Erstellung wissenschaftlich nachnutzbarer Volltexte für Präsentation und Analyse am Beispiel obersorbischer Drucke

WITO BÖHMAK

Serbski Sorbisches institut Institut

ROBERT SACHUNSKY

SLUB Wir führen Wissen. Sächsische Landesbibliothek - Staats- und Universitätsbibliothek Dresden

KAY-MICHAEL WÜRZNER

SLUB Wir führen Wissen. Sächsische Landesbibliothek - Staats- und Universitätsbibliothek Dresden

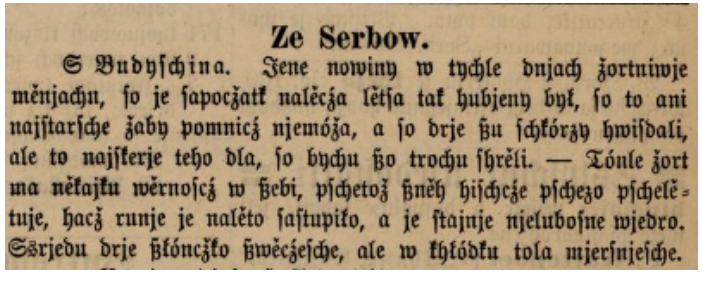
8. Bibliothekskongress Leipzig, 1.6.2022

https://hackmd.io/@bertsky/bibkon22-hsb-si-slub

Wissenschaftlich nachnutzbare Volltexte

- strukturierter Volltext mit digitaler Präsentation
    - u.a. für Wissenschaftler, Heimatforscher, Journalisten, breiten Bildungsbereich
    - u.a. für Search/Retrieval, Textkorpora, hist. Wörterbücher
- Belieferung mit Forschungsdatenrepos (Ground-Truth)
  und Sprachressourcen (Modelle)

Ein Beispiel: Serbske Nowiny 23.3.1878 (überwiegend Fraktur)



Ze Serbow.

S Budyſchina. Jene nowiny w tychle dnjach žortniwje
měnjachu, ſo je ſapočžatk nalěcża lětſa tak hubjeny byl, ſo to ani
najſtarſche žaby pomnicź njemóža, a ſo drje ßu ſchkórzy hwiſdali,
ale to najſkerje teho dla, ſo bychu ßo trochu ſhrěli. — Tónle žort
ma nětajku wěrnoſcź w ßebi, pſchetož ßněh hiſchcże pſchezo pſchelě-
tuje, hacž runje je naléto ſaſtupiko, a je ſtajnje njelubosne wjedro.
Sſrjedu drje ßkónczko ßwěcżeſche, ale w khłódku tola mjerſnjeſche.

