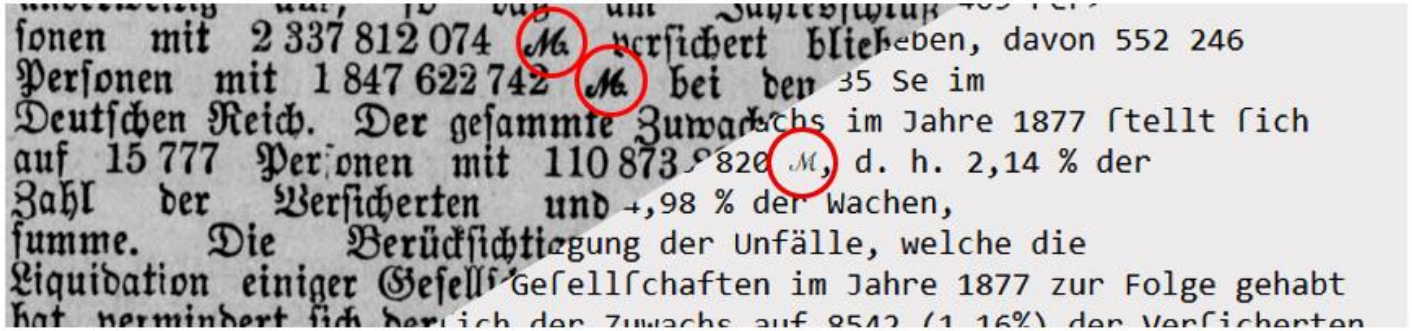
Der Weg zum nutzbaren Volltext. Werkspezifisches Training als Baustein der OCR-Volltexterkennung für Alte Drucke



Jan Kamlah und Thomas Schmidt

Der Weg zum nutzbaren Volltext. Werkspezifisches Training als Baustein der OCR-Volltexterkennung für Alte Drucke

UB
UNIVERSITÄTSBIBLIOTHEK
MANNHEIM

1. OCR und werkspezifisches Training an der UB Mannheim
2. Einführung
3. Training mit synthetischer Ground Truth (Ansatz A)
4. Training mit realer Ground Truth (Ansatz B)
5. Fazit

Der Weg zum nutzbaren Volltext. Werkspezifisches Training als Baustein der OCR-Volltexterkennung für Alte Drucke

UB
UNIVERSITÄTSBIBLIOTHEK
MANNHEIM

1. OCR und werkspezifisches Training an der UB Mannheim
2. Einführung
3. Training mit synthetischer Ground Truth (Ansatz A)
4. Training mit realer Ground Truth (Ansatz B)
5. Fazit

1. OCR und werkspezifisches Training an der UB Mannheim

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- 2022: werkspezifisches Training (eScriptorium + Kraken) mit theologischen Werken (DigiTheo Ground Truth: https://github.com/UB-Mannheim/DTGT)
- Seit 2021: Modulprojekt „Werkspezifisches Training“ (für Tesseract + Calamari) im Rahmen der 3. DFG-Förderphase von OCR-D
- 2013–2019: Digitalisierung und OCR „Aktienführer-Datenarchiv“
- 2019: Werkspezifisches Training für „Weisthümer“ (von Jacob Grimm, 1840) https://github.com/UB-Mannheim/Weisthuemer
- 2011–2018: Digitalisierung und OCR der Sammlung „Desbillons“ (Reiseberichte & juristische Texte)

OCR-D
Koordinierte Förderinitiative zur Weiterentwicklung von Verfahren der Optical Character Recognition (OCR)

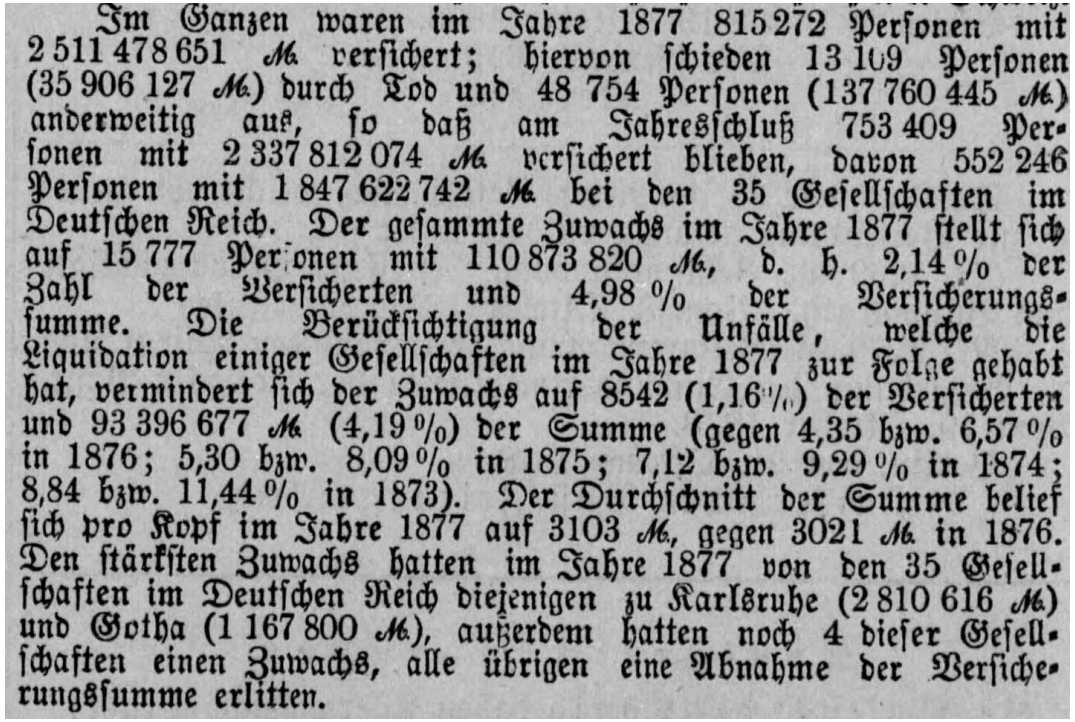
DFG

Der Weg zum nutzbaren Volltext. Werkspezifisches Training als Baustein der OCR-Volltexterkennung für Alte Drucke

UB
UNIVERSITÄTSBIBLIOTHEK
MANNHEIM

1. OCR und werkspezifisches Training an der UB Mannheim
2. Einführung
3. Training mit synthetischer Ground Truth (Ansatz A)
4. Training mit realer Ground Truth (Ansatz B)
5. Fazit

