Jan Kamlah und Thomas Schmidt

Transkriptionsregeln und Guidelines zur Layoutbearbeitung

im DFG-Projekt "Workflow für werkspezifisches Training auf Basis generischer Modelle mit OCR-D sowie Ground-Truth-Aufwertung"

18.04.2023

1. Transkriptionsregeln

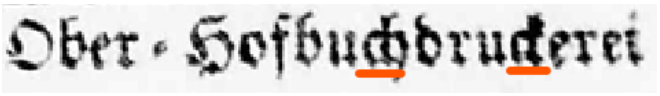
Die Grundlage zur Transkription bilden die OCR-D Ground Truth Richtlinen auf Level 2 (https://ocr-d.de/de/gt-guidelines/trans/tr_level_2_4.html). Abweichungen werden im Folgenden hervorgehoben.

1. Alle Zeichen, die in der Bildvorlage existieren, werden vorlagengetreu transkribiert. Aus diesem Grund verlangt die Transkription eine aufmerksame und konzentrierte Prüfung jedes einzelnen Buchstabens. Bei Unklarheiten bitte Rücksprache halten; spez. Entscheidungen werden danach hier dokumentiert.

2. **Konsonantische Ligaturen** (drucktechnische Verbindung bzw. Verschmelzung von 2 Konsonanten) werden in Einzelkonsonanten (Einzelbuchstaben) aufgespalten:

   a. 

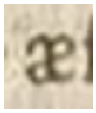
   Thätigkeit geſetzt,

   b. 

   Ober=Hofbuchdruckerei

c. 

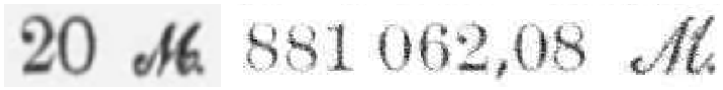
Die Fraktur ist eine gebrochene Schrift.

3. **Vokalische Ligaturen** werden wie in der Vorlage übernommen:
   a. 

   æ

   b. 

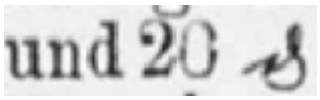
   Æ

4. **Währungssymbole** werden durch eigene Sonderzeichen dargestellt und werden gemäß der Vorlage übernommen. Sonderzeichen können in Transkribus im Reiter “custom” der “Virtuellen Tastatur” hinterlegt werden.
   a. 

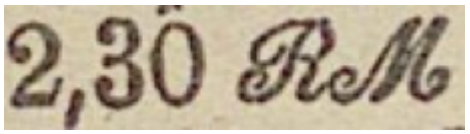
   20 ℳ 881 062,08 ℳ

   b. 

   340 000 ℳ.

   c. 

   und 20 ₰ (= Pfennig)

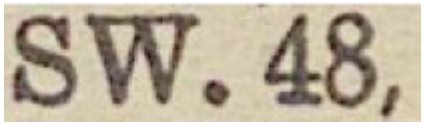
   d. 

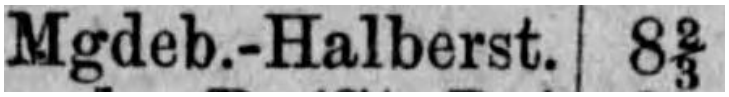
   2,30 ℛℳ

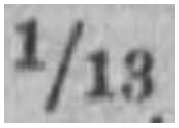
   e. 

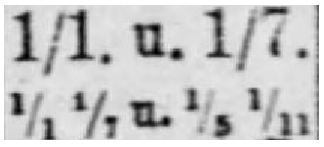
   Crt. (U+F2F1, https://mufi.info/m.php?p=mufi charinfo&i=4944)

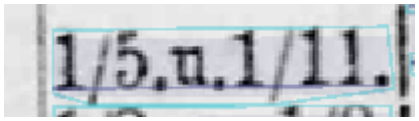
5. **Besondere Abkürzungen**
   a. 
   nen ⬚c. (“etc.”)

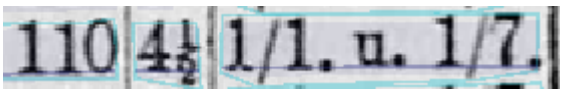
6. Auf Satzzeichen wie Punkt, Komma, Fragezeichen etc. folgt ein **Leerzeichen** - selbst dann, wenn dieses in der Vorlage fehlt:
   a. 
   SW. 48, statt SW.48,

7. Vor Symbolen wie **Brüchen, Prozentzeichen%, Währungssymbolen** etc. steht ein Leerzeichen:
   a. 
   Mgdeb.-Halberst. 8 ⅔

8. **Brüche** werden vorlagengetreu transkribiert; Brüche, die nicht durch ein Unicode-Sonderzeichen dargestellt werden können, werden ausgeschrieben. Wird der Bruch durch einen hochgestellten Nenner und tiefgestellten Zähler dargestellt, findet **nicht** der normale Slash (/) Verwendung, sondern der dezimale Bruchstrich (⁄).
   a. 
   1⁄13 (Dezimal-Bruchstrich, da hochgestellter Nenner und tiefgestellter Zähler)

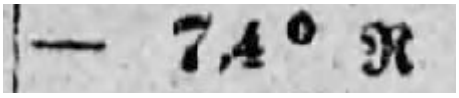
   b. 
   1/1. u. 1/7.
   1/1 ⬚ u. ⅐ 1/11

   c. 
   1/5. u. 1/11.

d. 

110 4 ½ 1/1. u. 1/7.

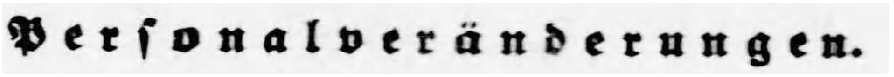
9. Das **Gradzeichen** (°) wird
   a. bei folgendem Einheitenzeichen mit diesem verbunden und durch ein Leerzeichen von der Zahl getrennt:



– 7,4 °R

10. Wörter und Sätze in **Großbuchstaben** werden wie dargestellt erfasst:
    a. 

CAMERA DI COMMERCIO ED ARTI

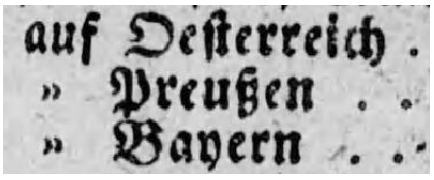
11. **Gesperrte Wörter** und Sätze werden ohne zusätzliche Trenner erfasst:
    a. 

Perſonalveränderungen.

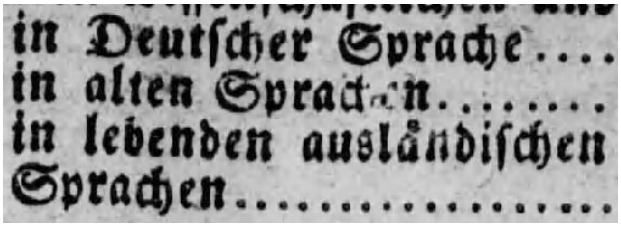
12. **Größere Leerräume** zwischen Wörtern gelten als ein Leerzeichen:
    a. 

(L. S.) Dr. Teßmann.

13. Als **Unterführungszeichen** wird ein dt. Anführungszeichen („) verwendet:



auf Oeſterreich
„ Preußen
„ Bayern

14. **Mehrere hintereinander folgende Punkte**, die keinen semantischen Gehalt haben, gelten als Separatoren und werden nicht erfasst:
    a. 

    in Deutſcher Sprache
    in alten Sprachen
    in lebenden ausl̊ändiſchen
    Sprachen

15. **Silbentrennung:** Im Reichsanzeiger erfolgt die Silbentrennung meist mit einem “double oblique hyphen” (⸗). Steht dieses Zeichen, muss es auch in der Transkription verwendet werden, obwohl es einem Istgleich-Zeichen (=) ähnelt. Istgleich-Zeichen kommen jedoch ebenfalls in anderen Kontexten vor. Daher ist auf diese Unterscheidung zu achten. **(Abweichung von OCR-D Level 2!)**

    a. 

    herrlich⸗

    b. 

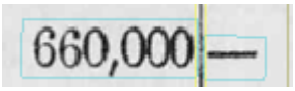
    auserlesener Zwei=Deutig-

16. **Minuszeichen (-), Bindestriche (-), Trennstriche (—):** lange Trennstriche (“Geviertstriche”) (—) kommen im Reichsanzeiger am häufigsten vor und unterscheiden sich vom kürzeren Bindestrich (-).

Februar-März —, per März-April

    a. **Geviertstrich**:
    

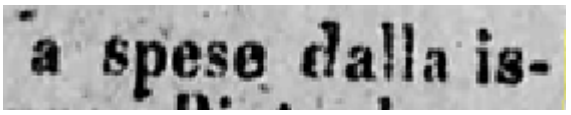
    Rhede. — Von den



660,000 —

— Die Deputation

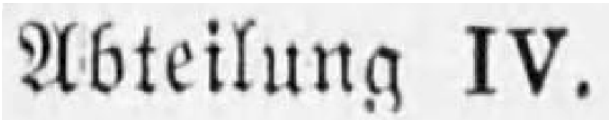
b. **Bindestrich**:



a spese dalla is-

Fonds- und Staats-Papiere.

17. ~~**Römische Zahlen:** Römische Zahlen sehen wie die Großbuchstaben X I M V aus, besitzen aber eigene Zeichen, die nicht auf einer gängigen Tastatur gefunden werden können und deshalb als Sonderzeichen in die Transkription eingefügt werden müssen (siehe Sonderzeichenliste). **(Abweichung von OCR-D Level 2: Eventuell muss im Nachgang eine Normalisierung erfolgen)**:~~

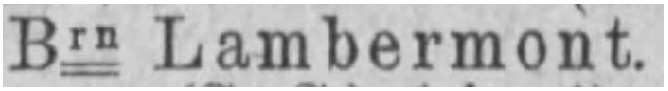
~~a.~~ 

~~*Abteilung I V.*~~

~~b.~~ 

~~*Neubau, I I . Stock,*~~

18. **Hochgestellte Buchstaben**, die als Abkürzung dienen, werden **nicht** hochgestellt transkribiert

a. 

Brn Lambermont.

19. **Hoch- und tiefgestellte Zahlen** werden vorlagentreu transkribiert, indem die entsprechenden Sonderzeichen (siehe 3. Sonderzeichenliste für Transkribus) verwendet werden. Es ist darauf zu achten, die Baseline korrekt zu setzen: die tief- oder hochgestellte Zahl orientiert sich nicht an der Baseline, sondern an der zugehörigen Zahl in Normalschreibung.

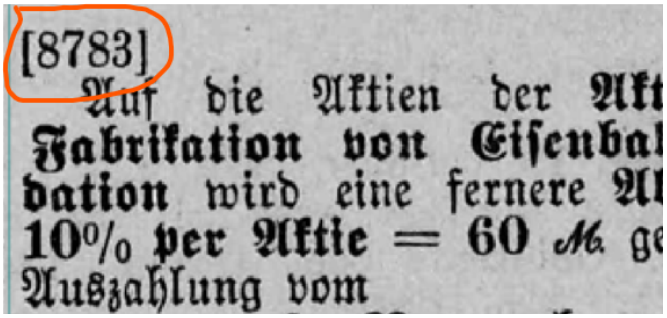
a. 

3₄

b. 

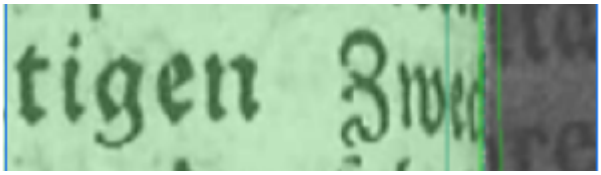
2²

20. **Eckige Klammern** werden ebenfalls transkribiert:

a. 

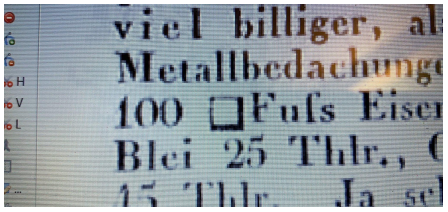
[8783]

21. Verhindert die **Buchbindung** die Lesbarkeit, wird nur bis zu jenem Buchstaben transkribiert, der sichtbar im grünen Rechteck liegt:

a. 

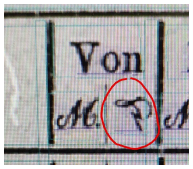
tigen Zwec

22. Quadrat als Symbol:

a. 

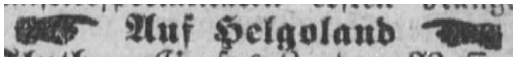
100 □ Fufs Eise

23. Symbol die um 90° oder 180° rotiert gedruckt wurden, werden, solange es sich nicht um ein systematischen Vorkommen handelt, ohne Rotation transkribiert. Systematische Fälle müssen abgeklärt werden. **Bitte immer in der Excel-Datei vermerken.**

    a. 

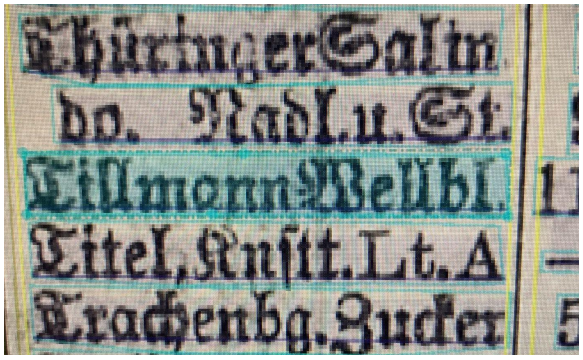
    M ℔

24. Das Symbol der **zeigenden Finger** wird ebenfalls transkribiert:

    a. 

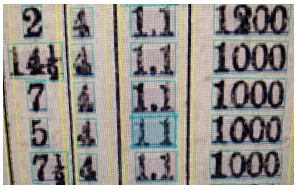
    ☛ *Auf Helgoland* ☚

25. Trennung von Wörtern **bei geringem oder keinem Leerraum**:
    Bei geringem Zeichenabstand (besonders in Tabellenspalten) sind aufeinanderfolgende Wörter teils ohne eingeschaltetes Leerzeichen gedruckt. Bei existierender Interpunktion (Kommata, Punkt) zwischen den Wörtern oder bei einem folgenden Nomen wird ein Leerraum eingefügt. Kann nur intellektuell eine Trennung vorgenommen werden, wird der Leerraum nicht transkribiert.

    a. 

