für I/J)	
4	E3, Z. 2	leiblichen Söhnen	leiblichenSöhnen	Konventionsebene: Spatium im eng gesetzten Druck, am Druckbild nicht eindeutig entscheidbar	
5	E3, Z. 4	fahe (Schaft-s)	fahe (f)	Abweichung der Referenz vom Druckbild: Vorlage zeigt Schaft-s	A2
6	E3, Z. 4	das Land	dasLand	Konventionsebene: Spatium, am Druckbild nicht eindeutig	
7	E3, Z. 5	vergoffenen	ve2goffenen	r-Form ($r/2$), maschinenseitig	
8	E3, Z. 5	vergoffenen Blut	ve2goffenenBlut	Konventionsebene: Spatium, am Druckbild nicht eindeutig	
9	E3, Z. 6	vozhält	vorhält	r-Form ($r/2$), maschinenseitig	
10-12	E3, Z. 6	Ef. 1,15.59,3.	Ef. 1, 15. 59, 3.	Spatienregel (Leerzeichen nach Satzzeichen) im Zitat nicht angewandt, maschinenseitig (3 Edits)	
13	E3, Lat. Z. 2	Chananæi	Chanænæi	Abweichung der Referenz vom Druckbild	A3

Nr.	Stelle	Maschinelle Fassung	Referenz	Klassifikation	Beleg
14	E3, Lat. Z. 3	cta,	cta.	Abweichung der Referenz vom Druckbild (Komma mit Unterlänge sichtbar)	A4
15	E3, Z. 8	Jfrael	Ifrael	Konventionsebene wie Nr. 3	
16	E3, Z. 12	Mozdthat	Mordthat	r-Form (r/2), maschinenseitig	
17	E3, Z. 26	Mozd-Predigt	Mord-Predigt	r-Form (r/2), maschinenseitig	

Einordnung: 9 Edits sind maschinenseitig (1 Fehllösung der Buchstabenidentität, 5 Wahl der r-Form bei abgenutzten Typen, 3 Zitatspatien). 5 betreffen Konventions- und Grenzfälle (I/J: Nr. 3, 15; Spatien: Nr. 4, 6, 8). 3 betreffen Stellen, an denen die Referenz vom Druckbild abweicht (Nr. 5, 13, 14; Belege in Anhang A); Letzteres ist kein Befund gegen das Double-Keying-Verfahren. Die r-Form- und Konventionsklassen (Nr. 2-4, 6-12, 15-17) hängen von der projektspezifischen Festlegung der Erfassungskonventionen ab.

5. Grenzen der Messung

Das Blindprinzip und das Festschreiben stellen sicher, dass die Auswertung die Referenz nicht kannte. Sie schließen nicht aus, dass öffentlich verfügbare DTA-Texte in den Trainingsdaten des zugrunde liegenden Modells enthalten waren; eine Memorierung ist prinzipiell nicht auszuschließen. Gegen eine bloße Wiedergabe der Referenz sprechen die dokumentierten Abweichungen von ihr: die drei Stellen, an denen die Ausgabe der Referenz widerspricht und das Druckbild die Ausgabe stützt (Nr. 5, 13, 14), sowie die systematischen Konventionsdivergenzen (I/J, Zitatspatien). Abschließend klären lässt sich die Frage nur an Material ohne öffentliche Referenz, etwa an neuen Vorlagen.

Nicht Gegenstand dieser Messung: Layoutanalyse, Lesereihenfolge, Wortfehlerrate (WER), TEI-Strukturauszeichnung. Das Ergebnis gilt für die drei gemessenen Seiten; es ist kein Bestandsdurchschnitt.

6. Reproduzierbarkeit

Folgende Unterlagen liegen vor und werden auf Wunsch vollständig übergeben: die vorregistrierten Konventionen, die drei festgeschriebenen Transkriptionsdateien samt vollständigen SHA-256-Prüfsummen und Zeitstempel, die Referenz-Seitenschnitte, Belegausschnitte zu den Abweichungen sowie die Vergleichslogik als lauffähiges Skript (NFC, Levenshtein). Transkriptionen und Referenzschnitte werden als Klartext übergeben; das Ergebnis ist mit Standardwerkzeugen nachrechenbar (z. B. dinglehopper). Jede Zeile ist am öffentlichen Digitalisat nachprüfbar:

- E1: media.dwds.de/dta/images/511528/511528_0020_1600px.jpg
- E2: media.dwds.de/dta/images/509377/509377_0012_1600px.jpg
- E3: media.dwds.de/dta/images/421583/421583_0012_1600px.jpg

Ein früherer Einzelseiten-Blindtest (DTA 386427, Breslau 1641, 1.558 Zeichen) mit gesondertem Protokoll wird auf Wunsch ebenfalls übergeben.