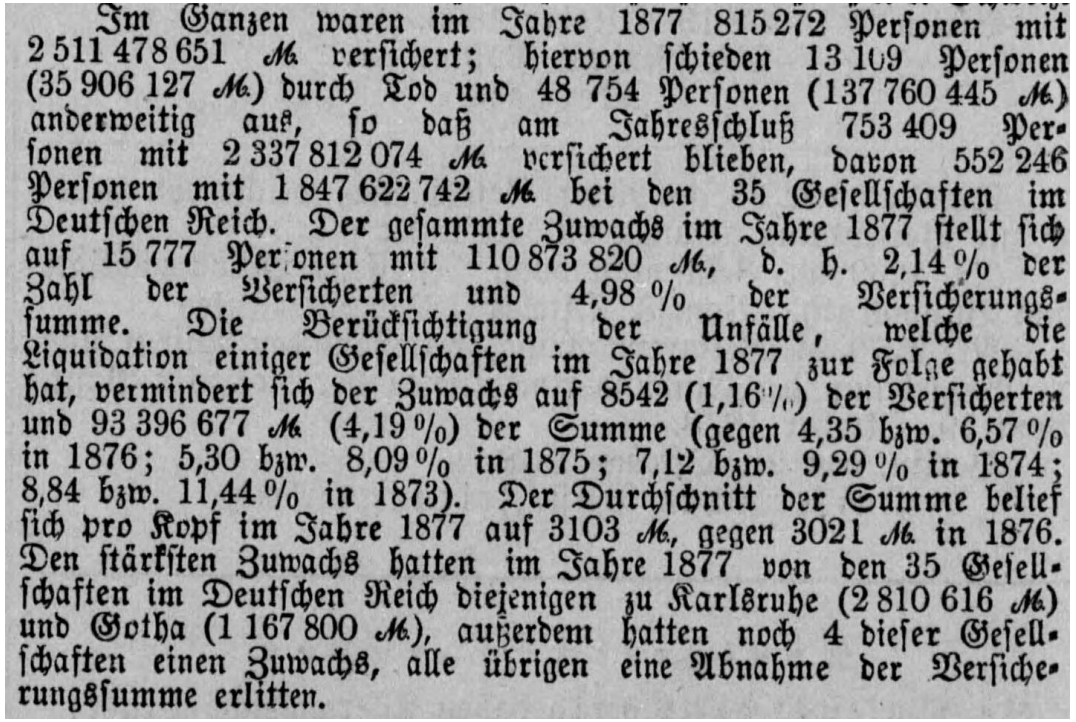
2. Einführung



Im Ganzen waren im Jahre 1877 815 272 Personen mit 2 511 478 651 ℳ versichert; hiervon schieden 13 109 Personen (35 906 127 ℳ) durch Tod und 48 754 Personen (137 760 445 ℳ) anderweitig aus, so daß am Jahresschluß 753 409 Personen mit 2 337 812 074 ℳ versichert blieben, davon 552 246 Personen mit 1 847 622 742 ℳ bei den 35 Gesellschaften im Deutschen Reich. Der gesammte Zuwachs im Jahre 1877 stellt sich auf 15 777 Personen mit 110 873 820 ℳ, d. h. 2,14 % der Zahl der Versicherten und 4,98 % der Versicherungssumme. Die Berücksichtigung der Unfälle, welche die Liquidation einiger Gesellschaften im Jahre 1877 zur Folge gehabt hat, vermindert sich der Zuwachs auf 8542 (1,16 %) der Versicherten und 93 396 677 ℳ (4,19 %) der Summe (gegen 4,35 bzw. 6,57 % in 1876; 5,30 bzw. 8,09 % in 1875; 7,12 bzw. 9,29 % in 1874; 8,84 bzw. 11,44 % in 1873). Der Durchschnitt der Summe belief sich pro Kopf im Jahre 1877 auf 3103 ℳ, gegen 3021 ℳ in 1876. Den stärksten Zuwachs hatten im Jahre 1877 von den 35 Gesellschaften im Deutschen Reich diejenigen zu Karlsruhe (2 810 616 ℳ) und Gotha (1 167 800 ℳ), außerdem hatten noch 4 dieser Gesellschaften einen Zuwachs, alle übrigen eine Abnahme der Versicherungssumme erlitten.

Deutscher Reichsanzeiger und Preußischer Staatsanzeiger, Jg. 1878, Ausgabe 248

2. Einführung



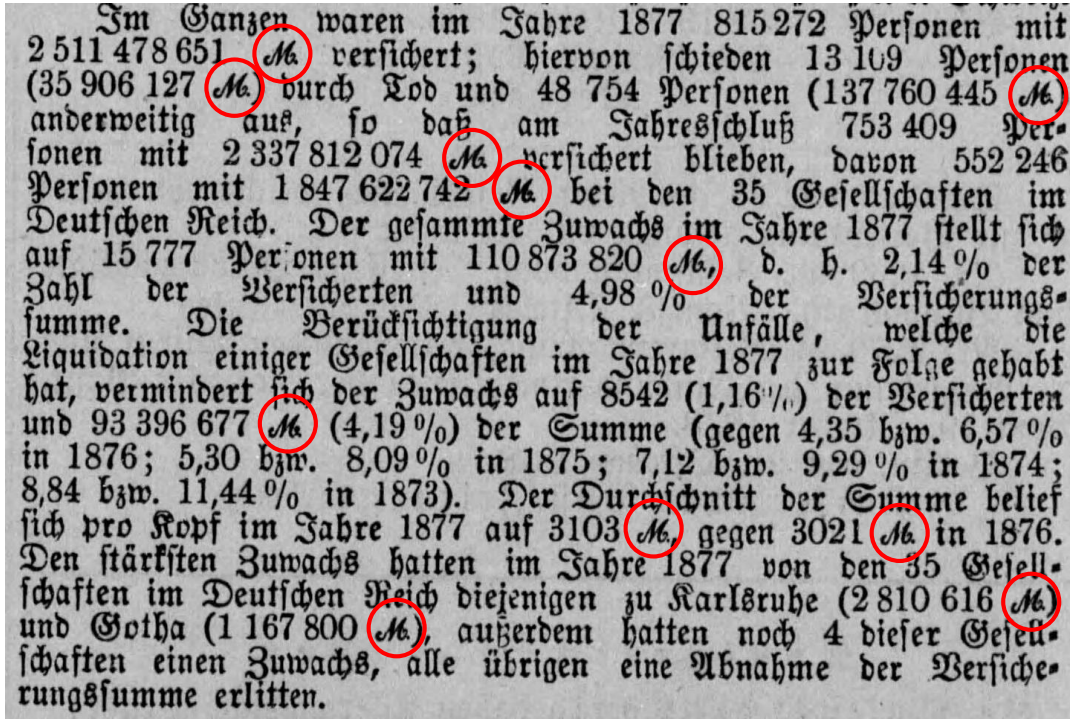
Im Ganzen waren im Jahre 1877 815 272 Personen mit 2 511 478 651 ℳ versichert; hiervon schieden 13 109 Personen (35 906 127 ℳ) durch Tod und 48 754 Personen (137 760 445 ℳ) anderweitig aus, so daß am Jahresschluß 753 409 Personen mit 2 337 812 074 ℳ versichert blieben, davon 552 246 Personen mit 1 847 622 742 ℳ bei den 35 Gesellschaften im Deutschen Reich. Der gesammte Zuwachs im Jahre 1877 stellt sich auf 15 777 Personen mit 110 873 820 ℳ, d. h. 2,14 % der Zahl der Versicherten und 4,98 % der Versicherungssumme. Die Berücksichtigung der Unfälle, welche die Liquidation einiger Gesellschaften im Jahre 1877 zur Folge gehabt hat, vermindert sich der Zuwachs auf 8542 (1,16 %) der Versicherten und 93 396 677 ℳ (4,19 %) der Summe (gegen 4,35 bzw. 6,57 % in 1876; 5,30 bzw. 8,09 % in 1875; 7,12 bzw. 9,29 % in 1874; 8,84 bzw. 11,44 % in 1873). Der Durchschnitt der Summe belief sich pro Kopf im Jahre 1877 auf 3103 ℳ, gegen 3021 ℳ in 1876. Den stärksten Zuwachs hatten im Jahre 1877 von den 35 Gesellschaften im Deutschen Reich diejenigen zu Karlsruhe (2 810 616 ℳ) und Gotha (1 167 800 ℳ), außerdem hatten noch 4 dieser Gesellschaften einen Zuwachs, alle übrigen eine Abnahme der Versicherungssumme erlitten.

Deutscher Reichsanzeiger und Preußischer Staatsanzeiger, Jg. 1878, Ausgabe 248

```
Im Ganzen waren im Jahre 1877 815272 Perſonen mit
2511478 651 -ÄZ erſichert; hiervon ſchieden 13 109 Perſonen
(35 906 127 ìì) durch Tod und 48 754 11 (137 760 445 κìì
anderweitig aus, ſo daß am Jahresſchluß 753 409 Per≈
ſonen mit 2337 812 074 &α verſichert blieben, davon 552 246
Perſonen mit 1847 622 742 bei den 35 re 157 fteh im
Deutſchen Reich. Der geſammte Zuwachs im Jahre 1877 ftellt ſich
auf 15 777 Perionen mit 110 873 820 , d. h. 2,14 %0 der
Zahl der Verſicherten und 4,98 % der Wunche die
ſumme. Die Berückſichtigung der Unfälle, welche die
Liquidation einiger Geſellſchaften im Jahre 1877 zur Folge gehabt
hat, vermindert ſich der Zuwachs auf 8542 (1,16¾) der Verſicherten
und 93 396 677 α (4,19%) der Summe (gegen 4,35 bzw. 6,57 /
in 1876; 5,30 bzw. 8,09 % in 1875; 7,12 bzw. 9,29 % in 1874;
8,84 bzw. 11,44% in 187, Der Durchſchnitt der Summe belief
ſich pro Kopf im Jahre 1877 auf 3103 α, gegen 302 σ in 1876.
Den ſtärkſten Zuwachs hatten im Jahre 1877 von den 35 Geſell≈
ſchaften im Deutſchen Reich diejenigen zu Karlsruhe (2810 616
und Gotha (1 167 800 ), außerdem hatten noch 4 dieſer Geſell≈
ſchaften einen Zuwachs, alle übrigen eine Abnahme der Verſiche≈
rungsſumme erlitten.
```

OCR-Ergebnis
Tesseract (tesseract 5.1.0-20)
Tesseract-Model: frak2021 (UB Mannheim)