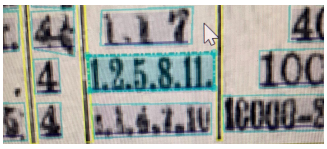
    Thüringer Salin
    do. Nadl. u. St.
    Tillmann Wellbl.
    Titel, Knftt. Lt. A
    Trachenbg. Zucker

26. **Fehlende Interpunktion**:
    Bei fehlenden Interpunktionen, beispielsweise durch unvollständigen Druck, wird diese Interpunktion nicht ergänzt. Entsteht durch das Fehlen der Interpunktion ein Leerraum, muss aus dem Kontext erschlossen werden, ob ein Leerraum eingefügt wird oder nicht.

a. 

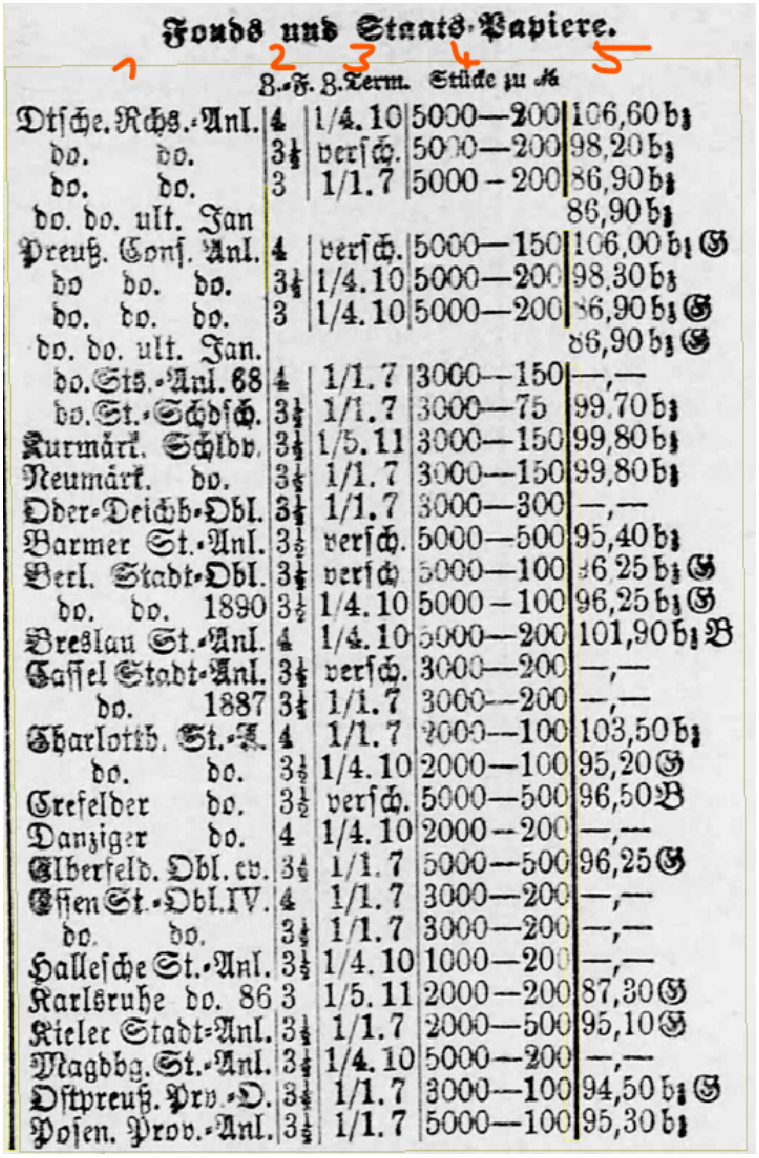
11 oder 1 1

27. **Leerräume nach Interpunktionen** bei Zahlen:
Werden Zahlen durch Punkte getrennt, wird kein Leerzeichen transkribiert. Bei Aufzählungen mittels Kommata hingegen wird ein Leerraum eingefügt.

a. 

1.2.5.8.11.

28. Erklärung einer **“Fonds- und Staats-Papier”-Tabelle** im Reichsanzeiger



Fonds und Staats-Papiere.

	Z.-F.	Z. Term.	Stücke zu ℳ	
Dtsche. Rchs.-Anl.	4	1/4. 10	5000—200	106,60 bz
do. do.	3½	versch.	5000—200	98,20 bz
do. do.	3	1/1. 7	5000 – 200	86,90 bz
do. do. ult. Jan				86,90 bz
Preuß. Conf. Anl.	4	versch.	5000—150	106,00 bz G
do do. do.	3½	1/4. 10	5000—200	98,30 bz
do. do. do.	3	1/4. 10	5000—200	86,90 bz G
do. do. ult. Jan.				86,90 bz G
do. Sts.-Anl. 68	4	1/1. 7	3000—150	—,—
do. St.-Schdsch.	3½	1/1. 7	3000—75	99,70 bz
Kurmärk. Schldv.	3½	1/5. 11	3000—150	99,80 bz
Neumärk. do.	3½	1/1. 7	3000—150	99,80 bz
Oder-Deichb.-Obl.	3½	1/1. 7	3000—300	—,—
Barmer St.-Anl.	3½	versch.	5000—500	95,40 bz
Berl. Stadt-Obl.	3½	versch	5000—100	96,25 bz G
do. do. 1890	3½	1/4. 10	5000 – 100	96,25 bz G
Breslau St.-Anl.	4	1/4. 10	5000—200	101,90 bz B
Cassel Stadt-Anl.	3½	versch.	3000—200	—,—
do. 1887	3½	1/1. 7	3000—200	—,—
Charlottb. St.-A.	4	1/1. 7	2000—100	103,50 bz
do. do.	3½	1/4. 10	2000—100	95,20 G
Crefelder do.	3½	versch.	5000—500	96,50 B
Danziger do.	4	1/4. 10	2000—200	—,—
Elberfeld. Obl. ev.	3½	1/1. 7	5000—500	96,25 G
Essen St.-Obl. IV.	4	1/1. 7	3000—200	—,—
do. do.	3½	1/1. 7	3000—200	—,—
Hallesche St.-Anl.	3½	1/4. 10	1000—200	—,—
Karlsruhe do. 86	3	1/5. 11	2000—200	87,30 G
Kieler Stadt-Anl.	3½	1/1. 7	2000—500	95,10 G
Magdbg. St.-Anl.	3½	1/4. 10	5000—200	—,—
Ostpreuß. Prv.-O.	3½	1/1. 7	3000—100	94,50 bz G
Posen. Prov.-Anl.	3½	1/1. 7	5000—100	95,30 bz

Spalte 1: Das ist der Name der Anleihe. Das do. steht für dito, also wiederholen. Somit sind die ersten drei Zeilen, erste Spalte „Deutsche Reichs-Anleihe“.

Spalte 2: Das ist die Coupon Rate, also die Höhe der Verzinsung, erste drei Zeilen entsprechend 4%, 3.5%, 3%.

Spalte 3: Es handelt sich um die Zeitpunkte der Zinszahlungen.

Bsp. „1 /4. 10“ -> Zinszahlungen am 1. April und am 1. Oktober

Das Wort „versch.“, also verschiedene, bezieht sich darauf, dass es keine festen zwei Tage im Jahr für die Zinszahlungen gibt.

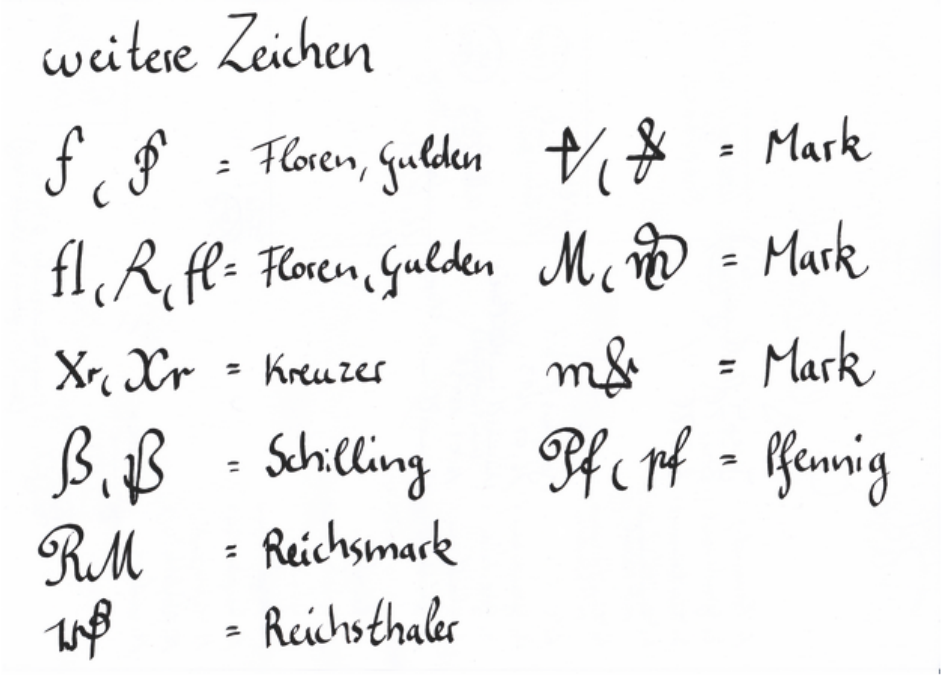
Spalte 4: Anzahl der ausgegebenen Anteilsscheine und deren nominaler Wert (face value). Beispielsweise wurde in Zeile 1 eine Anleihe über eine Million ausgegeben, bestehend aus 5000 Anteilen zu je 200.

Spalte 5: Preis zu dem gehandelt wurde, falls Angebot und Nachfrage sich getroffen haben. Die Kürzel hinter dem Wert sind „bz“, was für bezahlt, also gehandelt steht und das „G“ steht für Gesucht.

29. **Symbole aus dem Nationalsozialismus**
    Die in den Zeiten des Nationalsozialismus häufig verwendeten Symbole werden in einer normalisierten Form transkribiert.
    a. 
       Reichsführer SS und

30. **Briefsymbole**
    Die Postanrschift wird teilweise mit einem Briefsymbol repräsentiert. Das Symbol wird ✉ (https://www.compart.com/de/unicode/U+2709) mit Unicode konform erfasst.
    a. 
       ✉ u. N

31. Abkürzungen und Symbole



Quelle: https://www.ahnenwiki.at/eine-reise-durch-die-vergangenheit/symbole-im-kirchenbuch/

a. Für Münzen.

B. A. = Bankassignation, Bankanweisung.
B. B. = Bankbillet.
Bco., B° = Banko.
B. Z. = Bankzettel.
mk = Bankomark.
C. A. = Kassenanweisung.
C. B. = Kassenbillet.
Crt., Cour. = Courant.
Ck = Courantmark.
c., ct., cs., cts. = Centime, Cent, Centesimo.
d., ₰ = denier, denaro, penny, Pfennig; ₰ vls. = Grot (Pfennig) vlämisch.
Duc. ♯ = Ducaten.
Duc. d. C., Duc. d. R. = Ducato de Cambio, Ducato del Regno.
Drchm. = Drachme.
f., fl. = Gulden.
C-fl. = Courant-Gulden.
Fd'or., Friedrd'or = Friedrichsd'or.
Fr., Frs., Frcs. = Franc, Francs.
Gr. = Groschen.
Gt. = Grot.
Kop. = Kopeken.
Kr., kr., Kr. = Kreuzer.
Krthlr., Bbthlr. — Kronthaler, Brabanterthlr.
£ = Lira, Lire.
Lsterl. = Livre (Pfund) Sterling.
Ls. ts. = Livres tournois.
Ld'or = Louisd'or.
Ld'orthlr. = Louisd'or-Thaler, Goldthaler.
M. Z. = Meßzahlung.
Ngr., ngr. = Neugroschen.
Oesterr. W. = Oesterreichische Währung.
Pr. Crt. = Preußisch Courant.
Rthr. thlr. = Thaler, Vereinsthaler.
Rbdlr., Rbthlr. = Rigsbankdaler, Reichsbankthaler.
Roñ., Rpta. = Reales de Vellou, Reales de Plata.
Rs. = Reïs, Rees; m, ‖, : = Milreïs.
Rbl. B°, Rbl. S. = Rubel Banco, Silberrubel.
S. = Soldo, Soldi.
Sch., s, ßl. = Schilling.
Sgr = Silbergroschen.
Spthlr. = Speciesthaler.
St. = Stüber, Stüver.
Südd W. = Süddeutsche Währung.
W. W. = Wiener Währung.
W. G., W. Z. = Wechselgeld, Wechselzahlung.
Dollar, Piaster.
⊙ = Gold, ☽ = Silber, ○ = Kupfer.

b. Für Raummaße.

Ell., ϱ = Elle.
aune = (franz.) Elle.
Bbϱ. = Brabanter Elle.
R., (°) = Ruthe.
F., (′) = Fuß.
Z., (″) = Zoll.
L., (‴) = Linie.

1*

c. Für Gewichte.

Arroba.
Avdp. = Avoir du poids (engl. Handelsgewicht).
car. = (engl.) carat, carats.
Ctr. = Centner.
Cwt. = Hundredweight, (engl.) Centner.
Cant. = Cantaro, Centner.
dwt. = Pennyweight, (engl.) Pfenniggewicht.
gr. = grain.
Gr. = Gramme, Grän.
K°., Kilo. = Kilogramme.
Lth. = Loth.
mk. = Mark.
℔ = Pfund.
Sch℔. = Schiffspfund.

d. Für Handelsgebräuche.