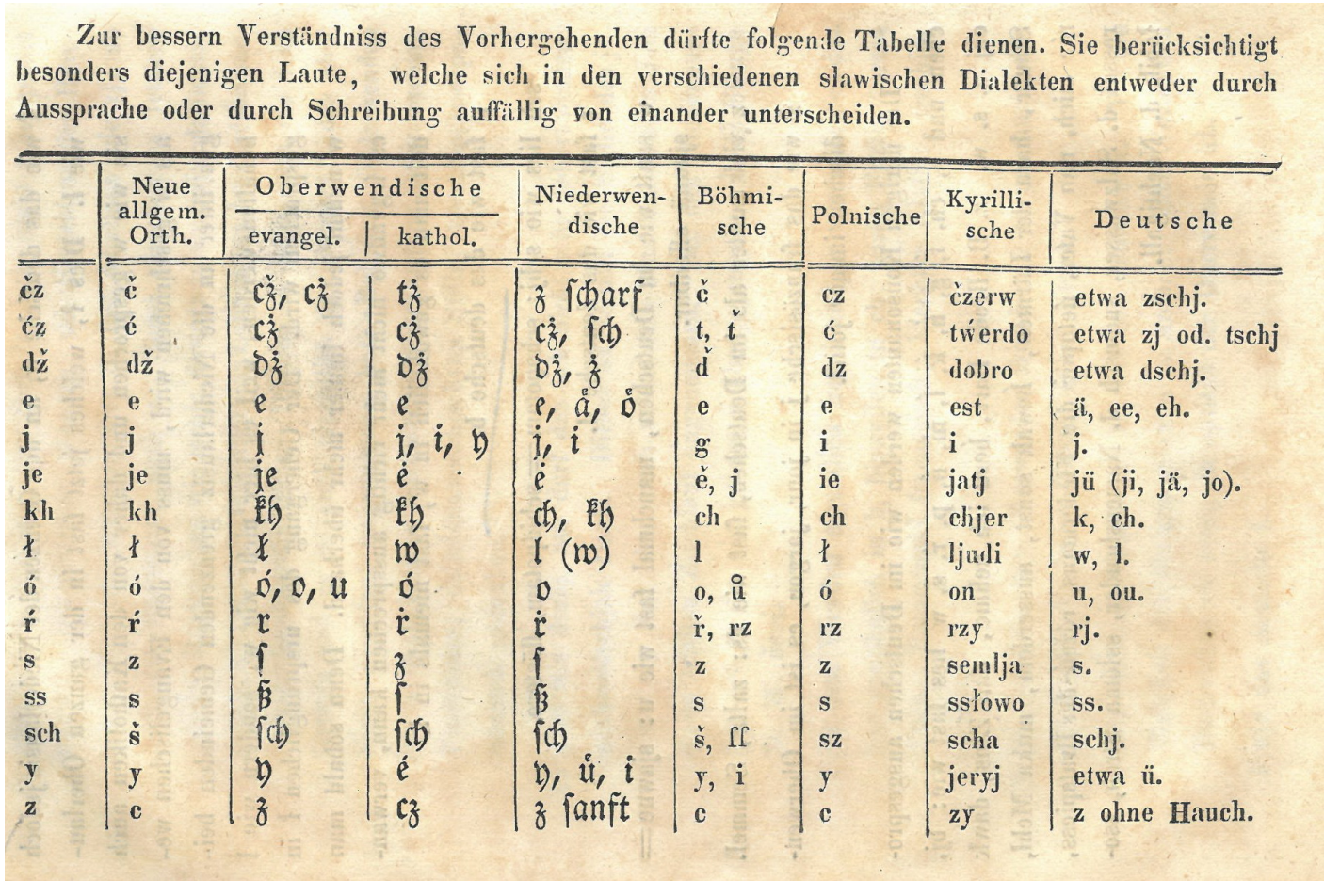
```
▼<TextLine ID="region_0010_line_0001" HEIGHT="66" WIDTH="367" HPOS="2359" VPOS="3326">
  ▼<Shape>
     <Polygon POINTS="2359,3326 2726,3328 2725,3392 2359,3389"/>
   </Shape>
  ▼<String ID="region_0010_line_0001_word0000" HEIGHT="53" WIDTH="74" HPOS="2360" VPOS="3330" CONTENT="Ze">
    ▼<Shape>
       <Polygon POINTS="2360,3330 2434,3330 2434,3383 2360,3383"/>
     </Shape>
   </String>
   <SP/>
  ▼<String ID="region_0010_line_0001_word0001" HEIGHT="55" WIDTH="251" HPOS="2473" VPOS="3329" CONTENT="Serbow.">
    ▼<Shape>
       <Polygon POINTS="2473,3329 2724,3329 2724,3384 2473,3384"/>
     </Shape>
   </String>
 </TextLine>
</TextBlock>
▼<TextBlock ID="region_0005" HEIGHT="702" WIDTH="1586" HPOS="1746" VPOS="3393" TAGREFS="layouttag-paragraph" IDNEXT="region_0002">
```

```
<head><lb/>Ze Serbow.</head>
<p>
<lb/>
S Budyſchina. Jene nowiny w tychle dnjach žortniwje
<lb/>
měnjachu, ſo je ſapočžatk nalěcża lětſa tak hubjeny był, ſo to ani
<lb/>
najſtarſche žaby pomnicź njemóža, a ſo drje ßu ſchkórzy hwiſdali,
<lb/>
ale to najſkerje teho dla, ſo bychu ßo trochu ſhrěli...
</p>
```