2. Einführung



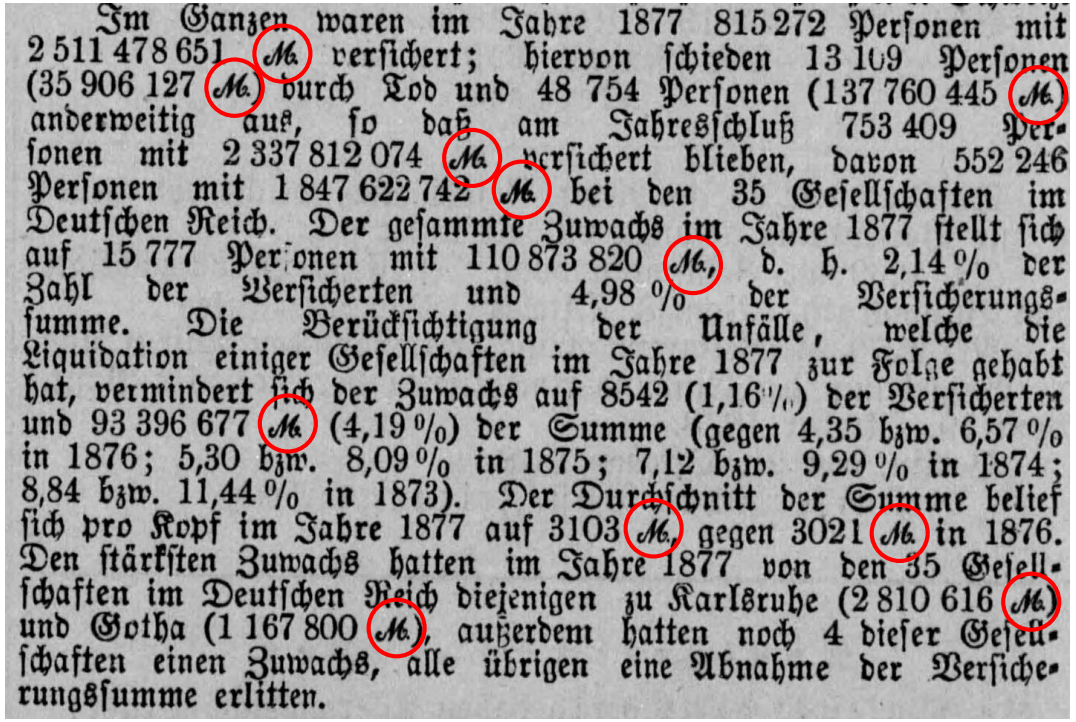
Im Ganzen waren im Jahre 1877 815272 Personen mit 2 511 478 651 ℳ versichert; hiervon schieden 13 109 Personen (35 906 127 ℳ) durch Tod und 48 754 Personen (137 760 445 ℳ) anderweitig aus, so daß am Jahresschluß 753 409 Personen mit 2 337 812 074 ℳ versichert blieben, davon 552 246 Personen mit 1 847 622 742 ℳ bei den 35 Gesellschaften im Deutschen Reich. Der gesammte Zuwachs im Jahre 1877 stellt sich auf 15 777 Personen mit 110 873 820 ℳ, d. h. 2,14 % der Zahl der Versicherten und 4,98 % der Versicherungssumme. Die Berücksichtigung der Unfälle, welche die Liquidation einiger Gesellschaften im Jahre 1877 zur Folge gehabt hat, vermindert sich der Zuwachs auf 8542 (1,16 %) der Versicherten und 93 396 677 ℳ (4,19 %) der Summe (gegen 4,35 bzw. 6,57 % in 1876; 5,30 bzw. 8,09 % in 1875; 7,12 bzw. 9,29 % in 1874; 8,84 bzw. 11,44 % in 1873). Der Durchschnitt der Summe belief sich pro Kopf im Jahre 1877 auf 3103 ℳ, gegen 3021 ℳ in 1876. Den stärksten Zuwachs hatten im Jahre 1877 von den 35 Gesellschaften im Deutschen Reich diejenigen zu Karlsruhe (2 810 616 ℳ) und Gotha (1 167 800 ℳ), außerdem hatten noch 4 dieser Gesellschaften einen Zuwachs, alle übrigen eine Abnahme der Versicherungssumme erlitten.

Deutscher Reichsanzeiger und Preußischer Staatsanzeiger, Jg. 1878, Ausgabe 248

```
Im Ganzen waren im Jahre 1877 815272 Perſonen mit
2511478 651 -ÄZ erſichert; hiervon ſchieden 13 109 Perſonen
(35 906 127 ii) durch Tod und 48 754 11 (137 760 445 kìì
anderweitig aus, ſo daß am Jahresſchluß 753 409 Per=
ſonen mit 2337 812 074 &a verſichert blieben, davon 552 246
Perſonen mit 1847 622 742 bei den 35 re 157 fteh im
Deutſchen Reich. Der geſammte Zuwachs im Jahre 1877 ftellt ſich
auf 15 777 Perionen mit 110 873 820 , d. h. 2,14 %0 der
Zahl der Verſicherten und 4,98 % der Wunche die
ſumme. Die Berückſichtigung der Unfälle, welche die
Liquidation einiger Geſellſchaften im Jahre 1877 zur Folge gehabt
hat, vermindert ſich der Zuwachs auf 8542 (1,16%) der Verſicherten
und 93 396 677 α (4,19%) der Summe (gegen 4,35 bzw. 6,57 /
in 1876; 5,30 bzw. 8,09 % in 1875; 7,12 bzw. 9,29 % in 1874;
8,84 bzw. 11,44% in 187, Der Durchſchnitt der Summe belief
ſich pro Kopf im Jahre 1877 auf 3103 α, gegen 302 σ in 1876.
Den ſtärkſten Zuwachs hatten im Jahre 1877 von den 35 Geſell=
ſchaften im Deutſchen Reich diejenigen zu Karlsruhe (2810 616
und Gotha (1 167 800 ), außerdem hatten noch 4 dieſer Geſell=
ſchaften einen Zuwachs, alle übrigen eine Abnahme der Verſiche=
rungsſumme erlitten.
```

OCR-Ergebnis
Tesseract (tesseract 5.1.0-20)
Tesseract-Model: frak2021 (UB Mannheim)

2. Einführung



Im Ganzen waren im Jahre 1877 815 272 Personen mit 2 511 478 651 ℳ versichert; hiervon schieden 13 109 Personen (35 906 127 ℳ) durch Tod und 48 754 Personen (137 760 445 ℳ) anderweitig aus, so daß am Jahresschluß 753 409 Personen mit 2 337 812 074 ℳ versichert blieben, davon 552 246 Personen mit 1 847 622 742 ℳ bei den 35 Gesellschaften im Deutschen Reich. Der gesammte Zuwachs im Jahre 1877 stellt sich auf 15 777 Personen mit 110 873 820 ℳ, d. h. 2,14 % der Zahl der Versicherten und 4,98 % der Versicherungssumme. Die Berücksichtigung der Unfälle, welche die Liquidation einiger Gesellschaften im Jahre 1877 zur Folge gehabt hat, vermindert sich der Zuwachs auf 8542 (1,16 %) der Versicherten und 93 396 677 ℳ (4,19 %) der Summe (gegen 4,35 bzw. 6,57 % in 1876; 5,30 bzw. 8,09 % in 1875; 7,12 bzw. 9,29 % in 1874; 8,84 bzw. 11,44 % in 1873). Der Durchschnitt der Summe belief sich pro Kopf im Jahre 1877 auf 3103 ℳ, gegen 3021 ℳ in 1876. Den stärksten Zuwachs hatten im Jahre 1877 von den 35 Gesellschaften im Deutschen Reich diejenigen zu Karlsruhe (2 810 616 ℳ) und Gotha (1 167 800 ℳ), außerdem hatten noch 4 dieser Gesellschaften einen Zuwachs, alle übrigen eine Abnahme der Versicherungssumme erlitten.

Deutscher Reichsanzeiger und Preußischer Staatsanzeiger, Jg. 1878, Ausgabe 248

```
Im Ganzen waren im Jahre 1877 815272 Perſonen mit
2511 478 651 ℳ verſichert; hiervon ſchieden 13 109 Perſonen
(35 906 127 ℳ) durch Tod und 48 754 14. (137 760 445 ℳ)
anderweitig aus, ſo daß am Jahresſchluß 753 409 Per≈
ſonen mit 2 337 812 074 ℳ verſichert blieben, davon 552 246
Perſonen mit 1 847 622 742 ℳ bei den 35 Se im
Deutſchen Reich. Der geſammte Zuwachs im Jahre 1877 ſtellt ſich
auf 15777 Per onen mit 110 873 820 ℳ, d. h. 2,14 % der
Zahl der Verſicherten und 4,98 % der Wachen,
ſumme. Die Berückſichtigung der Unfälle, welche die
Liquidation einiger Geſellſchaften im Jahre 1877 zur Folge gehabt
hat, vermindert ſich der Zuwachs auf 8542 (1,16%) der Verſicherten
und 93 396 677 ℳ (4,19 %) der Summe (gegen 4,35 bzw. 6,57 0%
in 1876; 5,30 bzw. 8,09% in 1875; 7,12 bzw. 9,29 in 1874;
8,84 bzw. 11,44 % in 71. Der Durchſchnitt der Summe belief
ſich pro Kopf im Jahre 1877 auf 3103 ℳ., gegen 3021 ℳ in 1876.
Den ſtärkſten Zuwachs hatten im Jahre 1877 von den 35 Geſell≈
ſchaften im Deutſchen Reich diejenigen zu Karleruhe (2 810 616 ℳ)
und Gotha (1 167 800 ℳ), außerdem hatten noch 4 dieſer Geſell≈
ſchaften einen Zuwachs, alle übrigen eine Abnahme der Verſiche≈
rungsſumme erlitten.
```

OCR-Ergebnis
Tesseract (tesseract 5.1.0-20)
Tesseract-Model: frak2021 (UB Mannheim) nach werkspezifischem Training

2. Einführung

Gründe für ein werkspezifisches Training …

2. Einführung

Gründe für ein werkspezifisches Training …

- Verbesserung der Texterkennungsqualität

2. Einführung

Gründe für ein werkspezifisches Training …

- Verbesserung der Texterkennungsqualität
- Erweiterung des zu erkennenden Zeichenvorrates → historische Währungssymbole, astronomische oder mathematische Zeichen etc.