Av. = Avance, Gewinn.
Btto. = Brutto.
compt., Cont. = comptant, Contant, baar.
do. = ditto, desgleichen.
Fol. = Folio.
gez. = gezeichnet.
Ggw. = Gutgewicht.
K. S. = Kurze Sicht.
O|: = Ordre.
ord. = ordinair.
p. a., p. Mt. = per Anno; per Monat.
p. c., % = Procent.
p. m., ‰ = pro Mille.
Sig. = Signum; sign. = signirt, gezeichnet.
Sopta. = Sopratara.
Ta. = Tara.
ult. = ultimo.
u. r. = ut retro, wie umstehend.
u. s. = ut supra, wie oben.
|c. = Hundert.
|m. = Tausend.

e. Für Sprachen.

engl. = englisch.
fr. = französisch.
it. = italienisch.
sp. = spanisch.
holl. = holländisch.

f. Arithmetische Zeichen.

+ Zeichen der Addition.
— Zeichen der Subtraction.
× oder · Zeichen der Multiplication.
: Zeichen der Division.
= Zeichen der Gleichheit.

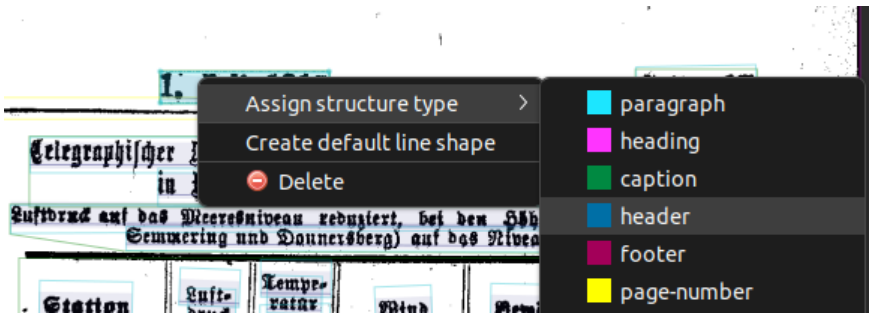
Quelle: Die gesammelten Handelswissenschaft. Friedrich Heinrich Schlössing, Berlin, 1863

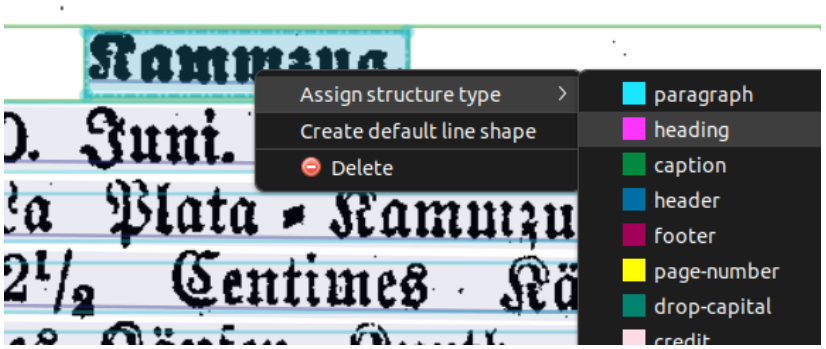
1.1. Ergänzung zum Teilprojekt ***Austrian Newspapers***

In dem folgenden Kapitel sind die wichtigsten Punkte bei der Korrektur der Transkription der *Austrian Newspapers* (https://github.com/UB-Mannheim/AustrianNewspapers) aufgelistet.

Hervorhebung der relevantesten Punkte, sowie werksspezifische Ergänzungen

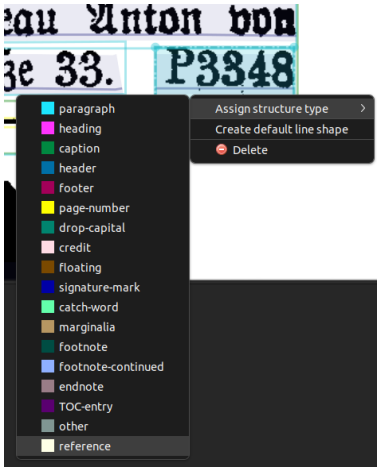
32. **Taggen der TextRegionen** (Header..)
    Regionen sollten in eine der folgenden Kategorien eingeordnet werden: *header*, *footer*, *footnote*, *heading*, *paragraph*, *page-number*, *reference*.

a)


b)


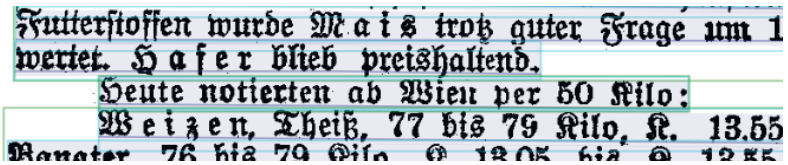
a) Tagging TextRegion als *header*
b) Tagging TextRegion als *heading*

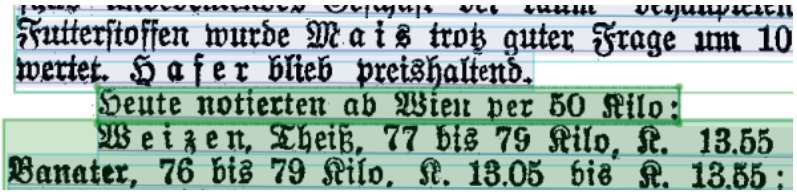
Heading und Reference können auch als TextLine-Attribut gesetzt werden.

a)


a) Tagging TextLine als *reference*

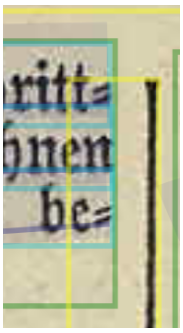
33. **Zusammenführen von zersplitterten TextRegionen**
Einige TextRegionen sind in sehr kleine Subregionen zersplittert und sollten zu passenden Regionen zusammengeführt werden.

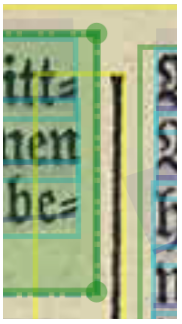
a)


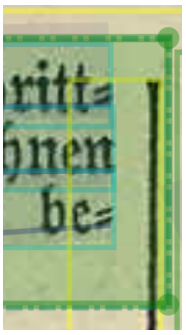
b)


a) Einzelne Zeile als TextRegion, obwohl kein offensichtlicher Grund für eine Zerteilung der theoretisch größeren Region erkennbar ist.
b) Zusammenführen der zwei TextRegionen. Bei der Auswahl mehrerer TextRegion ist die Reihenfolge entscheidend, bitte wählt die TextRegionen entsprechend der Lesereihenfolge hintereinander aus.

34. **Verbreiterung von TextRegionen**
TextRegionen sollten, wenn möglich, den gesamten Text umschließen und etwas Leerraum zwischen Text und Regiongrenze haben, ohne dabei Separatoren zu schneiden.

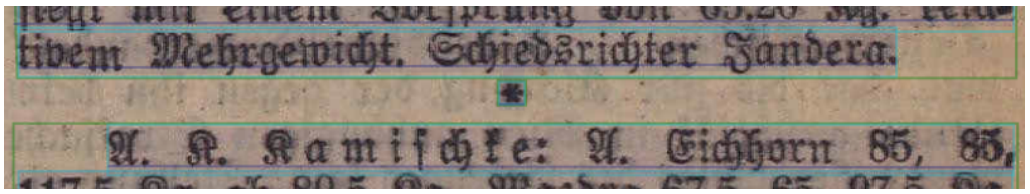
a)


b)


c)


a) TextRegion ist sehr nah am Text
b) TextRegion umschließt den Text ohne andere Strukturen zu schneiden (optimal)
c) TextRegion umschließt den Text und schneidet andere Strukturen (nicht optimal)

35. **Löschen von Separator TextRegions/TextLines**
TextLines die nur Graphic- oder Separatorelement enthalten und kein Textinhalt sollten entfernt werden.

a) 

a) Als Separator verwendeter Asterisk (nicht optimal)

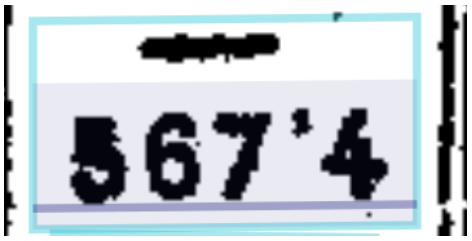
Graphic- oder Separatorelemente die zu Beginn einer Zeile stehen und bspw. als Aufzählungshilfe verwendet werden, sollen transkribiert werden.

b) 

b) Als Aufzählungszeichen verwendeter Asterisk -> * *Das Bedürfnis*

36. **Korrektur der Textline-Polygonzüge**

Der Text einer Zeile sollte vollständig umschlossen sein, ohne möglichst andere Zeilen oder Separatoren zu schneiden.

a) 

b 

c) 

d) 

a) TextLine Polygonzug umschließt ein Separator (nicht optimal)
b) TextLine Polygonzug schneidet multiple Separator (nicht optimal)
c) Der Polygonzug der unteren Zeile schneidet die darüber liegende Grafik und Zeile. (nicht optimal)
d) Der Polygonzug wurde korrigiert und umfasst nur noch den eigenen Text. (optimal)

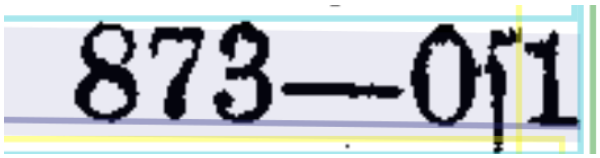
37. **Baselinekorrektur**
Die Baseline (Grundlinie) sollte alle zu transkribierenden Zeichen in einer TextLine abdecken und dabei nicht die Laufrichtung ändern. D.h. Korrekturen können in der Länge der Baseline sowie in der Laufrichtung vorgenommen werden.

a) 

a) Fehlerhafte Laufrichtung der Baseline (nicht optimal)

38. **Umgedrehtes langes s in Referenzen** (werksspezifisch)
In dem Werk gibt es Referenzen, die ein langes s bzw. ein umgedrehtes langes s aufweisen. In dem Werk wurde diese Variante mit dem UnicodeZeichen “Esh” ʃ transkribiert.

a) 

b) 

a) 12963 - 3ʃ3
b) 873-0ſ1

39. **Erfassen von speziellen Unicodes** (Zeigefinger, Quadrate)
Spezielle Unicode Zeichen wie der Zeigefinger, Quadrate, Dreiecke und der Vierpunkt sollen transkribiert werden.

a) 

b) 

c) 

a) :: *Eis- u* Zeigefinger wurde nicht erfasst (nicht optimal)
b) ☛ :: *Ei*
c) *Sattler* ☐ *Riemer*

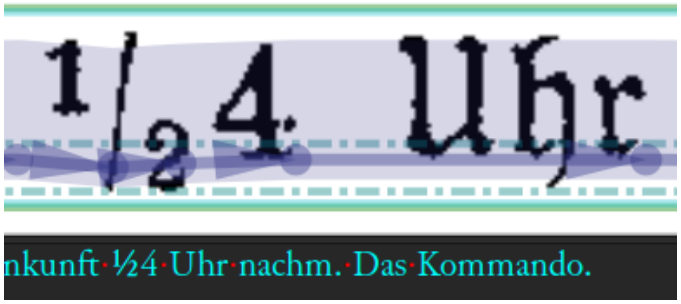
40. **Tausender Trenner: Punkt** (werksspezifisch)
    Im Vergleich zu dem Reichsanzeiger, bei dem der Tausendertrenner mit einem Leerzeichen transkribiert wurde, wird in diesem Werk ein Punkt gesetzt.

a) 

b) 

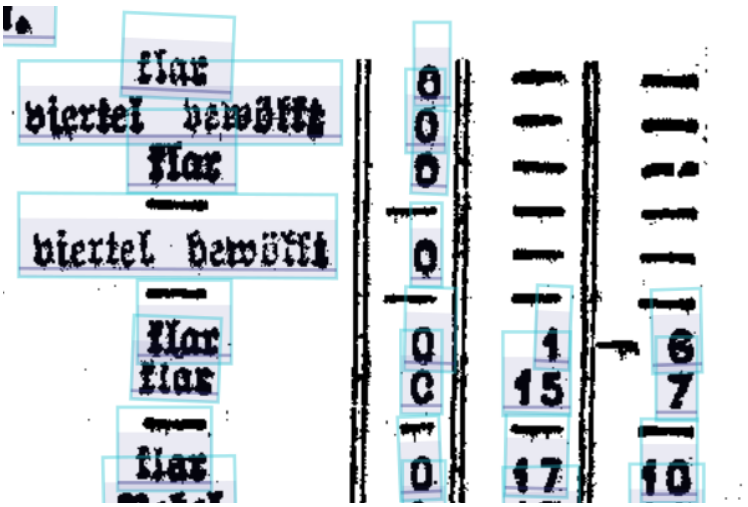
a) *59.904,400*
b) *120.273,155*

41. **Leerzeichentrenner bei Prozent, Bruch**
    Bei der vorliegenden Transkription wurden häufig die Leerzeichen zwischen Bruch und Zahl oder Zahl und Prozent nicht konsequent eingesetzt.

a) 

a) *½ 4 Uhr*

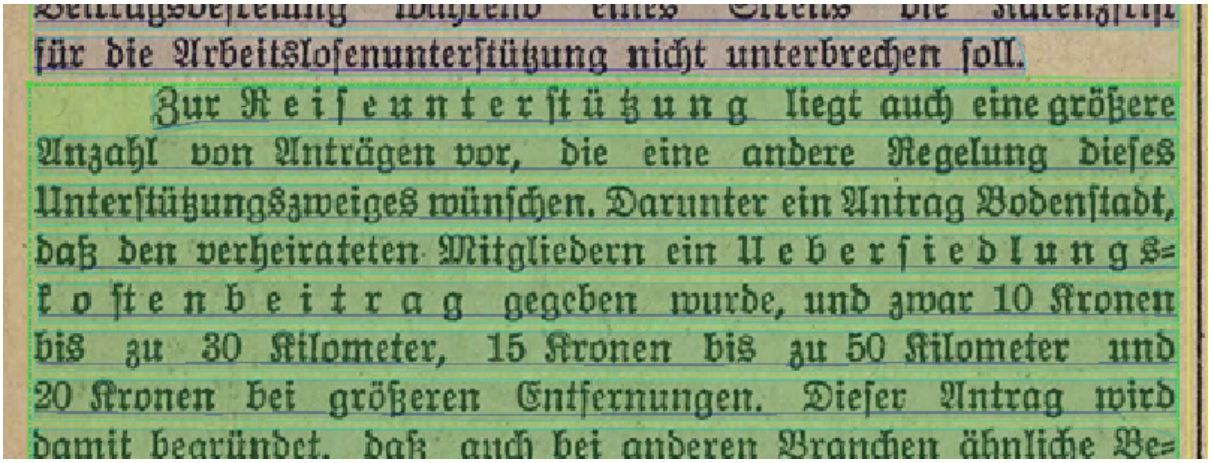
42. **Auslassungsstrichen in Tabelle ergänzen**
    Bei Tabellen wurden die Auslassungsstriche nicht erfasst.