Ausgangssituation

- Digitalisierungen am SI seit 15 Jahren
- Landesdigitalisierungsprogramm (LDP) Sachsen,
  koordiniert am Dresdner Digitalisierungszentrum der SLUB
- 2016-2017 im Rahmen des LDP Digitalisierung historischen Schrifttums:
    - Antiqua – ungenügende Volltext-Qualität
    - Fraktur – keine Volltext-Erzeugung
- besonderes Problem **Fraktur Sorbisch**:
    - slawische Diakritika, bisher keine verlässliche automatische Texterkennung (OCR)
    - mehrere Schreibweisen, Übergänge

Sorbische Druckzeichen Antiqua/Fraktur

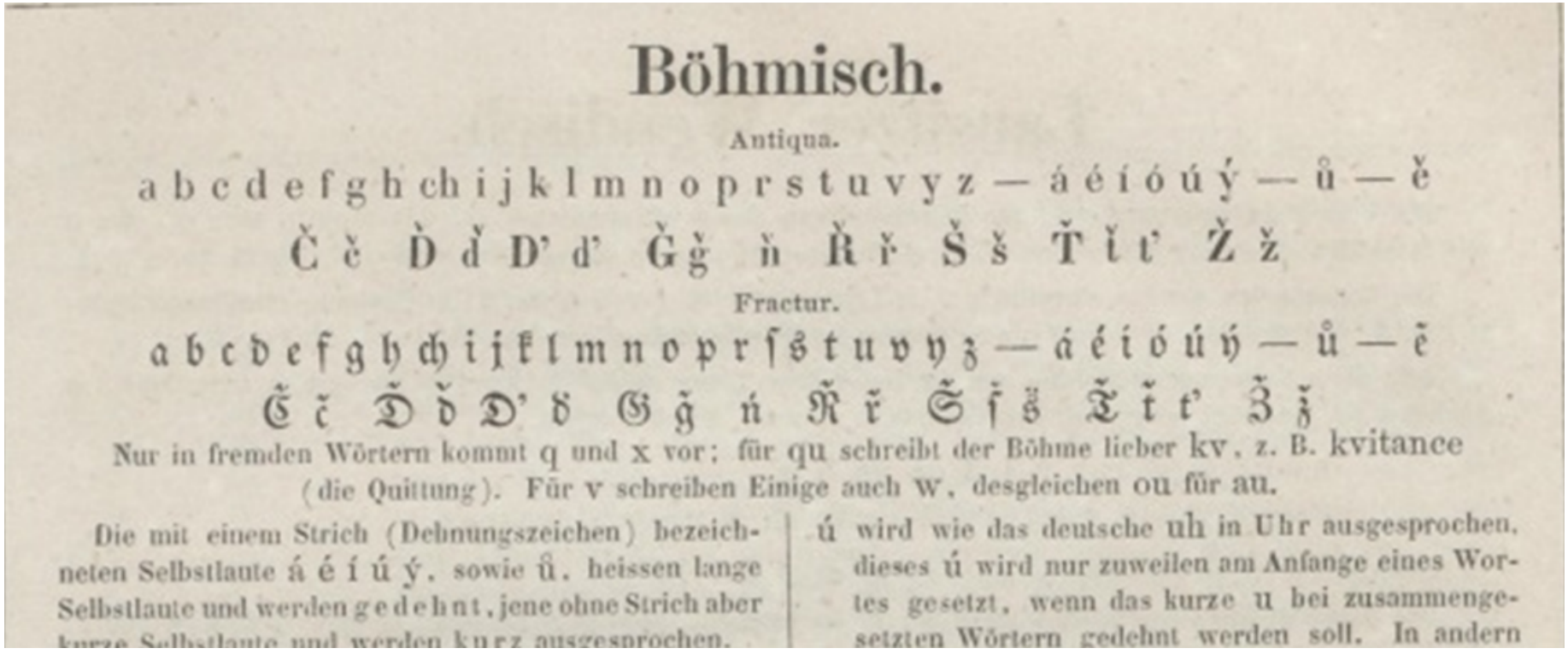


Zur bessern Verständniss des Vorhergehenden dürfte folgende Tabelle dienen. Sie berücksichtigt besonders diejenigen Laute, welche sich in den verschiedenen slawischen Dialekten entweder durch Aussprache oder durch Schreibung auffällig von einander unterscheiden.