2. Einführung

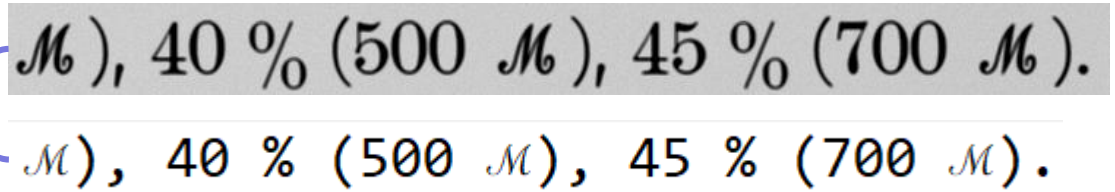
Gründe für ein werkspezifisches Training …

- Verbesserung der Texterkennungsqualität
- Erweiterung des zu erkennenden Zeichenvorrates → historische Währungssymbole, astronomische oder mathematische Zeichen etc.
- Spezielle Mappings → Normalisierungen, Auflösung von Kürzeln, Umgang mit Leerräumen (gesperrte Schrift) etc.

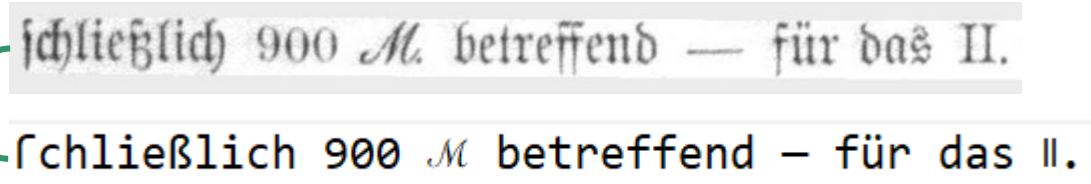
2. Einführung

Zwei getestete Ansätze für werkspezifisches Training:

Werkspezifisches Training mit **synthetischer** Ground Truth



Werkspezifisches Training mit **realer** Ground Truth

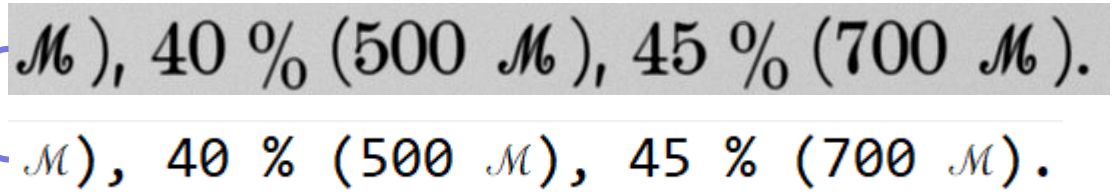


2. Einführung

Zwei getestete Ansätze für werkspezifisches Training:

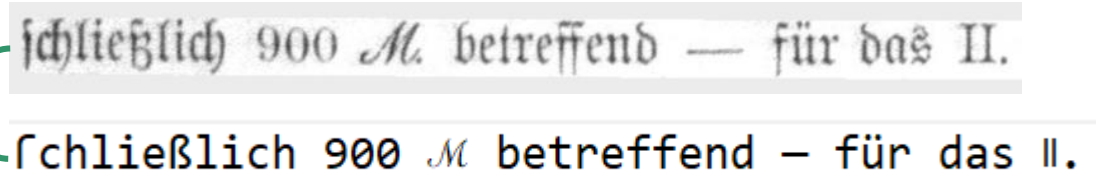
Durch Font-Rendering künstlich erzeugte GT

→ Werkspezifisches Training mit **synthetischer** Ground Truth



Auf realem Digitalisat (bspw. Buchseite) basierende GT

→ Werkspezifisches Training mit **realer** Ground Truth



Der Weg zum nutzbaren Volltext. Werkspezifisches Training als Baustein der OCR-Volltexterkennung für Alte Drucke

UB
UNIVERSITÄTSBIBLIOTHEK
MANNHEIM

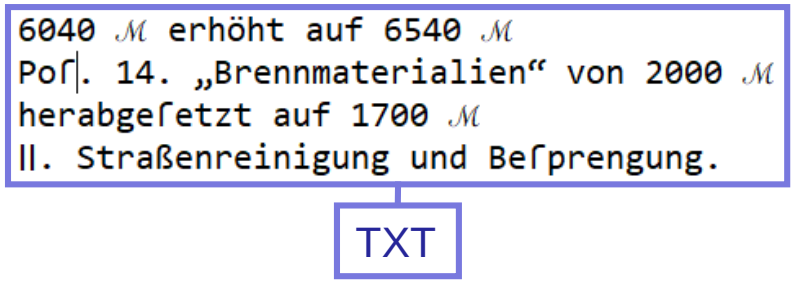
1. OCR und werkspezifisches Training an der UB Mannheim
2. Einführung
3. Training mit synthetischer Ground Truth (Ansatz A)
4. Training mit realer Ground Truth (Ansatz B)
5. Fazit

3. Training mit synthetischer Ground Truth (Ansatz A)

- Tesseract-Tool **Text2Image** erzeugt aus Textinput und Font eine Bilddatei

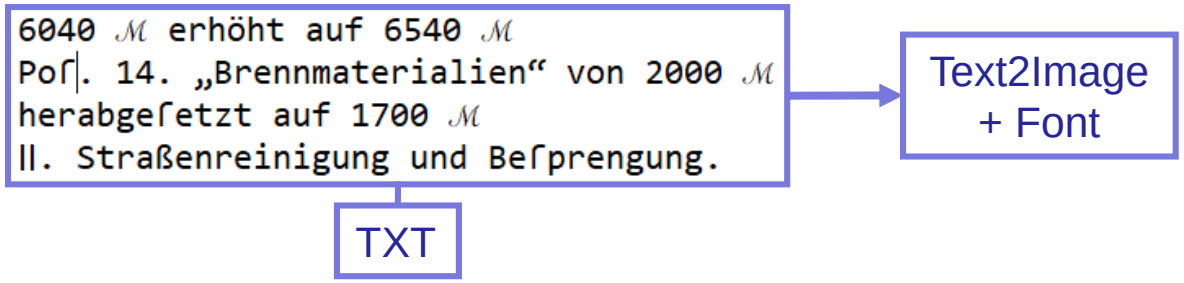
3. Training mit synthetischer Ground Truth (Ansatz A)

- Tesseract-Tool **Text2Image** erzeugt aus Textinput und Font eine Bilddatei



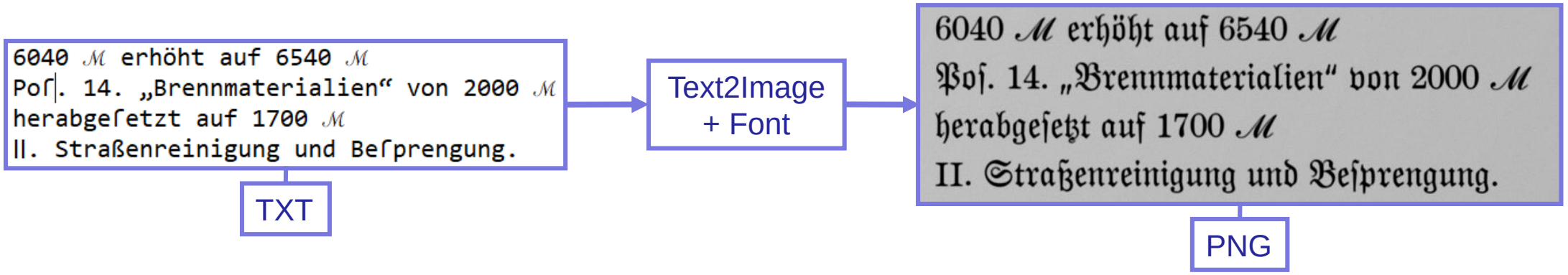
3. Training mit synthetischer Ground Truth (Ansatz A)

- Tesseract-Tool **Text2Image** erzeugt aus Textinput und Font eine Bilddatei



3. Training mit synthetischer Ground Truth (Ansatz A)

- Tesseract-Tool **Text2Image** erzeugt aus Textinput und Font eine Bilddatei



3. Training mit synthetischer Ground Truth (Ansatz A)

2 verwendete Fraktur-Fonts für Rendering:

Gründerkrach-Fraktur	Berthold Mainzer Fraktur
entspricht dem verwendeten Font des zu erkennenden Ausgangsmaterials	**ähnelt** dem Font des zu erkennenden Ausgangsmaterial
hoher Zeichenvorrat: langes ſ, rundes ꝛ, $\mathcal{M}$ etc.	hoher Zeichenvorrat: langes ſ, rundes ꝛ, $\mathcal{M}$ etc.