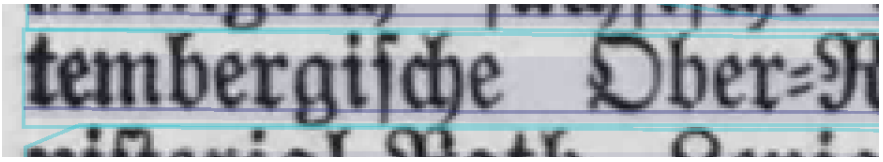
2. Layoutbearbeitung und -korrektur

1. Durch die Layoutkorrektur soll sichergestellt werden, dass Textregionen, Textlines und Baselines auf der zu transkribierenden Seite korrekt repräsentiert werden

2. Um die Korrektheit der Layouterkennung sicherzustellen, müssen 3 Elemente überprüft werden:
    a. die **Textregion** (grüne Markierung)
    b. die **Textline** (hellblaue Markierung)
    c. die **Baseline** (violette Markierung)

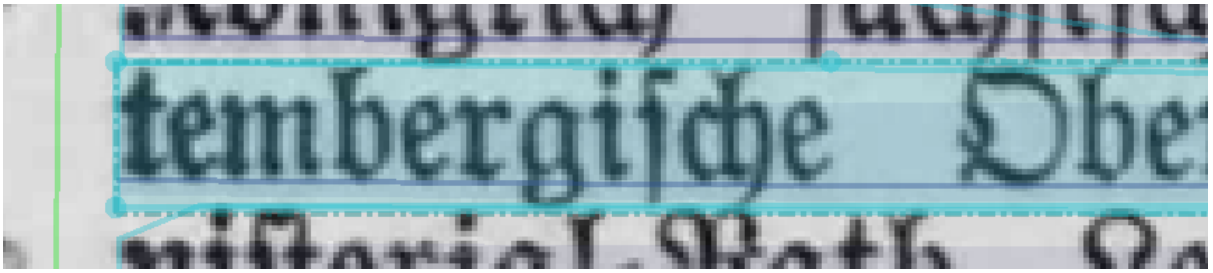


***Textregion**: grüne Markierung (rahmt als Rechteck oder Polygonzug einen Textabschnitt ein)*

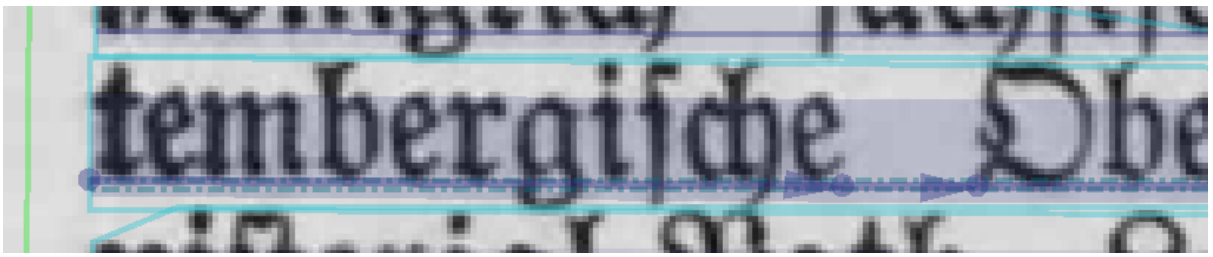


***Textline**: hellblaue Markierung (rahmt als Rechteck oder Polygonzug eine komplette Textzeile ein)*
***Baseline**: violette Linie (Grundlinie, auf der die Buchstaben "sitzen"; Unterlängen von Buchstaben ragen wie im Beispiel über die Baseline hinaus)*

3. Textregion, Textline und Baseline können per Linksklick separat angewählt werden:

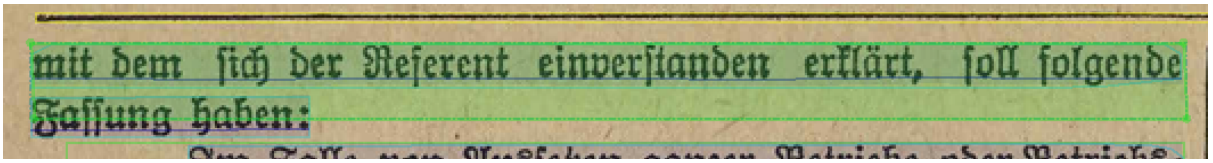


Ausgewählte Textline

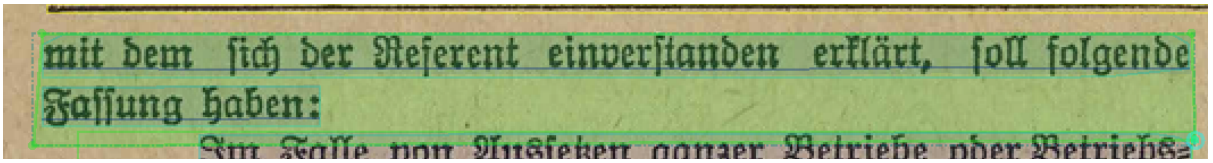


Ausgewählte Baseline

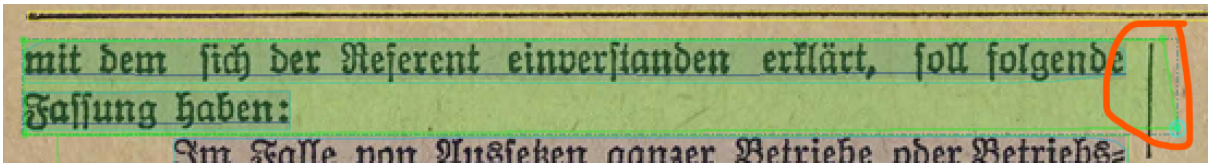
4. Korrektur der Textregionen:
    a. **Überprüfung der Textregion:** erste Textregion der Seite auswählen und überprüfen, ob der gesamte Textinhalt (alle Textlines) innerhalb des grünen Rechtecks liegen
    b. Im gleichen Arbeitsschritt überprüfen, ob die Textregion etwas Abstand zum Zeilenanfang- und ende besitzt
    c. ebenfalls überprüfen, ob die ausgewählte Textregion andere Textregionen oder Separatoren (schwarze Trennstriche) schneidet



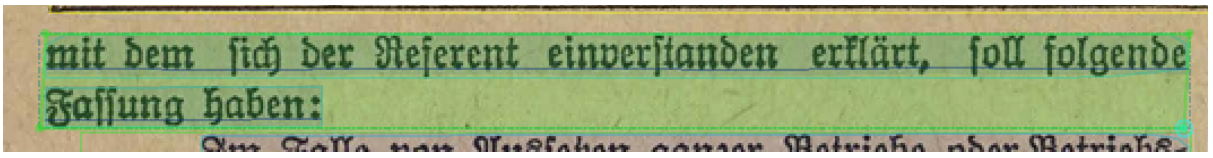
***Falsch**: ein Teil der zweiten Textzeile liegt außerhalb der Textregion (grüne Markierung)*



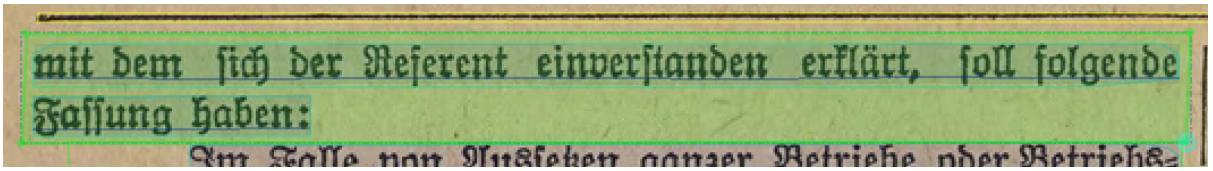
***Falsch**: die Textregion umfasst die ersten beiden Textzeilen, schneidet aber die folgende Textregion*



***Falsch**: die Textregion schneidet rechts einen Separator*

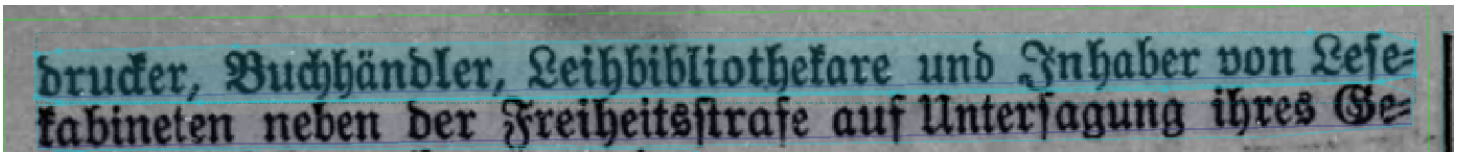


***Falsch**: die Textregion liegt zu eng an den Zeilen*

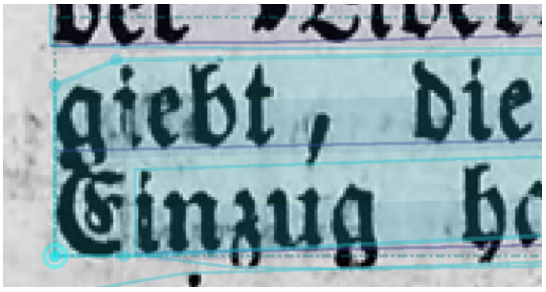


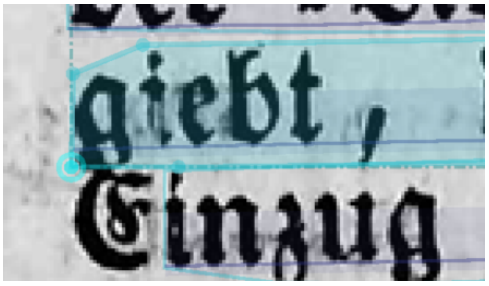
Richtig

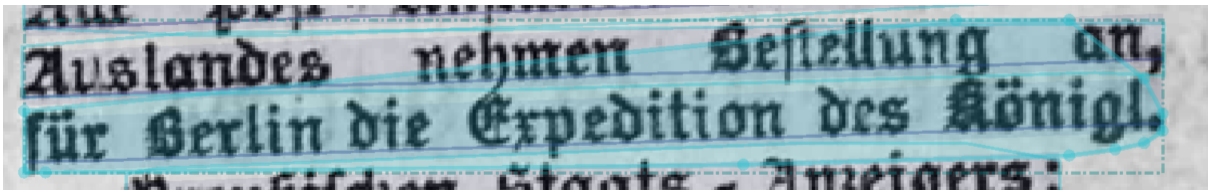
5. Korrektur von Textline und Baseline:
    a. **Überprüfung der Textline**: erste Textline auswählen und überprüfen, ob alle Buchstaben der Textzeile korrekt innerhalb der hellblauen Markierung liegen:

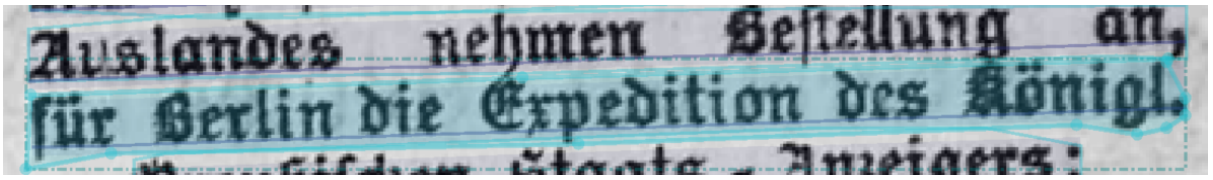


die Textline im Beispiel ist korrekt festgelegt

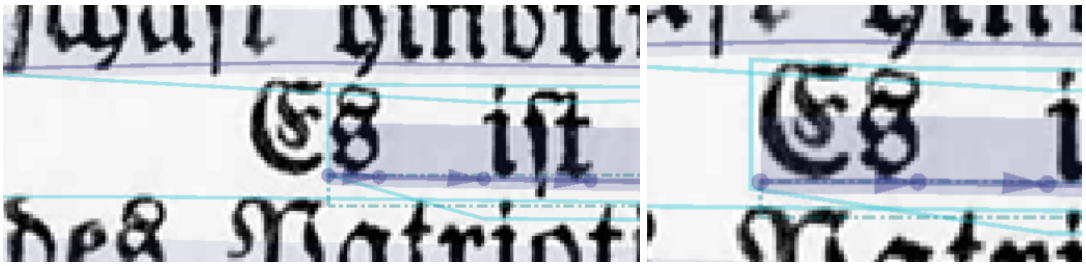
(a) 

(b) 

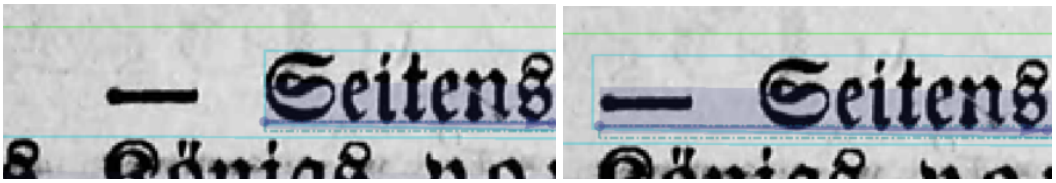
(c)


(d)


***Fehler (1)**: Beispiel (a) und (c) zeigen eine Textline, die falsch gesetzt ist und über zwei Textzeilen ragt. Mittels der hellblauen Punkte kann der Polygonzug so verändert werden, dass die Textline nur eine Zeile wie im Beispiel (b) und (d) umfasst*