	Neue allgem. Orth.	Oberwendische evangel.	Oberwendische kathol.	Niederwendische	Böhmische	Polnische	Kyrillische	Deutsche
čz	č	cž, cz	tz	z scharf	č	cz	čzerw	etwa zschj.
ćz	ć	cz	cz	cz, sch	t, ť	ć	tẃerdo	etwa zj od. tschj
dž	dž	dz	dz	dz, z	d	dz	dobro	etwa dschj.
e	e	e	e	e, ä, ö	e	e	est	ä, ee, eh.
j	j	j	j, i, y	j, i	g	i	i	j.
je	je	je	ė	ė	ě, j	ie	jatj	jü (ji, jä, jo).
kh	kh	kh	kh	ch, kh	ch	ch	chjer	k, ch.
ł	ł	ł	w	l (w)	l	ł	ljudi	w, l.
ó	ó	ó, o, u	ó	o	o, ů	ó	on	u, ou.
ŕ	ŕ	r	ŕ	ŕ	ř, rz	rz	rzy	rj.
s	z	s	z	s	z	z	semlja	s.
ss	s	ß	s	ß	s	s	ssłowo	ss.
sch	š	sch	sch	sch	š, ſſ	sz	scha	schj.
y	y	y	é	y, ů, i	y, i	y	jeryj	etwa ü.
z	c	z	cz	z sanft	c	c	zy	z ohne Hauch.

Jan Arnošt Smoler: Mały Sserb aby Serske a Njemske Rosmłowenja atd. = Wendisch-Deutsche Gespräche nebst einem wendisch-deutschen und deutsch-wendischen Wörterbuche, sowie einem Verzeichnisse von Ortsnamen, einer Darlegung der Aussprache und Orthographie und Zugabe der gebräuchlichen Eidesnormen, Bautzen 1841

Wo noch? – Fraktur und Antiqua Tschechisch



Böhmisch.

Antiqua.

a b c d e f g h ch i j k l m n o p r s t u v y z — á é í ó ú ý — ů — ě

Č č Ď ď D' d' Ǧ ǧ ň Ř ř Š š Ť ť t' Ž ž

Fractur.

a b c d e f g h ch i j k l m n o p r ſ s t u v y z — á é í ó ú ý — ů — ě

Č č Ď ď D' d Ǧ ǧ ń Ř ř Š ſ š Ť ť t' Ž ž

Nur in fremden Wörtern kommt q und x vor; für qu schreibt der Böhme lieber kv, z. B. kvitance (die Quittung). Für v schreiben Einige auch w, desgleichen ou für au.

Die mit einem Strich (Dehnungszeichen) bezeichneten Selbstlaute á é í ú ý, sowie ů, heissen lange Selbstlaute und werden gedehnt, jene ohne Strich aber kurze Selbstlaute und werden kurz ausgesprochen.

ú wird wie das deutsche uh in Uhr ausgesprochen, dieses ú wird nur zuweilen am Anfange eines Wortes gesetzt, wenn das kurze u bei zusammengesetzten Wörtern gedehnt werden soll. In andern

Alphabete orientalischer und occidentalischer Sprachen, zsgest. von Friedrich Ballhorn, 8., verb. Aufl., Leipzig 1859

Erste Studie zu Ground-Truth / Trainingsmodell in 2019

OCR für ressourcenarme Sprachen am Beispiel des Obersorbischen
(Kay-Michael Würzner & Wito Böhmak, Abstract zum Dt. Bibtag 2020 + DFG-Antrag):

- Bericht über manuelle (und iterative) Erstellung von **GT** für Fraktur-Obersorbisch und **Training** eines Tesseract-Modells per Finetuning von `Fraktur`
- Untersuchung mit ABBYY Recognition Server und Tesseract auf ihre **Eignung** für die Erzeugung wissenschaftlich nachnutzbarer Volltexte
- sowohl ABBYY Recognition Server mit dem Modell `Altdeutsch/Gothic` als auch Tesseract mit dem sprachunabhängigen Modell `Fraktur` erzielten **schlechte** Genauigkeit
- Erstellung von GT und eigenem Trainingsmodell `hsb_frak` mit einfachem Workflow: **Machbarkeit**