Anhang A: Belegausschnitte

Vergrößerte Ausschnitte der öffentlichen Digitalisate (Quellen: Abschnitt 6). A1 belegt eine maschinenseitige Fehllesung, A2 bis A4 die drei Stellen, an denen die Referenz vom Druckbild abweicht.

A1 · Nr. 1 · E1 (1591), Zeile 25



Maschinelle Fassung Glüchenden, Referenz Glüchenden: das Druckbild stützt die Referenz (c statt e, Fehllesung maschinenseitig).

A2 · Nr. 5 · E3 (1701), Zeile 4

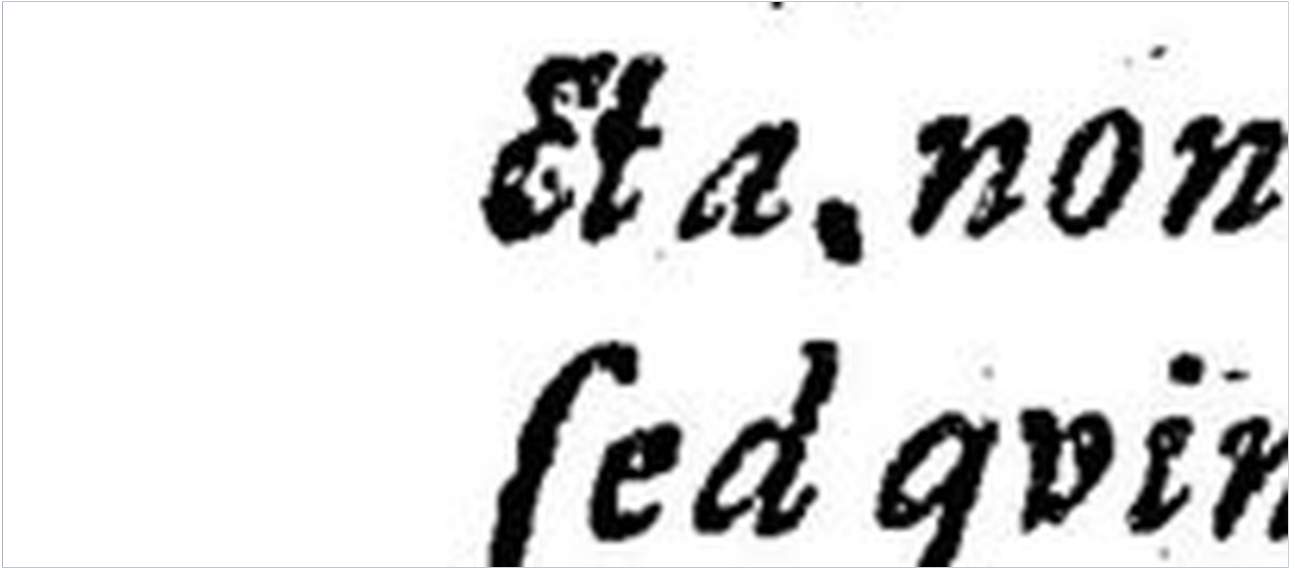


es fahe das Land: die Vorlage zeigt Schaft-s (f, ohne durchgehenden Querstrich); die Referenz kodiert fahe.

A3 · Nr. 13 · E3 (1701), lateinischer Einschub, Zeile 2



Die Vorlage zeigt Chananaei; die Referenz kodiert Chanænæi. Vergleichskriterium: das unstrittige æ (siebte Letter) zeigt zwei Binnenräume; die strittige fünfte Letter zeigt nur einen und entspricht dem a an dritter Stelle.



Die Vorlage zeigt nach cta ein Komma (Unterlänge sichtbar); die Referenz kodiert einen Punkt.