***Fehler (2)**: Das linke Beispiel zeigt eine Textline, die den ersten Großbuchstaben der Zeile nicht umfasst. Der Polygonzug muss so verändert werden, dass alle Buchstaben der Zeile innerhalb der hellblauen Markierung liegen und die Baseline muss entsprechend verlängert werden (siehe auch Bearbeitungsschritt b.) (rechtes Beispiel)*

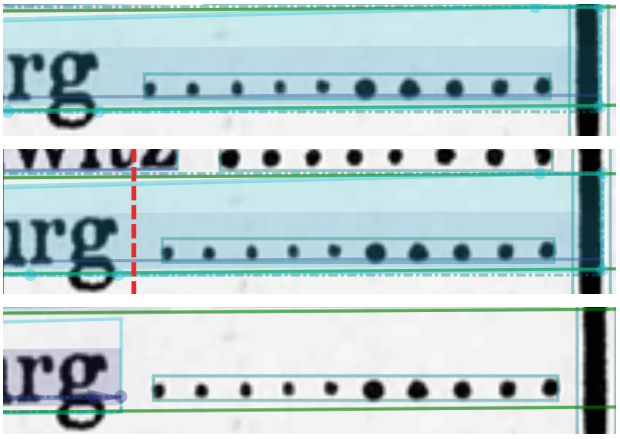


***Fehler (3)**: analog zu Fehler (3) spart das linke Beispiel einen Geviertstrich als Zeilenanfang aus. Auch hier muss die Textline (und die Baseline so verändert) werden, dass der Strich korrekt erfasst ist (rechtes Beispiel)*

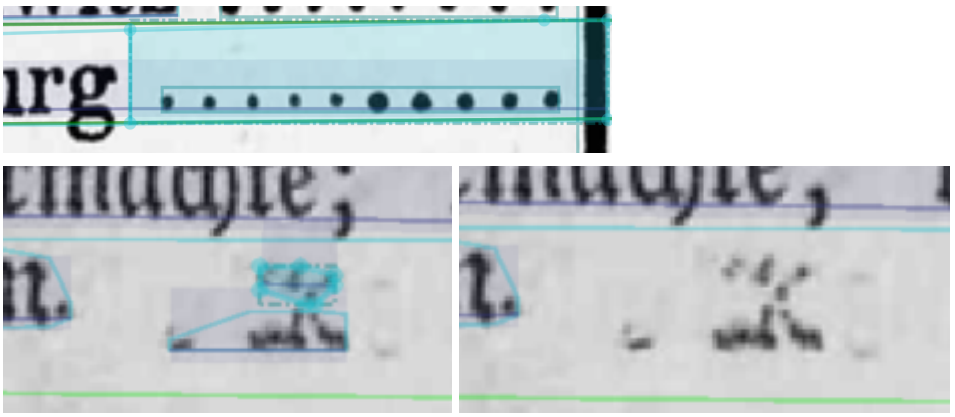
***Fehler (4)**: teilweise erkennt die Layouterkennung zwei getrennte Zeilen, obwohl es sich tatsächlich um eine zusammenhängende Textzeile handelt. Im linken Bsp. wurde das "p" als 1. Textline erkannt und das "flegt," als 2. Textline. Beide Textlines müssen verbunden werden. Hierfür werden zunächst beide Textlines ausgewählt (Mit Strg + Linksklick können mehrere Textlines ausgewählt werden). Analog kann man diesen Schritt auch über die Baseline machen. Danach werden diese über den "Merge"-Button in Transkribus verbunden (rechtes Beispiel):*

Eine manuelle Korrektur der Textline oder Baseline kann trotzdem notwendig sein

***Fehler (5)**: Separator-Elemente, die keinen Mehrwert zum textualen Kontext beitragen, sollen aus den Textlines ausgeschlossen werden. Dazu zählt etwa eine Vielzahl aufeinanderfolgender Punkte.*



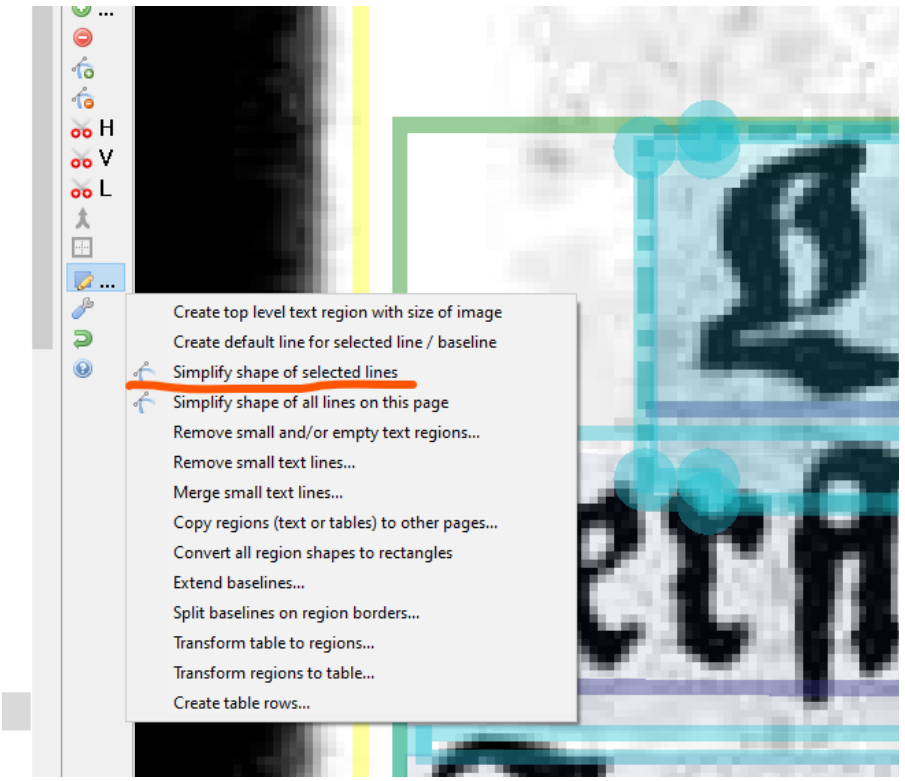
***Abtrennung von Separatoren:** Auswahl der Textline, Abtrennung mittels horizontalem Schneidwerkzeug, Löschen der abgetrennten Textline*



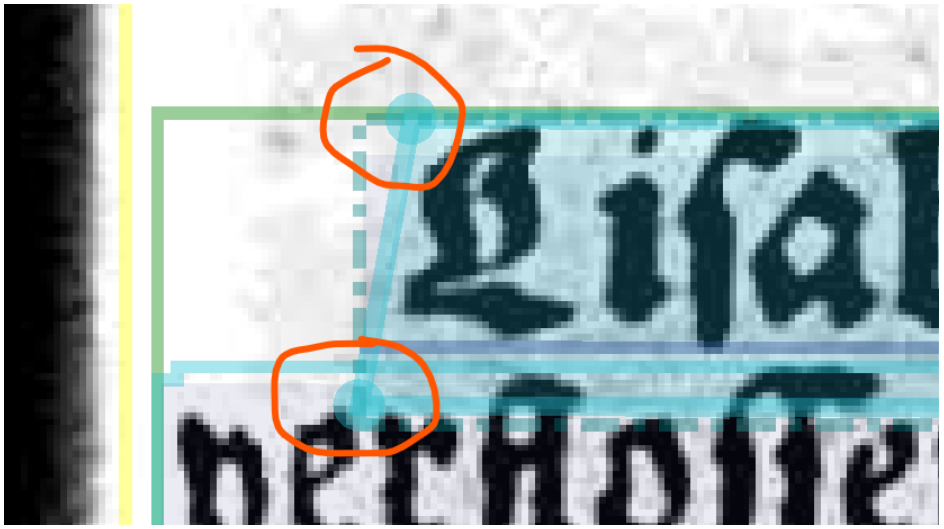
***Fehler (5)**: Hin und wieder werden Verschmutzungen als Textline erkannt. Diese müssen gelöscht werden: Textline per Klick auswählen und über Entf-Taste löschen*

b. **Bearbeitung mehrerer Punkte eines Polygonzugs:** teilweise erschweren zu viele Punkte im Rahmen einer Textline die schnelle Korrektur. Über die Transkribus-Funktion “Simplify shape of selected line” können überzählige Punkte gelöscht und dadurch die Bearbeitung vereinfacht werden:

(1) Textline mit überflüssigen Punkten im Polygonzug auswählen



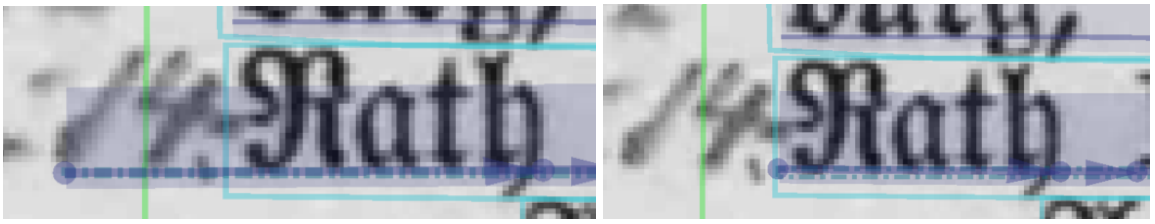
(2) Auswahl der “Simplify shape of selected Lines”-Funktion



(3) Der vereinfachte Polygonzug

c. **Überprüfung der Baseline**: Nachdem die Korrektheit der Textline überprüft worden ist, muss die Baseline überprüft werden. Hierfür wird die Baseline der ersten Textzeile ausgewählt und überprüft, ob alle Buchstaben korrekt auf der Grundlinie sitzen:

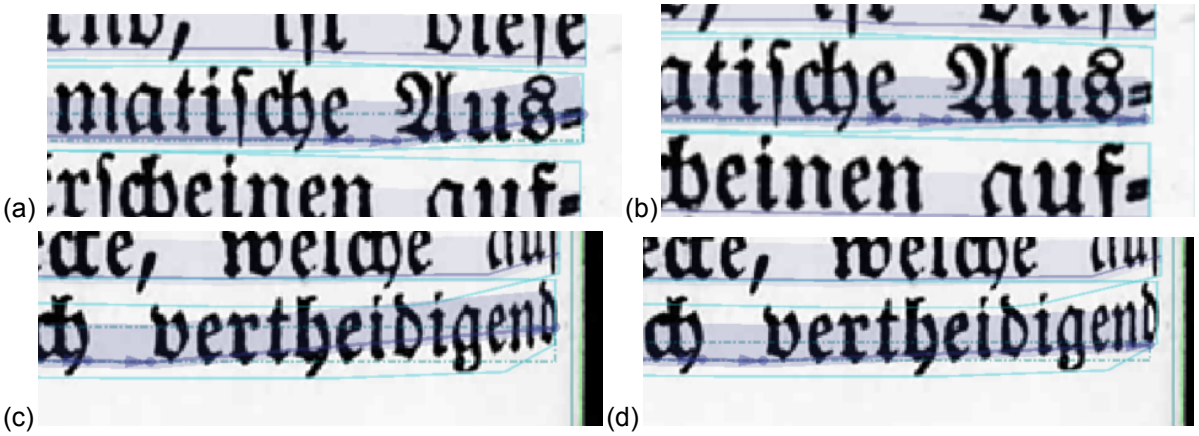
die Textline im Beispiel ist korrekt festgelegt



die Baseline im linken Beispiel ist nicht korrekt festgelegt: sie darf nicht über die Textline (hellblaue Markierung) hinausragen. Korrekt ist das rechte Beispiel

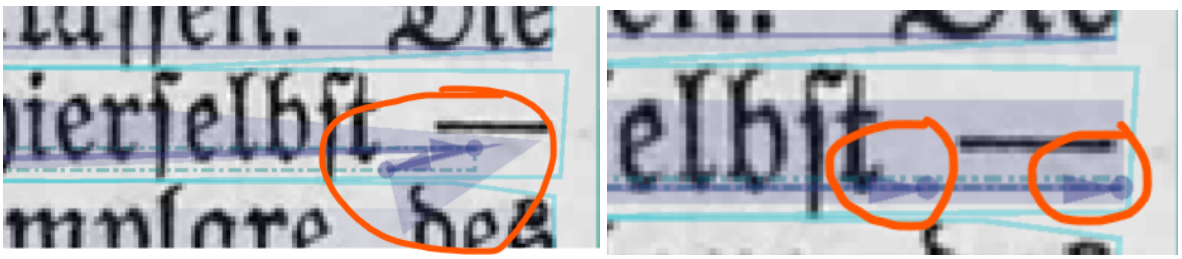
(a) (b)

(c) (d)



***Fehler (1)**: in (a) und (c) ist das Ende der Baseline verschoben: die Buchstaben sitzen nicht korrekt auf der Grundlinie. Über die violetten Punkte der Baseline kann diese so angepasst werden, dass die Buchstaben korrekt auf der Grundlinie wie in den Beispielen (b) und (d) sitzen*

***Fehler (2)**: im linken Beispiel beginnt die Baseline in der Buchstabenmitte und nicht am Anfang. Die korrekt angepasste Baseline zeigt das rechte Beispiel*

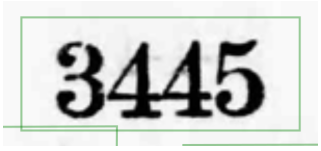


***Fehler (3)**: Die Baseline besitzt eine Lauf- bzw. Leserichtung (von links nach rechts). Die kleinen Pfeile geben an, wie die OCR die Buchstabenreihenfolge lesen wird und muss mit dem Ursprungstext übereinstimmen. Im linken Beispiel sind die Pfeile nicht korrekt ausgerichtet, so dass die OCR bis zum Geviertstrich vorgehen würde, um danach wieder zurückzukehren. Das rechte Beispiel zeigt die korrekt ausgerichtete Baseline: hier ist auch die korrekte Leserichtung gegeben*

d. **Textregionen ohne erkannte Textlines**:

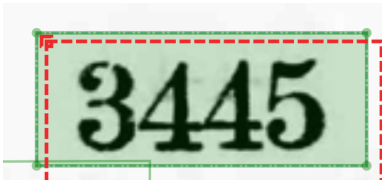
Wurden durch die Vorverarbeitung keine Textlines in einer Textregion erkannt, müssen die Textlines und die entsprechenden Baselines händisch eingefügt werden. An dieser Stelle gibt es zwei Standard-Vorgehensweisen: In der ersten Varianten wird über das Platzieren der Baseline eine Textregion automatisch erzeugt und die Textline muss händisch korrigiert werden, bei der zweiten wird zunächst die Textline und dann die Baseline definiert. Eine Schritt für Schritt Anleitung für Variante 2 (äquivalent kann Variante 1 durchgeführt werden):

I. Auswahl der Textregion durch Anklicken



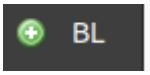
II. Hinzufügen der Textline durch Anklicken des “L”-Buttons

und Ziehen einer Textline, die den gesamten Textinhalt umfasst.