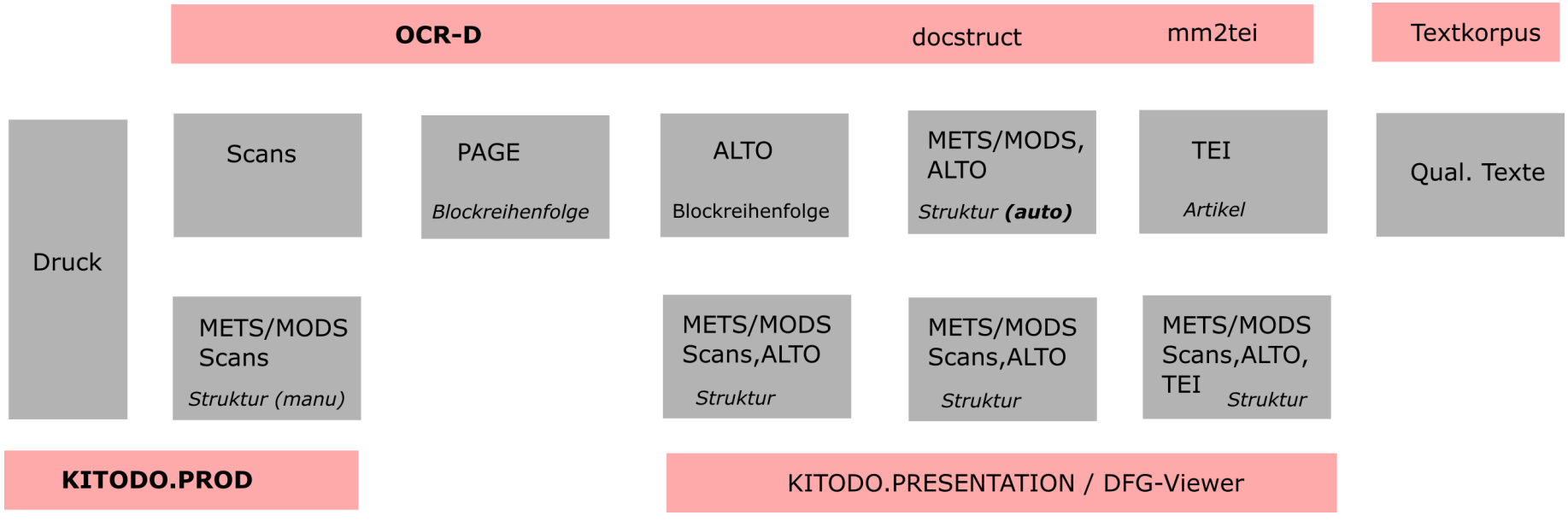
	ABBYY	**Tesseract**	**Tesseract (nachtrainiert)**
Zeichenfehlerrate:	12-17%	8-11%	0,5-3,7%