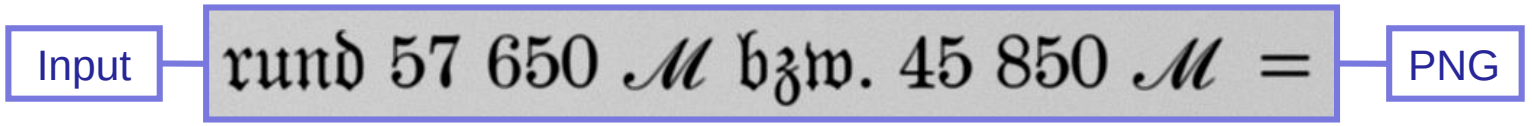
Tabelle 1: Verwendete Fonts

3. Training mit synthetischer Ground Truth (Ansatz A)

- Zusätzliches Skript, um Bildzeilen per data augmentation zu verändern und die GT mit Variationen anzureichern

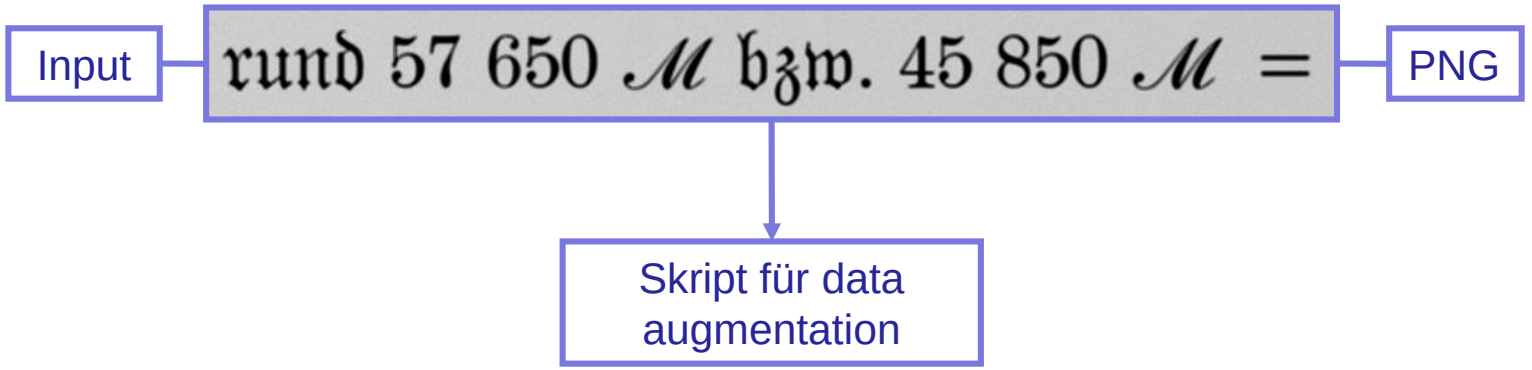
3. Training mit synthetischer Ground Truth (Ansatz A)

- Zusätzliches Skript, um Bildzeilen per data augmentation zu verändern und die GT mit Variationen anzureichern



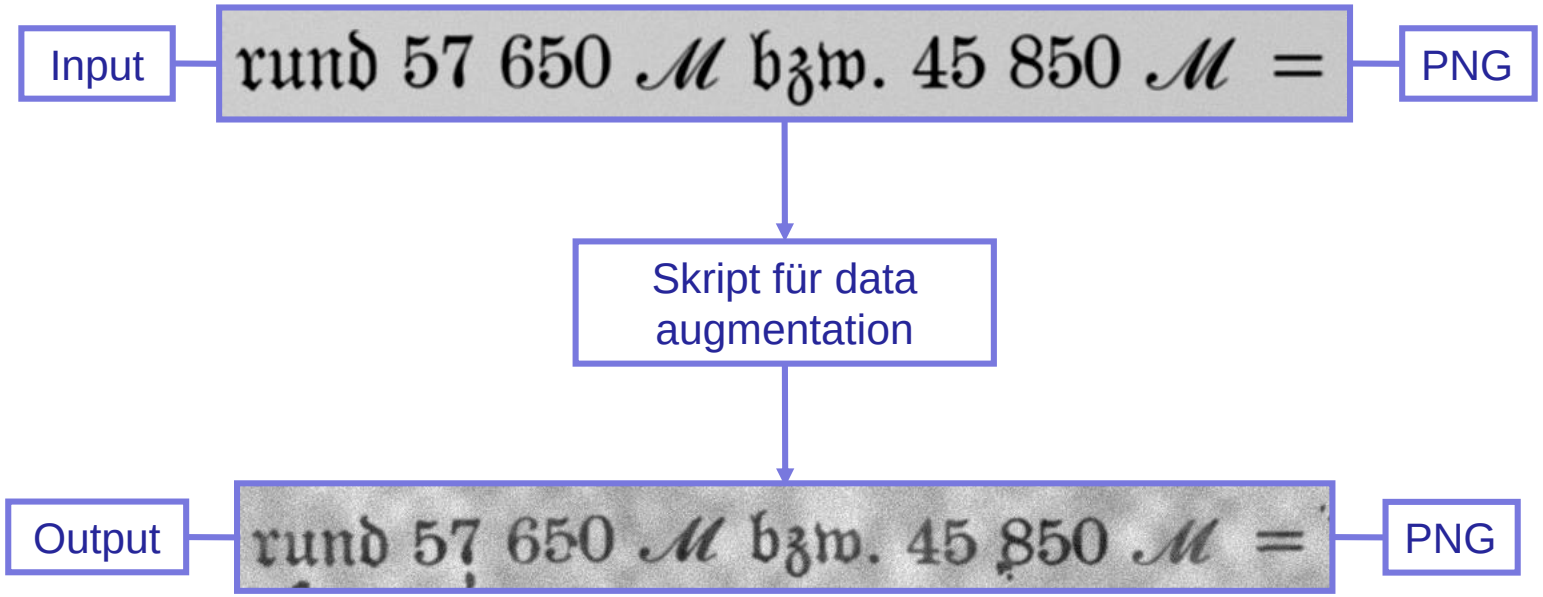
3. Training mit synthetischer Ground Truth (Ansatz A)

- Zusätzliches Skript, um Bildzeilen per data augmentation zu verändern und die GT mit Variationen anzureichern



3. Training mit synthetischer Ground Truth (Ansatz A)

- Zusätzliches Skript, um Bildzeilen per data augmentation zu verändern und die GT mit Variationen anzureichern



3. Training mit synthetischer Ground Truth (Ansatz A)

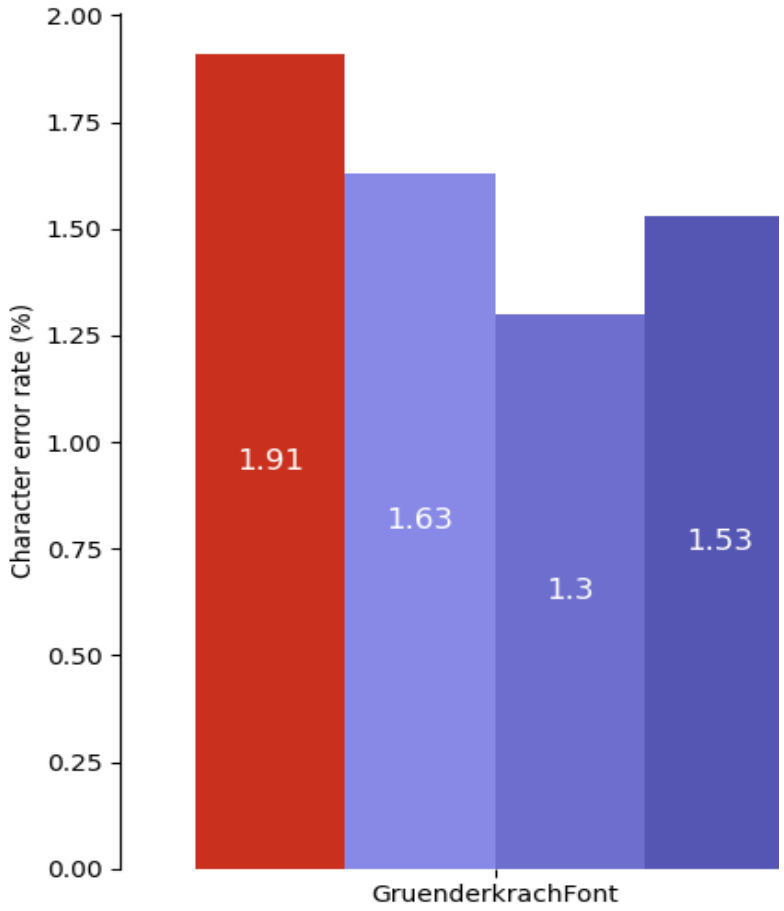
Tabelle 2: Trainingssets mit synthetischer Ground Truth