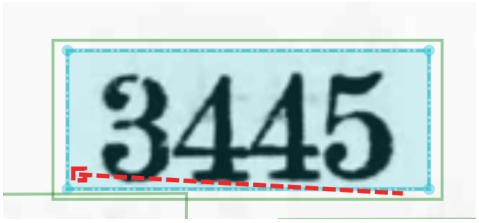
Sollte die Textline über die Textregion hinausragen, muss der Polygonzug noch über die Manipulation einzelner Polygonpunkte angepasst werden.

III. Um eine Baseline zu setzen, muss die eben erstellte Textline durch Anklicken aktiviert werden. Die Baseline kann dann durch Anklicken des “BL”-Buttons hinzugefügt werden.



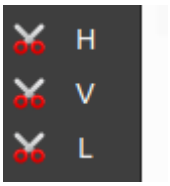
Der erste Punkt wird durch Maus-Links-Klick unterhalb des ersten Buchstaben gesetzt. Durch weitere einfach Klicks kann der Pfad exakt

definiert werden. Der letzte Punkt wird durch ein Maus-Links-Doppelklick unterhalb des letztens Buchstabens gesetzt.



IV. Zerschneiden der Textlines und Baselines mittels Werkzeugen

1. Werkzeuge zum Zerschneiden

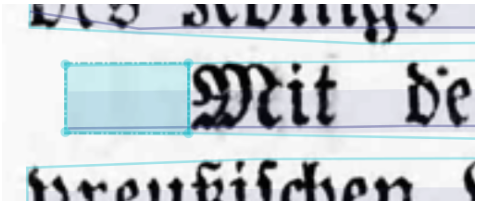


H - **H**orizontaler Schnitt
V - **V**ertikale Schnitt
L- Schnitt anhand einer benutzerdefinierten **L**inie

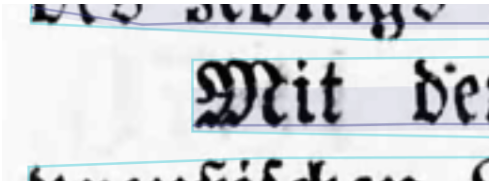
V. Wenn die Textline aktiviert ist, kann mittels der obig gezeigten Werkzeuge die Textline und die Baseline gleichzeitig gekürzt werden

Auswahl der Textline

Auswahl des vertikalen Schneidwerkzeugs



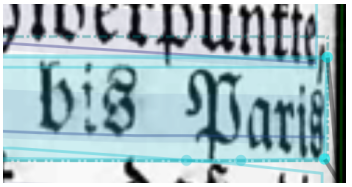
Aktivierung des zu entfernenden Teils



Entfernung durch Klicken der “Entf”-Taste

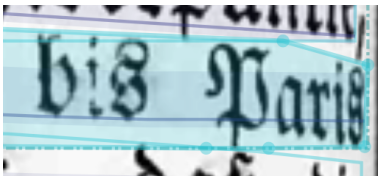
VI. Erweiterung/Verkürzung der Polygonzüge mittels Werkzeugen

1. Erweiterung



Auswahl des zu erweiternden Polygonzuges

Auswahl des Werkzeuges zum Erweitern um ein Polygon

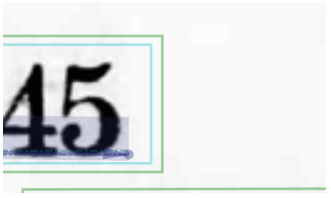


Durch klicken an der vorgesehen Stelle wird ein Polygon hinzugefügt und kann anschließend händisch angepasst werden

2. Verkürzung

Auswahl des zu bearbeitenden Polygonzuges

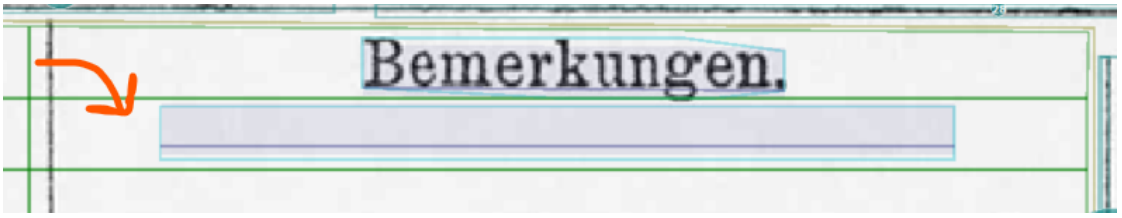
Auswahl des Werkzeuges zum Entfernen von Polygonen



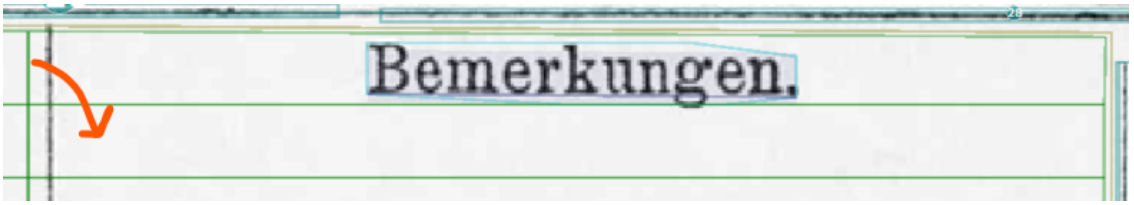
Durch Anklicken des Polygons wird dieser entfernt

e. **Zeilenkorrektur in Tabellen**: Komplexe Tabellen stellen die Layouterkennung aufgrund des komplizierten Aufbaus vor größere Probleme. Deshalb kommt es in Tabellenzellen häufiger zu einer falschen Zeilenerkennung. Folgende Probleme tauchen unter anderem auf:

(a)

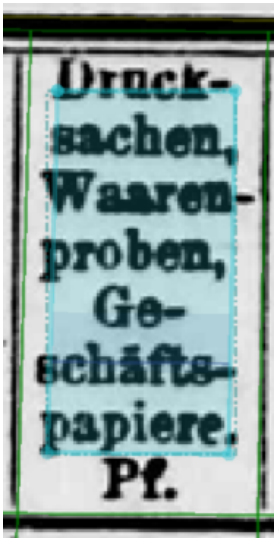


(b)

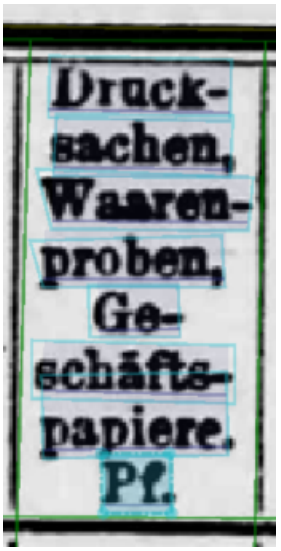


Fehler (1): **Leere Tabellenzellen mit Textline und Baseline**: Beispiel (a) zeigt eine leere Tabellenzelle, die dennoch Text- und Baseline besitzt. Hat eine Tabellenzelle keinen Inhalt, müssen Text- und Baseline gelöscht werden (b).

(a)

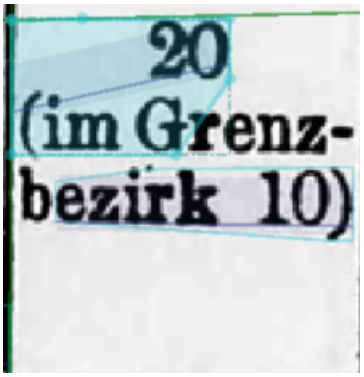


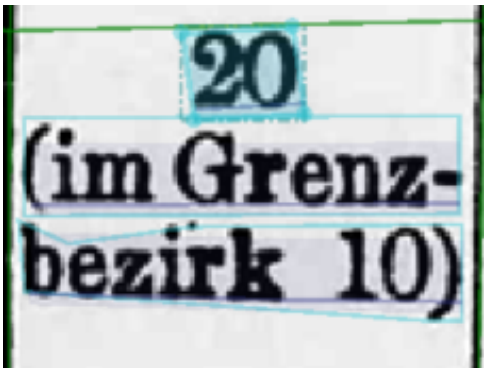
(b)



Fehler (2): **Tabellenzellen mit mehrzeiligem Text aber nur einer Text- und Baseline**: Beispiel (a) zeigt eine Tabellenzelle mit mehrfach umgebrochenem Text (mehrere Textzeilen), der von der Layouterkennung allerdings nur eine Text- und Baseline zugeordnet wurde. Um zeitsparend zum korrekten Ergebnis (b) zu kommen, sollten im ersten Schritt Text- und Baseline in Beispiel (a) gelöscht werden. Im zweiten Schritt können mit dem Baseline-Werkzeug 

für jede Textzeile neue Baselines angelegt werden, die gleichzeitig auch die Textlines erstellen (nach Bestätigung des PopUp-Fensters, das auf die Creation einer Parent-Region (= Textline) hinweist). Im dritten und letzten Schritt müssen nur noch die Textlines angepasst werden

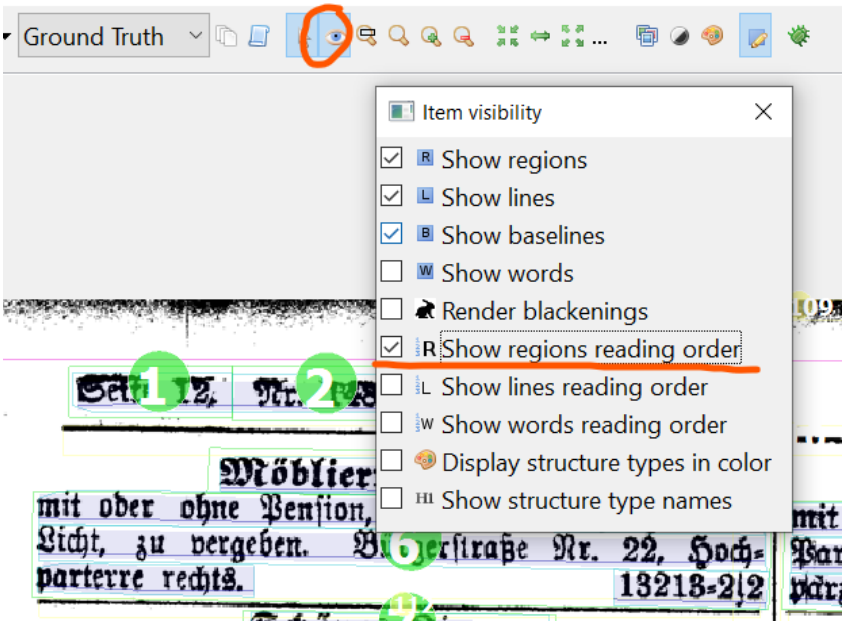
(a) 

(b) 

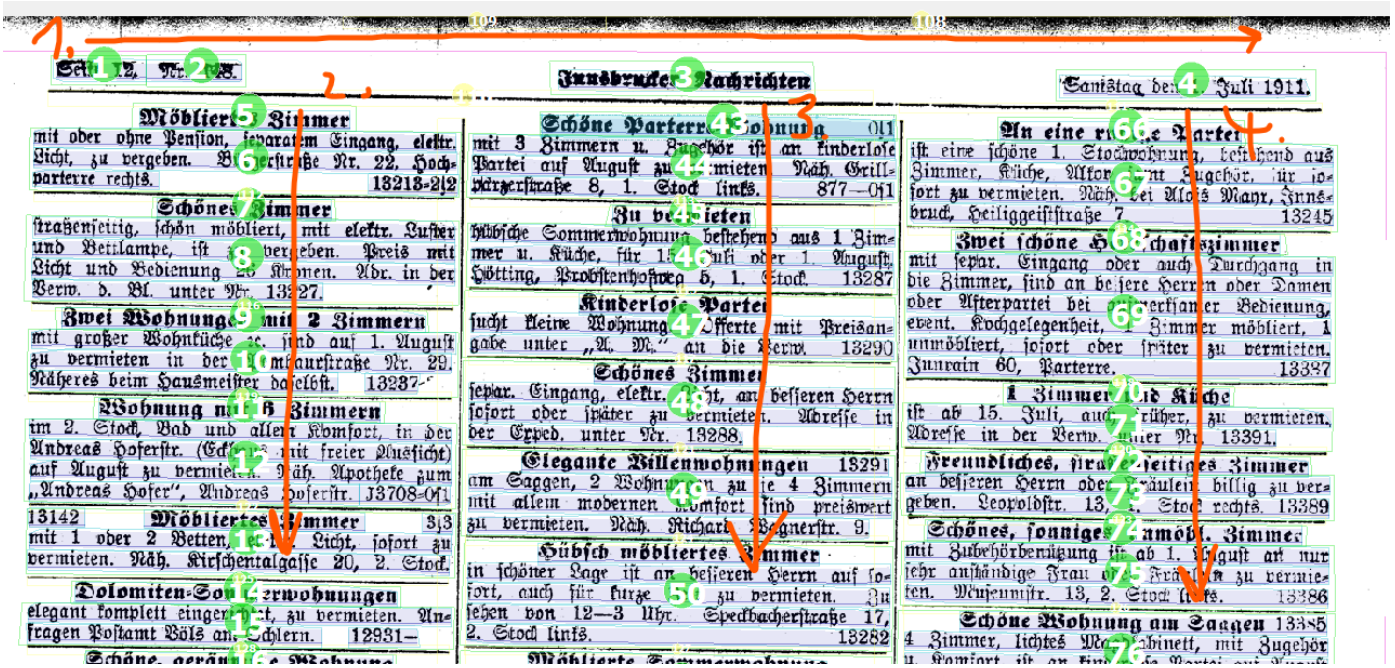
Fehler (3): **Tabellenzellen mit mehreren falschen Text- und Baselines**: teilweise werden in Tabellenzellen mit mehreren Textzeilen mehrere falsche Text- und Baselines wie in Beispiel (a) gesetzt. Auch hier ist es, wie bei Fehler (2), effizienter Text- und Baselines zu löschen und neu zu setzen wie in Beispiel (b)

f. Nach Überprüfung der Baseline wird mit der Cursor-Nach-Unten-Taste zur nächsten Zeile gewechselt und Arbeitsschritt (a) und (e) wiederholt, bis alle Textzeilen der Seite überprüft und ggfs. korrigiert worden sind