Entwicklung seit 2020

Verstetigung in anderen Projektkontexten

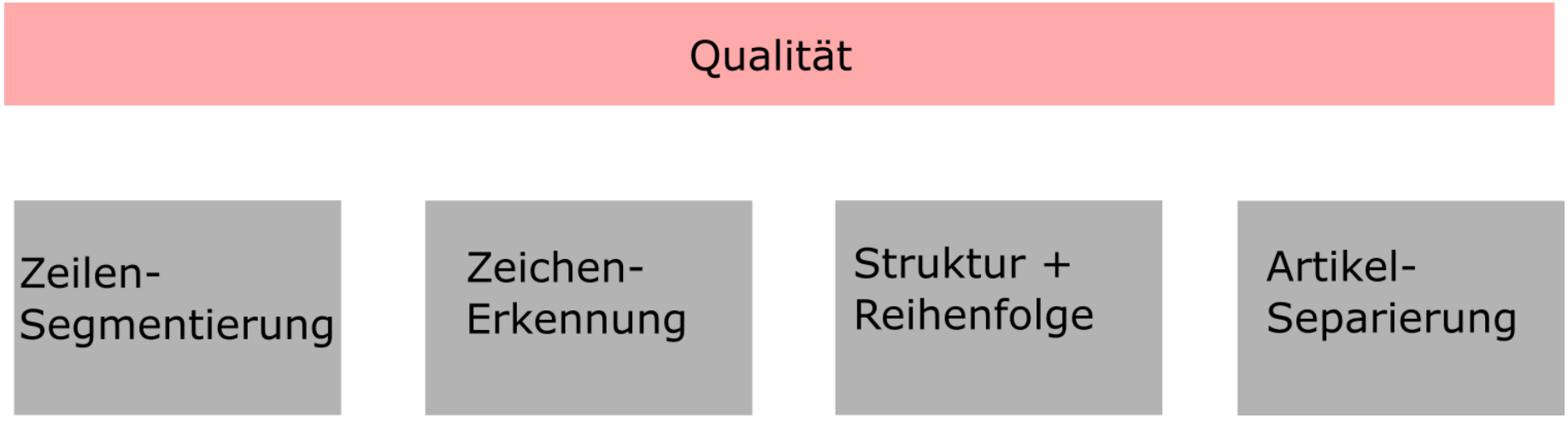
- SI: Retrodigitalisierung und Präsentation auf Basis von Kitodo
- SLUB: Beteiligung bei…
  - Entwicklung Kitodo und DFG-Viewer
  - DFG-Förderinitiative OCR-D
  - DDB Zeitungsportal

Einsetzbarkeit im produktiven Betrieb mit komplexem Workflow



Was macht eine hochqualitative Texterkennung aus?

- *gute Segmentierung*: Ist der Text richtig lokalisiert worden?
  (kein Text verloren, kein Nicht-Text verwechselt)
    - scheinbare/fehlende Wörter oder Zeilen
    - überlappende/abgeschnittene Zeichen
- *gute OCR*: Sind die Zeichen an sich richtig erkannt worden?
- *gute Strukturerkennung*: Sind die Blöcke und Zeilen in der richtigen Reihenfolge? Wurden Überschriften markiert? Wurden Kapitel/Artikel separiert?
- *weitere Analyse*: Schriftauszeichnung, Textnormalisierung



Maßnahmen zur Verbesserung der Textqualität für Obersorbisch

	OLR	OCR	AS	
Bereitstellung von Daten		✔		SI
Entwicklung von Werkzeugen	✔	✔	✔	SLUB / OCR-D
Entwicklung von Modellen		✔		SI / SLUB
Erarbeitung von Workflows	✔	✔	✔	SI / SLUB

OLR: Optische Layouterkennung (Segmentierung)
OCR: Optische Zeichenerkennung
AS: Artikelseparierung (Strukturerkennung)

Vorgehen bei Segmentierung

- Nutzung der Prozessoren aus OCR-D (hier `eynollah` und `ocrd_segment`)
- Erkennung von:
    - Seitenrändern
    - Linien und Ornamentierung
    - Blocksegmentierung und Lesereihenfolge
    - Zeilensegmentierung
- Adäquate Evaluierung: für Aussage über Gesamtqualität der Texterkennung ist **Anteil des Segmentierungsfehlers** mit zu betrachten!

Vorgehen bei Zeichenerkennung

- Nutzung der Werkzeuge und Workflows von OCR-D
- Nutzung von Multi-OCR-Alignierung zur weiteren Verbesserung der OCR-Qualität (z.B. Diplopie-Problem bei Tesseract)
- Training von eigenen Modellen für Obersorbisch (Fraktur + Antiqua), aufbauend auf Community-Modellen (Tesseract / Calamari OCR)
- Erstellung von GT-Material (eigenes Repository für `hsb`)
    - Transkription mit Aletheia
    - Transkription mit Larex
- iteratives Vorgehen Prozessierung – GT-Erstellung – Training

Iteratives Vorgehen

1. Verbesserung der OCR-D-Werkzeuge und Workflows
   (z.B. Binarisierung SBB)
2. Nutzung existierender Modelle für die OCR
3. Transkription von GT-Material
   (Level 2 nach Richtlinien OCR-D / DFG)
   - neue Seiten
   - Finden und Beseitigen von Transkriptionsfehlern
4. Neutrainieren der OCR-Modelle,
5. Wiederholung und stetige Evaluierung an ausgewähltem GT-Material