Werkspezifisches Training mit synthetischer Ground Truth (2 Fonts / 6 Datasets)			
Ausgangsmodell:	frak2021 (UB Mannheim)		
Textquelle:	„Charlottenburger Amtsschrifttum", 1879–1919 (ZLB)		
Fonts:	Gründerfrach-Fraktur © 2021 Philipp Poll	Berthold Mainzer Fraktur © Peter Wiegel	
Synthetische GT:	**LARGE*** 3.000 GT lines	**MEDIUM†** 1.000 GT lines	**SMALL†** 100 GT lines
	PNG, Graustufe, 50 % per data augmentation bearbeitet		
Trainingsziel:	Erweiterung des Zeichenvorrats um: $\mathcal{M}$		
Trainingslaufzeit:	60 Epochen mit Abbruchkriterium → Tatsächliche Laufzeit ca. 40–50 Epochen		

* enthält Lines mit und ohne $\mathcal{M}$; † enthält nur Lines mit $\mathcal{M}$

3. Training mit synthetischer Ground Truth (Ansatz A)

Evaluation: Gründerkrach-Fraktur



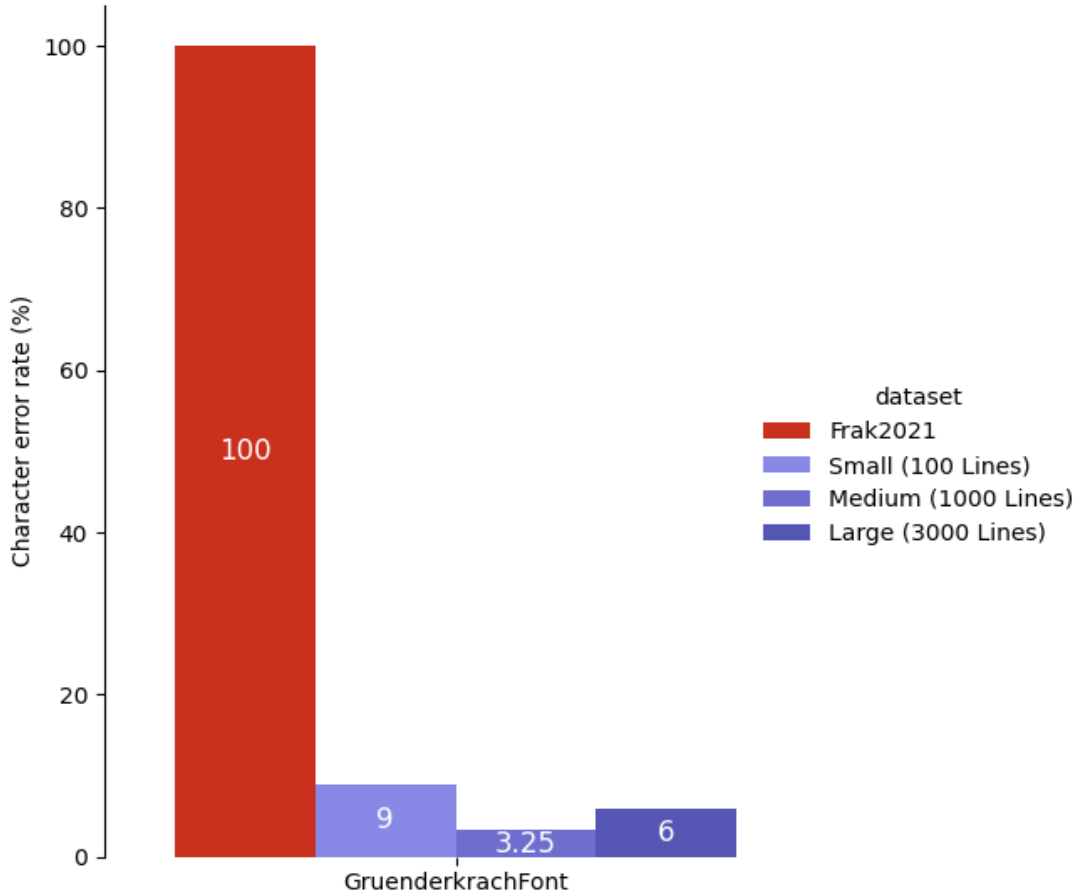


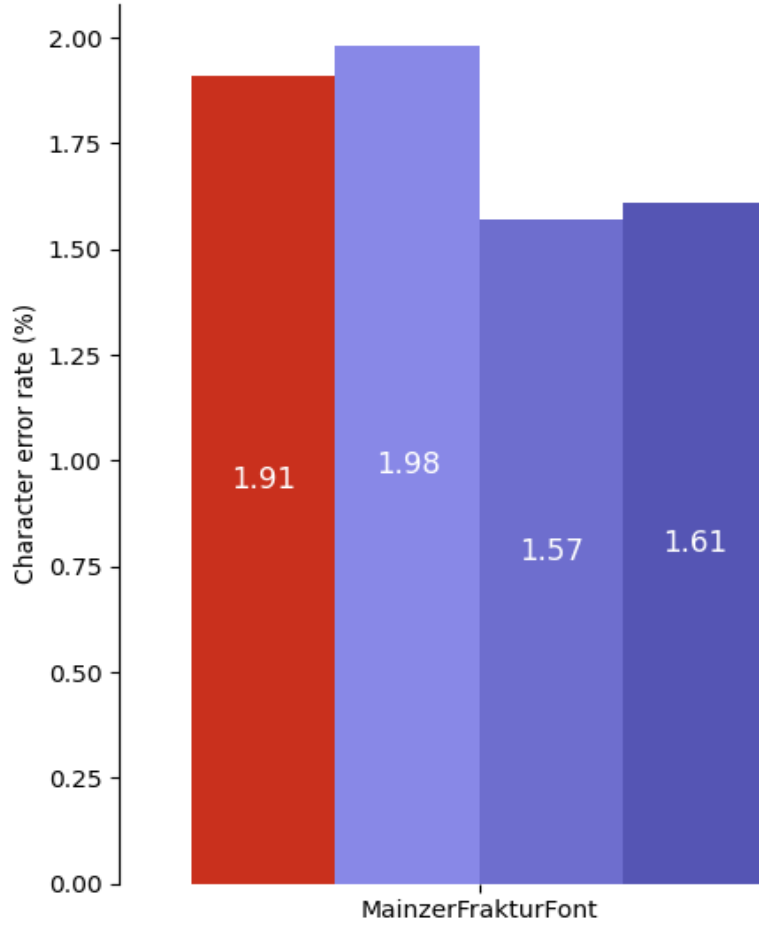
Fig. 1

Allgemeine CER

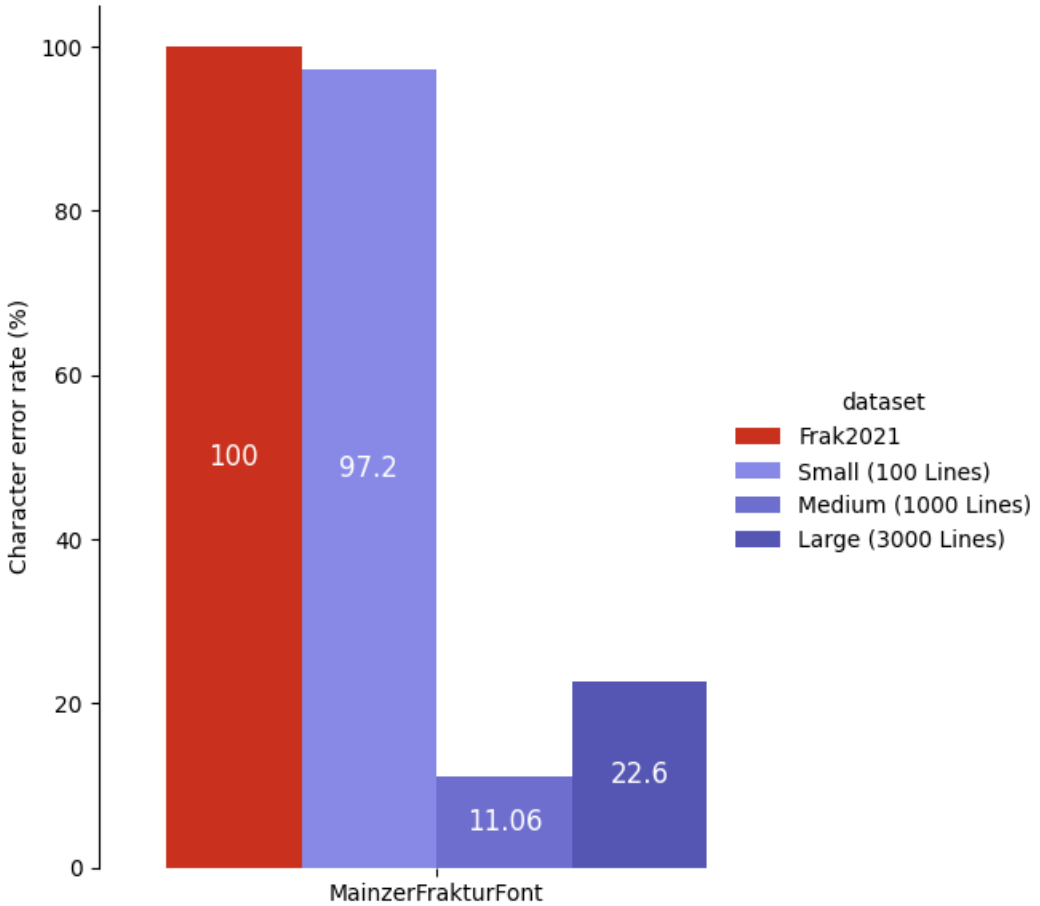
CER Marksymbole ($\mathcal{M}$)

3. Training mit synthetischer Ground Truth (Ansatz A)

Evaluation: Berthold Mainzer Fraktur



Allgemeine CER



CER Marksymbole ($\mathcal{M}$)

Fig. 2

Der Weg zum nutzbaren Volltext. Werkspezifisches Training als Baustein der OCR-Volltexterkennung für Alte Drucke

UB
UNIVERSITÄTSBIBLIOTHEK
MANNHEIM

1. OCR und werkspezifisches Training an der UB Mannheim
2. Einführung
3. Training mit synthetischer Ground Truth (Ansatz A)
4. Training mit realer Ground Truth (Ansatz B)
5. Fazit

4. Training mit realer Ground Truth (Ansatz B)

- Analog zum **synthetischen** Training wurde als Quelle der GT-Produktion das „Charlottenburger Amtsschrifttum”, 1879–1919 (Zentrale Landesbibliothek Berlin) verwendet.