6. **Korrektur der Reading Order:** Nach der Korrektur von Textregionen, Text lines und Baselines muss die Reading Order (Lesereihenfolge) der Regionen und Zeilen überprüft werden
   a. in der Toolbar das “Auge”-Symbol anklicken und die Option “Show regions reading order” auswählen

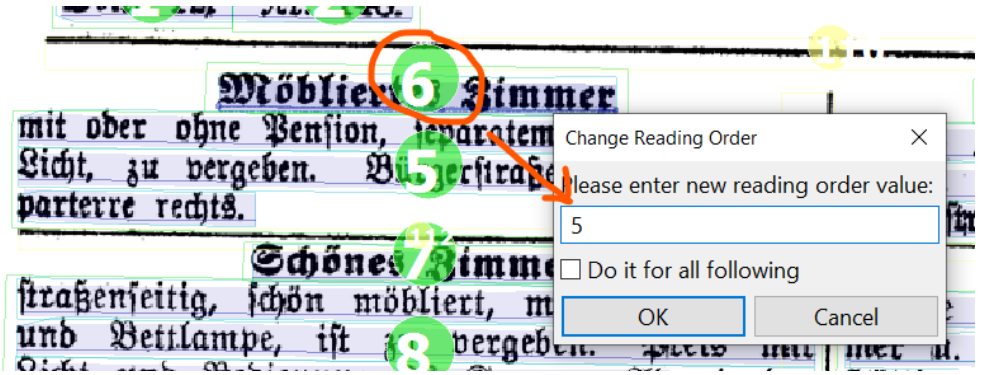


   b. überprüfen, ob die Reihenfolge der Textregionen der natürlichen Lesereihenfolge aller Textabschnitte der Seite entspricht; Die Zahlen in den grünen Kreisen geben die Abfolge der Textregionen an



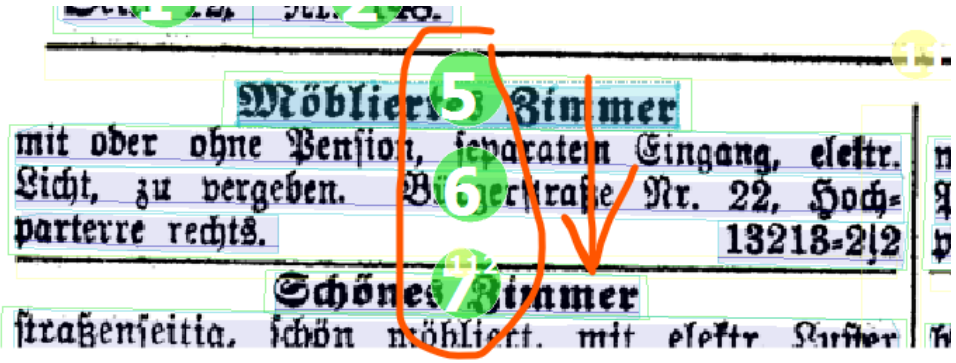
3-spaltige Beispielseite. Diese Reading-Order beginnt in de Kopfzeile (header) und verläuft von links nach rechts, um anschließend spaltenweise von oben nach unten zu verlaufen. Die Reading Order bildet somit die gewöhnliche Lesereihenfolge einer Zeitungsseite ab

c. Stimmt die Reading Order nicht, kann die Reihenfolge der Regionen verändert werden, indem auf die Zahl im grünen Kreis der jeweiligen Region geklickt wird
d. im Pop-Up-Fenster wird die korrekte Stelle der ausgewählten Region angegeben:



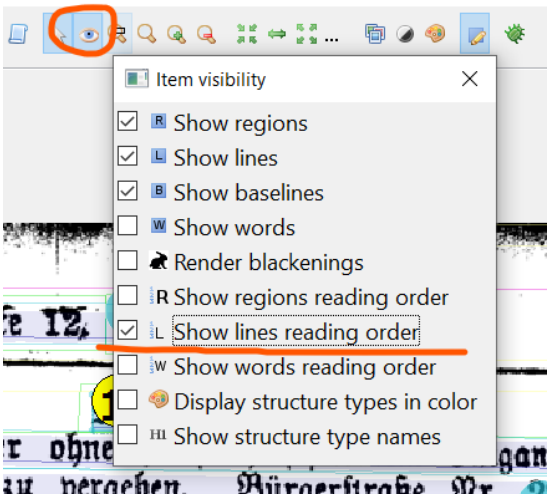
im Beispiel ist die Reihenfolge der Region 6 und 5 vertauscht und damit fehlerhaft

e. “OK” klicken und alle weiteren Regionen auf korrekte Reihenfolge überprüfen

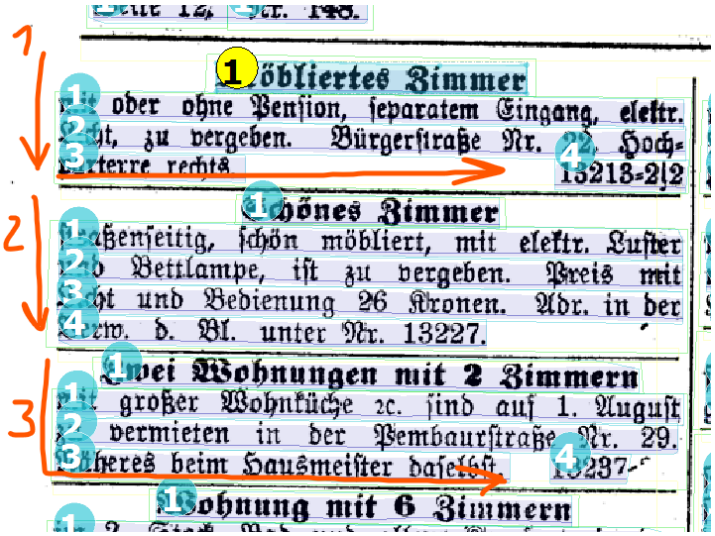


Nach Anpassung: korrekte Reading Order

f. Die **Reading Order der Textlines** wird analog zur Korrektur der Reading Order der Regionen durchgeführt
g. in der Toolbar auf das “Auge”-Symbol klicken und “Show line reading order” auswählen



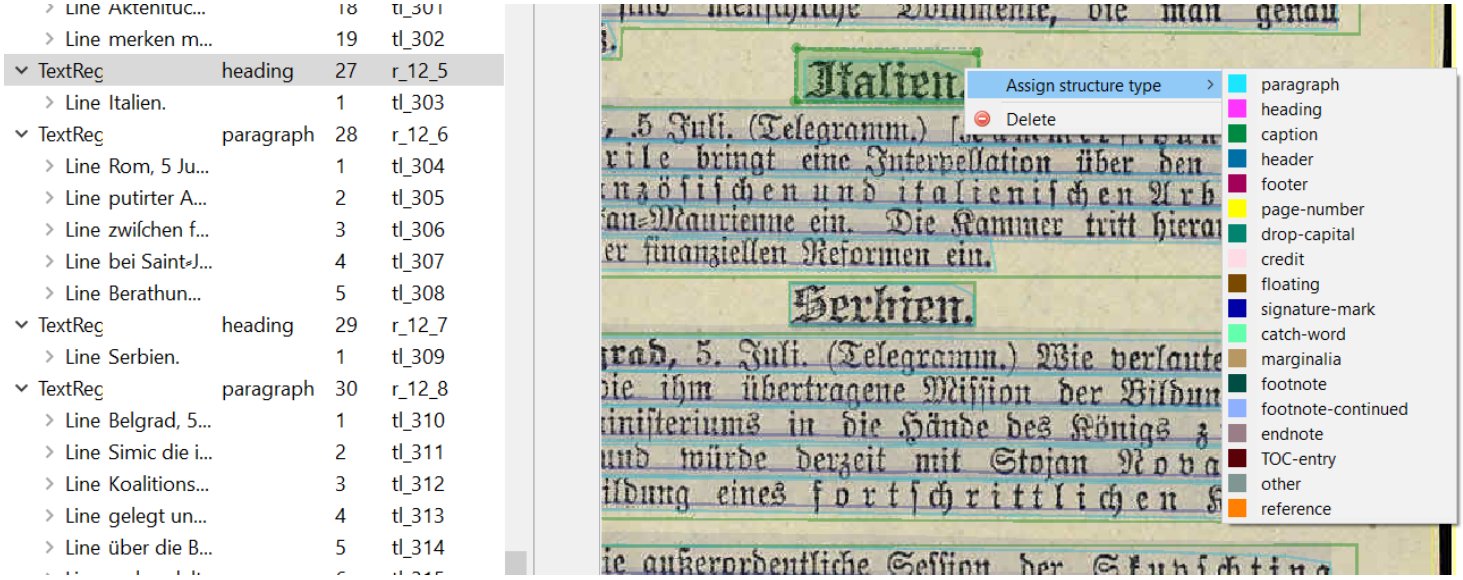
h. Überprüfen, ob die Lesereihenfolge der Textzeilen pro Textregion stimmt. Die Reihenfolge der Zeilen wird nicht für die gesamte Seite gezählt, sondern pro Textregion. D.h. alle Textzeilen, die zu einer Textregion gehören, müssen in der korrekten Reihenfolge angegeben sein:



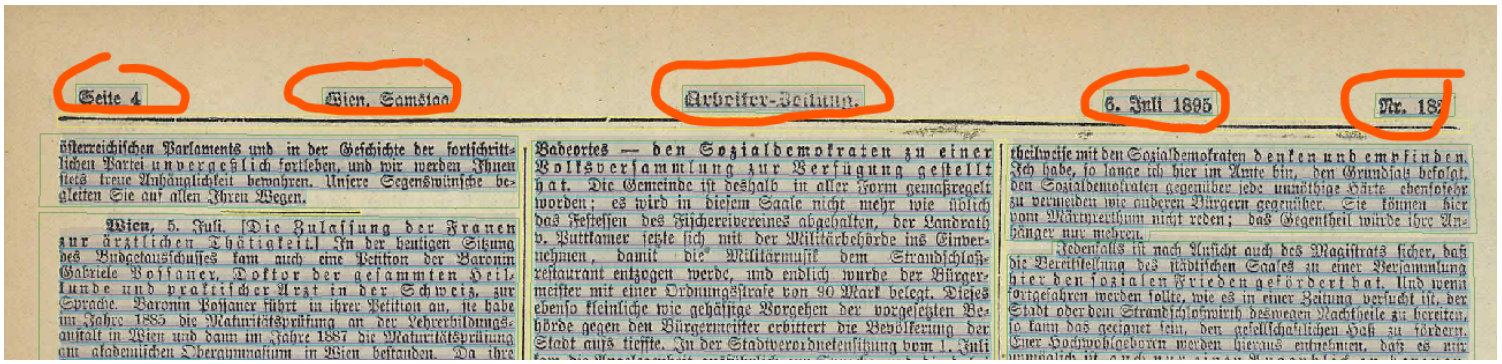
3. Tagging von Regionen und Textlines

Dieser Schritt erfolgt **nach** der [Layoutkorrektur](#).

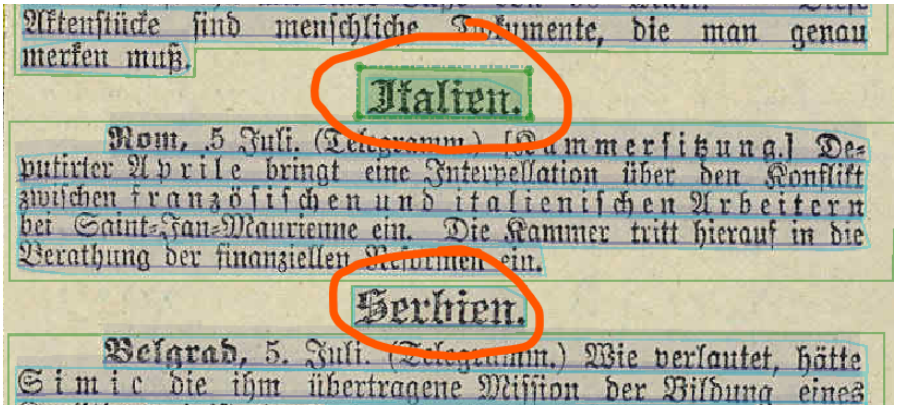
1. Ausgewählten Textregionen und Textlines kann per Rechtsklick ein “Strukturtyp” zugewiesen werden, d.h. sie können getaggt werden:



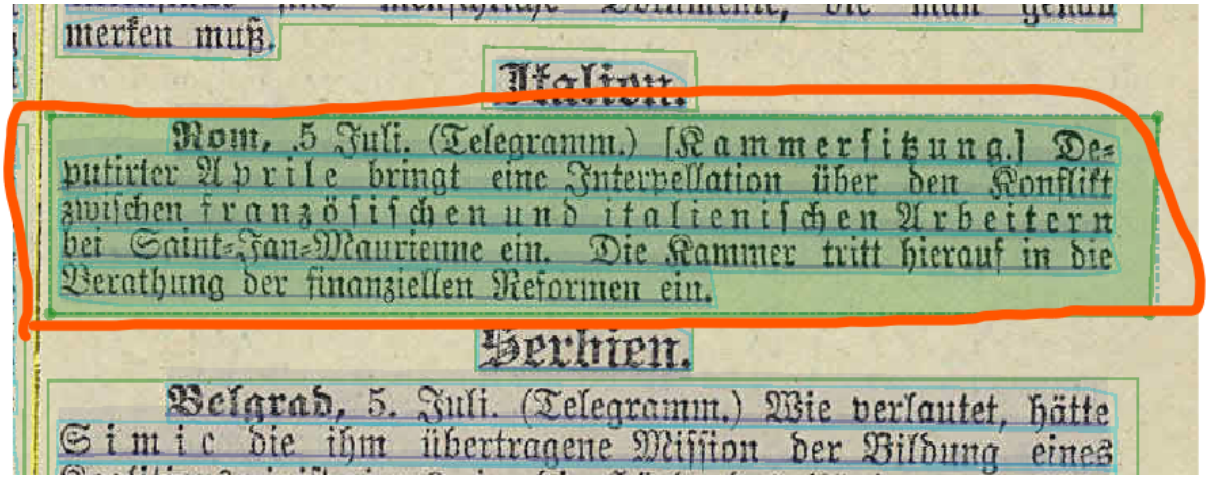
2. Folgende Strukturtypen werden erfasst:
   a. **header** (auf Regionen-Ebene)**:** Alle im Kopf der jeweiligen Seite befindlichen und von einem horizontalen Trennstrich vom Seiteninhalt getrennten Inhalte



   b. **heading** (auf Regionen-Ebene): (Artikel-)Überschriften



c. **paragraph** (auf Regionen-Ebene): Absätze



merken muß.