Training und Evaluierung

- Umfang der GT-Daten
  - Fraktur: > 16.000 Zeilen (1843-1911)
  - Antiqua: > 16.000 Zeilen (1880-1934, 1950-)
- Training von Modellen für Tesseract / Calamari …
- Zeichenfehlerraten (CER) auf Test-GT:

CER OCR	Abbyy Srv14*	Tesseract	hsb (Tess)	hsb (Cala)	hsb (multi)
Fraktur**	14,72%	9,01%	0.56%	0.45%	0.37%
Antiqua	***	2.17%	0.89%	0.52%	0.48%

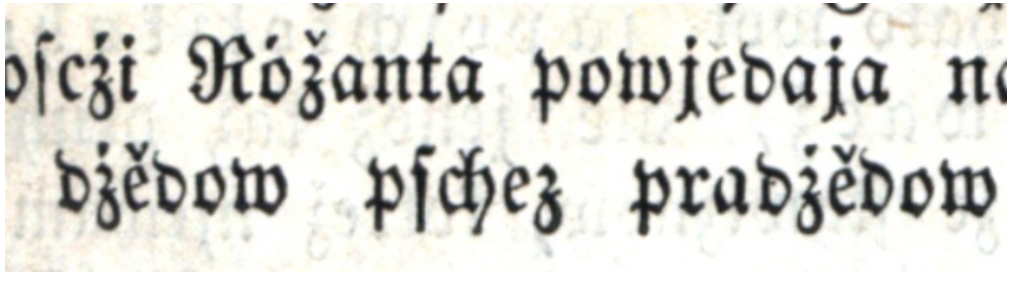
* Abbyy repräsentiert ſ als s (ca. 50-80% Fehleranteil)

** v.a. Überpunkt vs. Akut

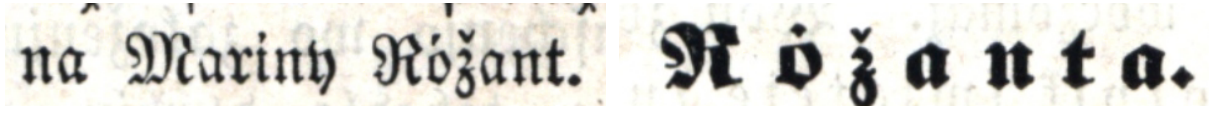
*** Stichproben mit Abbyy: bei Korrektur von ĕ (e breve) nach ě (e caron): < 1%

Herausforderungen bei GT-Erstellung

- Lücken in Transkriptionsrichtlinien
  z.B. **"** vs. **“** oder **—** vs. **–** oder **⸗** vs. **-**
- zu wenige Beispiele für große Fonts und für spezifische Zeichen
- Wandel historischer Schreibweisen
  z.B. am Übergang von Überpunkt zu Akut – Erkennung ob
    - “Druckschwäche”:



    - “gewollt” oder “beschränkter Drucksatz”:

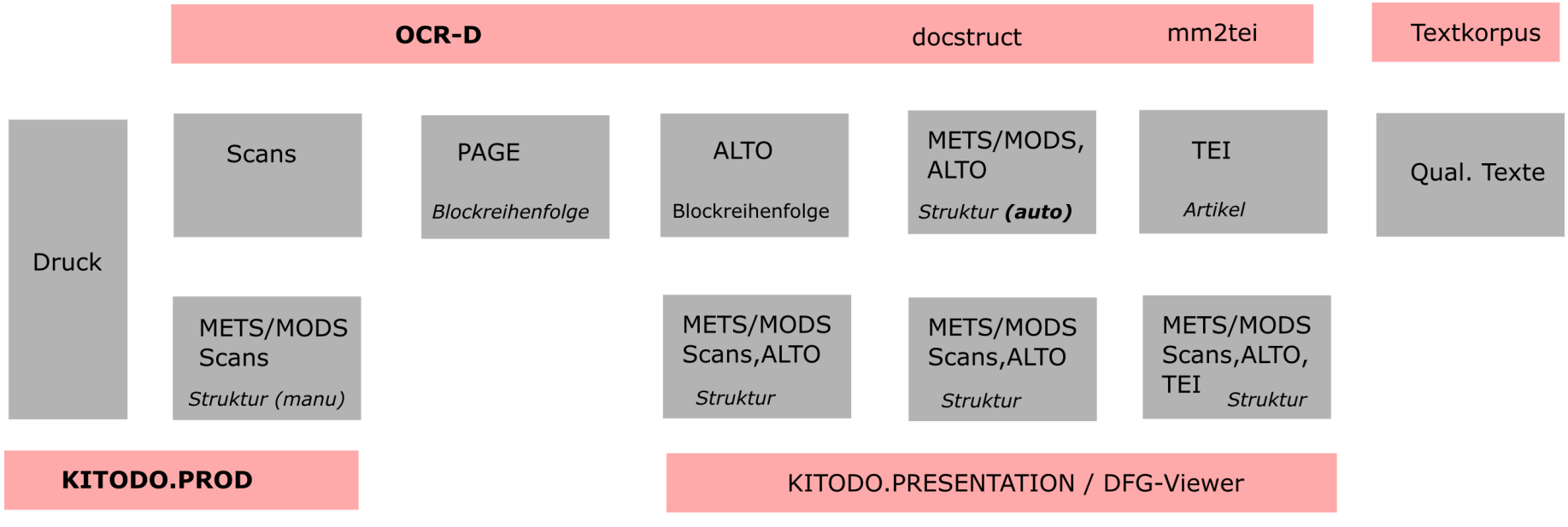


Strukturerkennung – warum?

Ziel ist Artikelseparierung

- Erstellung einer qual. hochwertigen Präsentation
- Artikelextraktion (in TEI) für Textkorpora
- siehe Masterplan Zeitungsdigitalisierung, Stufe 3a Artikelseparierung

Unser Workflow:



Strukturerkennung – was?

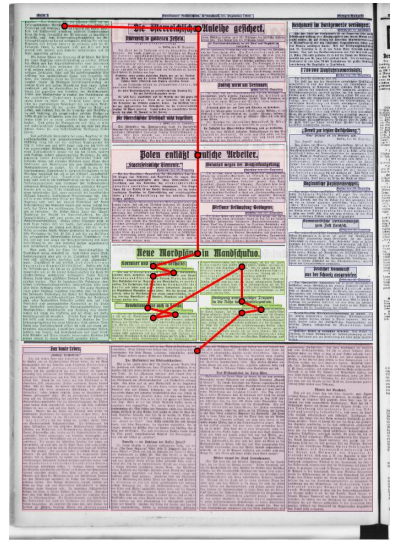
- Seitenstruktur (Segmentierung + Reihenfolge): PAGE oder ALTO
- Dokumentstruktur (Titel, Kapitel, Abschnitt, Legende, Seitenzahl, Artikel): Repräsentation…
    - nach **DFG-Anwendungsprofil** für METS/ALTO:
      `mets:div/@LABEL` + `mets:structLink` (mit ALTO-fileGrp)
      → ganze Seiten
    - nach **Europeana-Profil** für METS/ALTO:
      `mets:div/mets:fptr` (ALTO-fileID) `./mets:area` (ALTO-IDREF)
      → 1…n Segmente unterhalb+oberhalb von Seiten

Anwendungsbeispiel: Serbske Nowiny 23.3.1878 (DFG-Viewer)

Strukturerkennung – wie?

- automatische Strukturerkennung:
  - regelbasiert:
    vorherige seitenweise Textblock-Lesereihenfolge +
    heuristische (visuell-textuelle) Artikelseparierung
  - datengetrieben:
    neuronale Netze mit visuell-textueller Eingabe

- Unterstützung durch Präsentation:
  - Kitodo/DFG-Viewer: bisher nur `mets:structLink`
  - TEI-Konvertierung: bisher nur `mets:structLink`
  - kommerzielle Systeme: `mets:area` / beides
  - DDB-Zeitungsportal: noch offen



(Bsp.: Structify)

Fragen?

Vielen Dank für Ihre Aufmerksamkeit!

- wito.bejmak [[at]] serbski-institut.de
- robert.sachunsky [[at]] slub-dresden.de

https://hackmd.io/@bertsky/bibkon22-hsb-si-slub