4. Training mit realer Ground Truth (Ansatz B)

- Analog zum **synthetischen** Training wurde als Quelle der GT-Produktion das „Charlottenburger Amtsschrifttum”, 1879–1919 (Zentrale Landesbibliothek Berlin) verwendet.

- Die Transkription des Quellmaterials erfolgte in **Transkribus**.

4. Training mit realer Ground Truth (Ansatz B)

- Analog zum **synthetischen** Training wurde als Quelle der GT-Produktion das „Charlottenburger Amtsschrifttum”, 1879–1919 (Zentrale Landesbibliothek Berlin) verwendet.
- Die Transkription des Quellmaterials erfolgte in **Transkribus**.
- Im Gegensatz zum **synthetischen** Training erfolgte als data augmentation nur eine Graustufennormalisierung für die Bildinhalte.

4. Training mit realer Ground Truth (Ansatz B)

- Analog zum **synthetischen** Training wurde als Quelle der GT-Produktion das „Charlottenburger Amtsschrifttum”, 1879–1919 (Zentrale Landesbibliothek Berlin) verwendet.
- Die Transkription des Quellmaterials erfolgte in **Transkribus**.
- Im Gegensatz zum **synthetischen** Training erfolgte als data augmentation nur eine Graustufennormalisierung für die Bildinhalte.
- Sowohl Umfang der Ground Truth als auch Trainingslaufzeit (in Epochen) waren mit dem synthetischen Training identisch.

4. Training mit realer Ground Truth (Ansatz B)

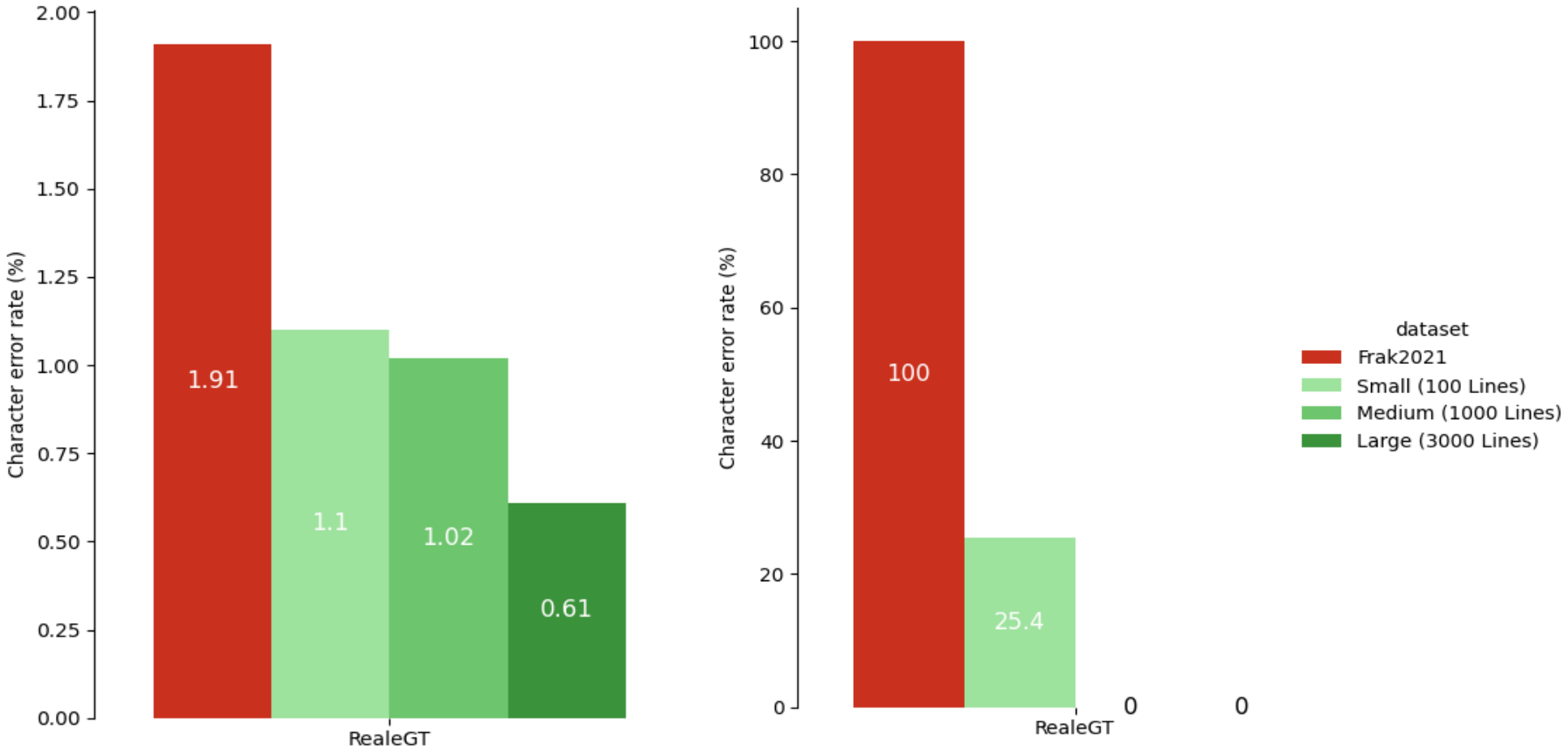
Tabelle 3: Trainingsaufbau reale Ground Truth

Werkspezifisches Training mit realer Ground Truth (3 Datasets)			
Ausgangsmodell:	frak2021 (UB Mannheim)		
Quelle:	„Charlottenburger Amtsschrifttum", 1881–1906 (ZLB)		
Ground Truth:	**LARGE*** 3.000 GT lines	**MEDIUM†** 1.000 GT lines	**SMALL†** 100 GT lines
	PNG, Graustufe (+ Graustufennormalisierung)		
Trainingsziel:	Erweiterung des Zeichenvorrats um: $\mathcal{M}$		
Trainingslaufzeit:	60 Epochen mit Abbruchkriterium → Tatsächliche Laufzeit ca. 40–50 Epochen		

* enthält Lines mit und ohne $\mathcal{M}$; † enthält nur Lines mit $\mathcal{M}$

4. Training mit realer Ground Truth (Ansatz B)

Evaluation: reale Ground Truth



Allgemeine CER

CER Marksymbol ($\mathcal{M}$)

Fig. 3

Der Weg zum nutzbaren Volltext. Werkspezifisches Training als Baustein der OCR-Volltexterkennung für Alte Drucke

UB
UNIVERSITÄTSBIBLIOTHEK
MANNHEIM

1. OCR und werkspezifisches Training an der UB Mannheim
2. Einführung werkspezifisches Training
3. Training mit synthetischer Ground Truth (Ansatz A)
4. Training mit realer Ground Truth (Ansatz B)
5. Fazit