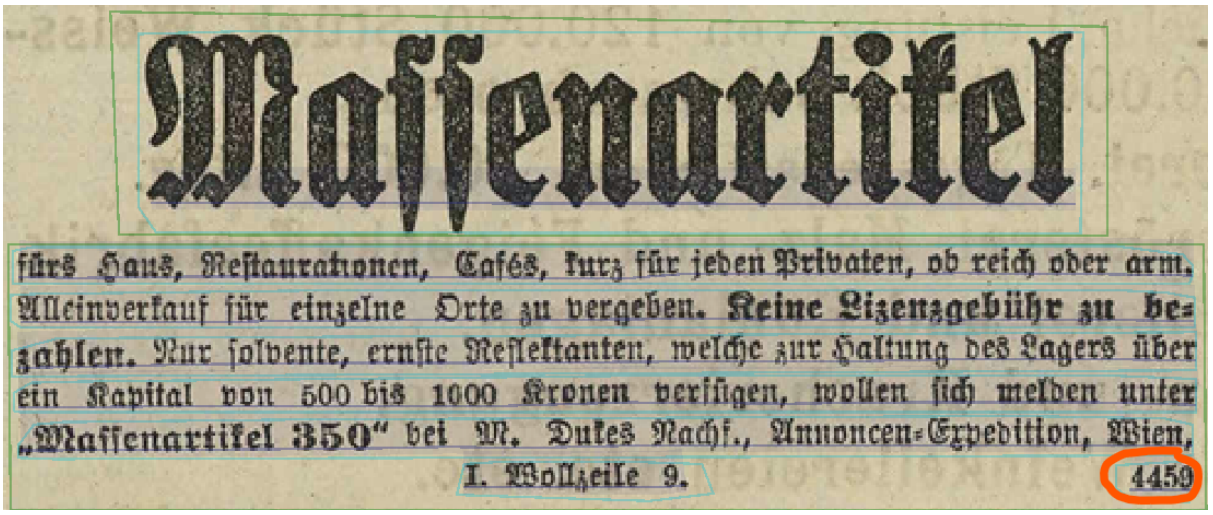
Italien.

Rom, 5 Juli. (Telegramm.) [Kammersitzung.] Deputirter Arile bringt eine Interpellation über den Konflikt zwischen französischen und italienischen Arbeitern bei Saint-Jan-Maurienne ein. Die Kammer tritt hierauf in die Berathung der finanziellen Reformen ein.

Serbien.

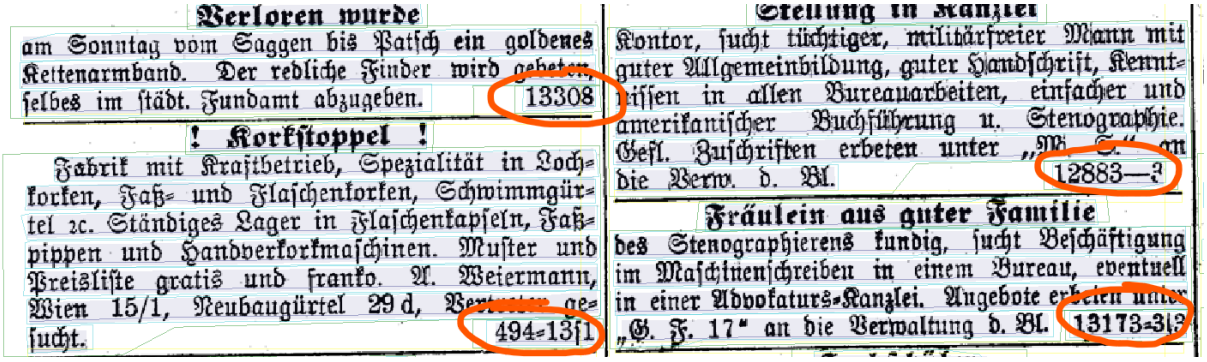
Belgrad, 5. Juli. (Telegramm.) Wie verlautet, hätte Simic die ihm übertragene Mission der Bildung eines

d. **reference** (auf Regionen- oder Textline-Ebene): Referenznummern / Verweise. References können sowohl au



Massenartikel

fürs Haus, Restaurationen, Cafés, kurz für jeden Privaten, ob reich oder arm. Alleinverkauf für einzelne Orte zu vergeben. **Keine Lizenzgebühr zu bezahlen.** Nur solvente, ernste Reflektanten, welche zur Haltung des Lagers über ein Kapital von 500 bis 1000 Kronen verfügen, wollen sich melden unter „Massenartikel 350" bei M. Dukes Nachf., Annoncen-Expedition, Wien, I. Wollzeile 9. 4459



Verloren wurde

am Sonntag vom Saggen bis Patsch ein goldenes Kettenarmband. Der redliche Finder wird gebeten, selbes im städt. Fundamt abzugeben. 13308

! Korkstoppel !

Fabrik mit Kraftbetrieb, Spezialität in Lochkorken, Faß- und Flaschenkorken, Schwimmgürtel 2c. Ständiges Lager in Flaschenkapseln, Faßpippen und Handverkorkmaschinen. Muster und Preisliste gratis und franko. A. Weiermann, Wien 15/1, Neubaugürtel 29 d, Vertreter gesucht. 494-13f1

Stellung in Kanzlei

Kontor, sucht tüchtiger, militärfreier Mann mit guter Allgemeinbildung, guter Handschrift, Kenntnissen in allen Bureauarbeiten, einfacher und amerikanischer Buchführung u. Stenographie. Gefl. Zuschriften erbeten unter „M. S." an die Verw. d. Bl. 12883—3

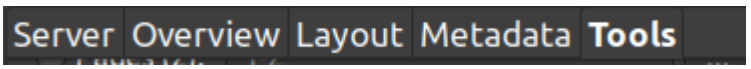
Fräulein aus guter Familie

des Stenographierens kundig, sucht Beschäftigung im Maschinenschreiben in einem Bureau, eventuell in einer Advokaturs-Kanzlei. Angebote erbeten unter „G. F. 17" an die Verwaltung d. Bl. 13173-3

4. Zwischenschritt: ReOCR der Textinhalte

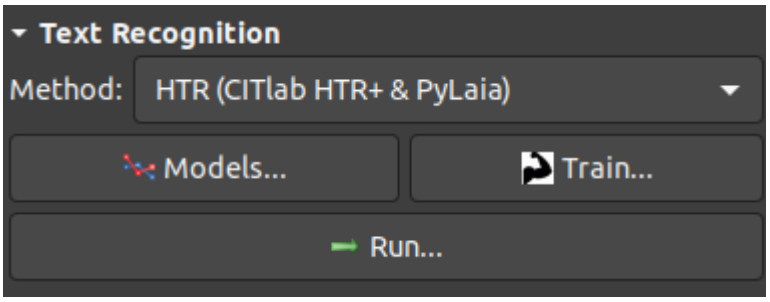
Dieser Schritt erfolgt **nach** der manuellen **Layoutkorrektur** und **vor** der **Zeichenfehlerkorrektur**. Mit einem neu trainierten Texterkennungs-Modell, basierend auf den bereits existierenden Transkriptionen, soll nun die Zeichenfehlerkennungsrate und somit auch die benötigte Bearbeitungszeit reduziert werden.

1. Auswahl des Reiters “Tools”

   Die Reiter-Auswahl befindet sich in dem Experten-Client von Transkribus auf der linken Seite des Hauptfensters.

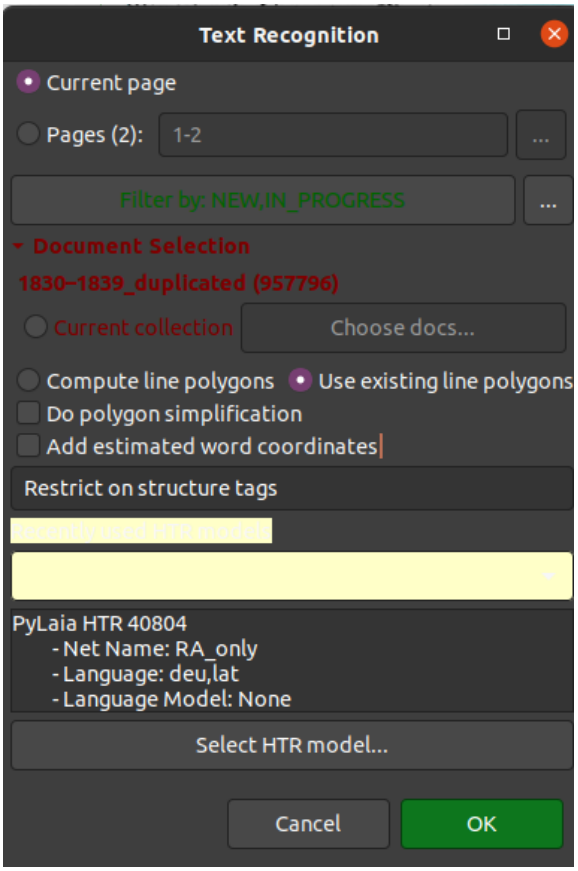
2. Text Recognition Tools

   Das Text Recognition Tool bietet mehrere ausführbare Aktionen:
   - “Method”: Auswahl des OCR-Verfahrens
   - “Models”: Übersicht aller verfügbarer Modelle für die Texterkennung
   - “Train” : Training eines OCR-Modells (an dieser Stelle nicht von Interesse)
   - “Run”: Starten der Texterkennung

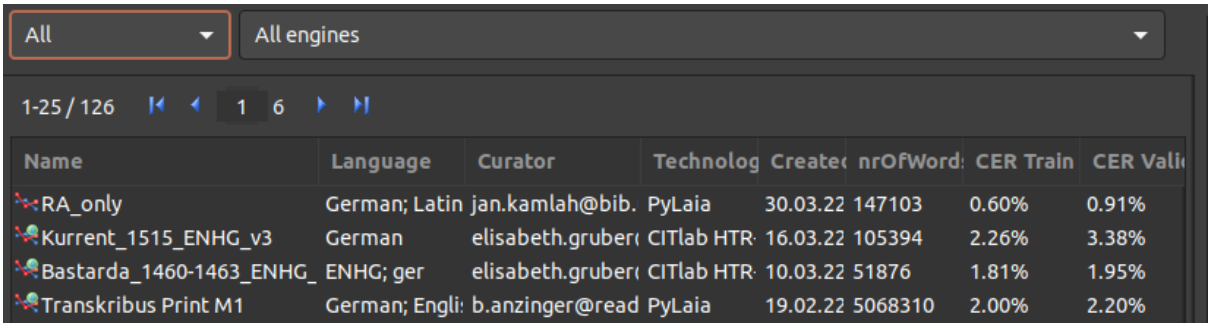
3. Auswahl der Method
   In dem Dropdown “Method” sollte HTR(CITlab HTR+ & PyLaia) ausgewählt sein

4. Parametrisierung und Starten der erneuten Texterkennung -> “Run”
   Nach dem Klicken auf den Button “Run” erscheint ein Fenster mit weiteren Einstellungsmöglichkeiten



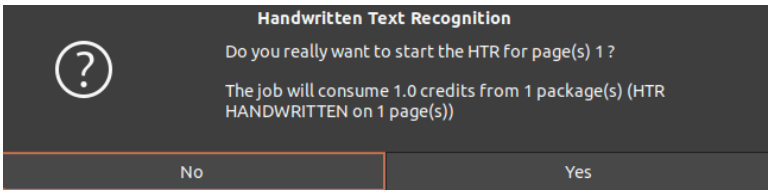
Die Parametrisierung sollte wie in dem Bild gezeigt erfolgen

- Current Page (aktiviert)
- Use existing line polygons (aktiviert)
- Restliche Einstellungsmöglichkeiten (deaktiviert)
- “Select HTR model..” -> Auswahl des Modells “RA_only”
    - Net Name: RA_only

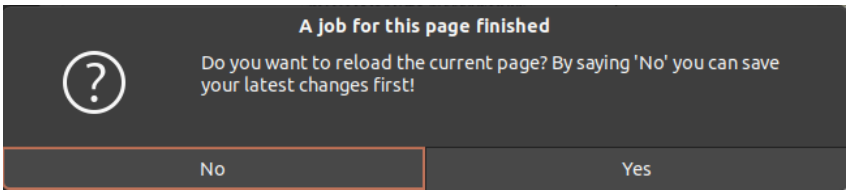


Mit dem grünen Button “OK” wird die Texterkennung gestartet

5. Akzeptieren der Credit-Kosten
   Die Erkennung einer Seite kostet 1.0 credits (ca. 10 Cent) und muss mit “YES” bestätigt werden.



6. Aktualisierung der Transkription
   Die Erkennung der Seite sollte zwischen 1-3 Min. dauern, danach erscheint ein Dialogfenster, welches mit “YES” bestätigt werden sollte.



7. Nun kann mit der Zeichenfehlerkorrektur begonnen werden