5. Fazit

Allgemeine CER von realer und synthetischer GT

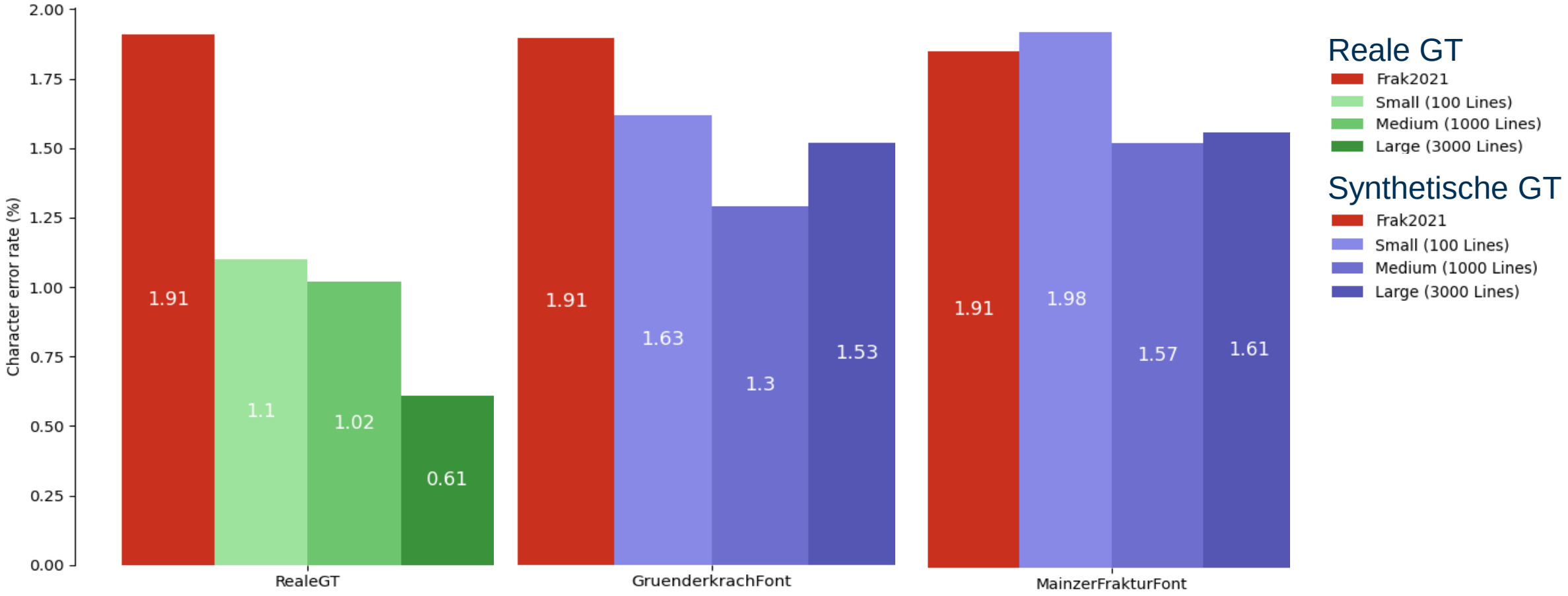


Fig. 4

5. Fazit

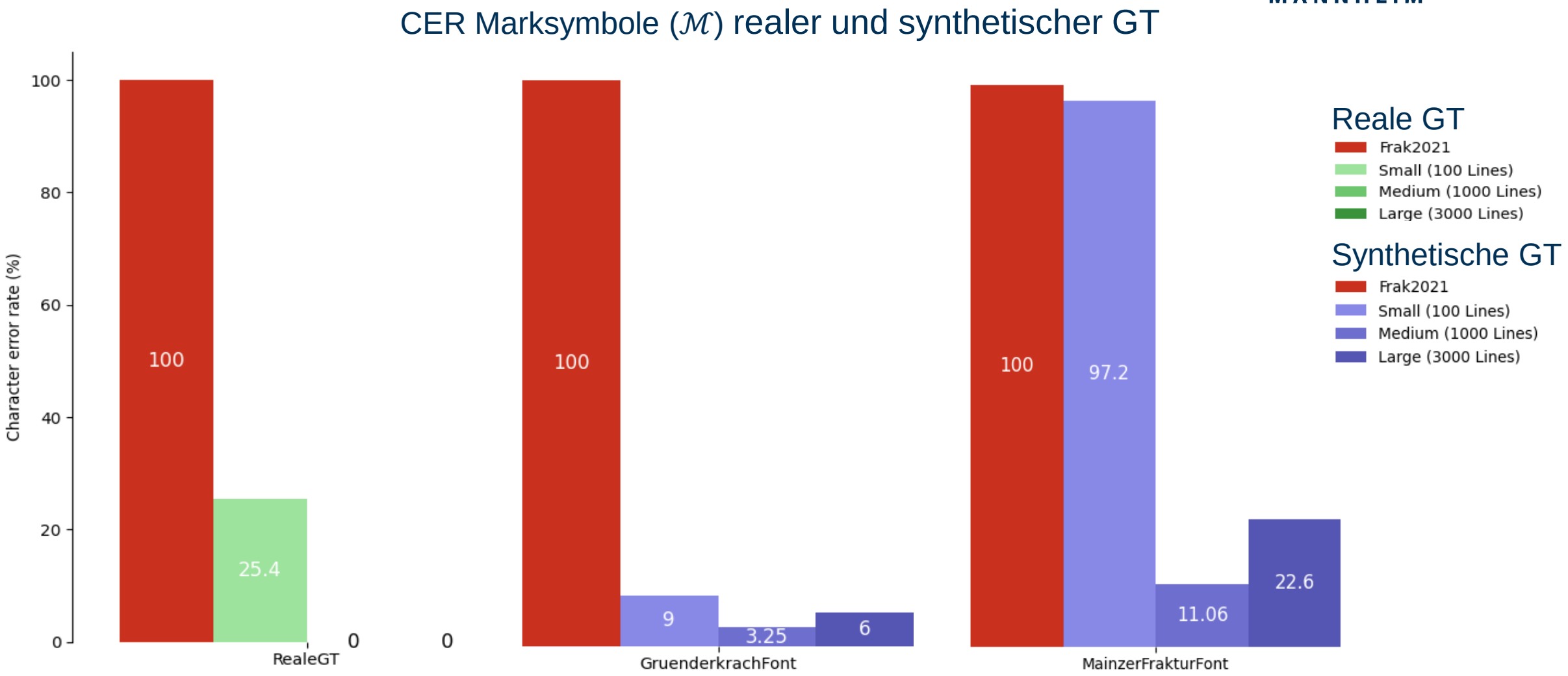

CER Marksymbole ($\mathcal{M}$) realer und synthetischer GT
Character error rate (%)
100
80
60
40
20
0
100
25.4
0
0
100
9
3.25
6
100
97.2
11.06
22.6
RealeGT
GruenderkrachFont
MainzerFrakturFont
Reale GT
Frak2021
Small (100 Lines)
Medium (1000 Lines)
Large (3000 Lines)
Synthetische GT
Frak2021
Small (100 Lines)
Medium (1000 Lines)
Large (3000 Lines)

Fig. 5

5. Fazit

Training mit synthetischer Ground Truth		Training mit realer Ground Truth	
Vorteile	**Nachteile**	**Vorteile**	**Nachteile**
Geringer zeitlicher Aufwand zur Erzeugung umfangreicher GT	Fonts, die Ausgangsmaterial entsprechen, sind selten verfügbar	Sehr gute Verbesserung des Ausgangsmodells	Relativ hoher Aufwand zur Erstellung der GT (Auswahl entsprechender Textstellen mit infrage kommenden Glyphen + Transkription + QA)
Gute Ergebnisse, wenn der genutzte Font dem Ausgangsmaterial entspricht	Zeichensatz verfügbarer Fonts meist unzureichend (besonders bei seltenen Glyphen)	Nachtrainiertes Modell ist Modellen mit synthetischer GT überlegen	
	Seltene Glyphen in verschiedenen Fonts uneinheitlich codiert		

Der Weg zum nutzbaren Volltext. Werkspezifisches Training als Baustein der OCR-Volltexterkennung für Alte Drucke

UB
UNIVERSITÄTSBIBLIOTHEK
MANNHEIM

Ground Truth, weitere Auswertungen und zusätzliche Informationen auf Github:
https://github.com/UB-Mannheim/charlottenburger-amtsschrifttum

Vielen Dank!

Jan Kamlah (Entwicklung): jan.kamlah@bib.uni-mannheim.de
Thomas Schmidt (Projektmanagement): thomas.schmidt@bib.uni-mannheim.de

5. Fazit

- Werkspezifisches Training ist ein effektives Verfahren zur Optimierung der Texterkennungsausgabe
- In der Regel sollte das Training mit realen GT-Daten bevorzugt werden
- Training mit geringem Material starten und ggf. erweitern
- Data augmentation regularisiert das Training und reduziert die Gefahren des Overfittings
- Min. 40 Epochen trainieren
  - besser wären 100 Epochen mit einem Abbruchkriterium wie etwa 0.01 % CER
  - Nachträgliche Evaluation der Modelle um die Verwendung von Overfitting-Modelle zu vermeiden
- Kontrolle des Zeichenvorrates des Trainingsmaterials

Literatur

- „Charlottenburger Amtsschrifttum”, 1879–1919 (ZLB), https://digital.zlb.de/viewer/metadata/16046633/1/
- BertholdMainzerFraktur, Peter Wiegel, http://www.peter-wiegel.de/MainzerFraktur.html
- Weil, Stefan: Neue Frakturmodelle für Tesseract, 2019, https://madoc.bib.uni-mannheim.de/53748
- Weil Stefan: 126 Jahre Zeitung online - Fundgrube für historisch Interessierte und Motor für die Bibliotheks-IT : 126 years of the newspaper online, 2018, https://madoc.bib.uni-mannheim.de/46507/
- Weil, Stefan und Kamlah, Jan: Forschungsdaten aus Digitalisaten, 2019, https://madoc.bib.uni-mannheim.de/52204