Universität Stuttgart • Institut für Literaturwissenschaft • Neuere Deutsche Literatur I Peggy Bockwinkel, M.A. • Keplerstr. 17 • Stockwerk 2b, Zi. 2.055 • 70174 Stuttgart Tel.: +49 (0)711-685-82279 • peggy.bockwinkel@ilw.uni-stuttgart.de

Dokumentation: Dilan Çakır

Finanzierung: www.uni-stuttgart.de/dda + QSM

Anleitung: Erstellen eines Korpus

Schritt	Hinweise und Lösungsvorschläge
1. Buch/ Text scannen	• Buch/ Text sollte wenig Flecken/ Knicke oder Risse haben • mit mind. 400 dpi und Vollfarbe scannen → das ist die Einstellung, mit der „Abbyy“ (für Frakturtexte) die besten Ergebnisse (d.h. die wenigstens Fehlern) liefert (→ s. Datei: **„01-Dokumentation_Vorarbeiten_Frakturerkennung** und im Detail: **Dokumentation_Abbyy_Krueger“**) • (am UB-Buchscanner kann man lediglich über die Farbe der Scans entscheiden. UB-Buchscanner scannt nur mit 300 dpi) • wenn möglich, keine Doppelseiten scannen, sondern jede Seite für sich (das ist später bei der Umwandlung mit „Abbyy“ praktischer) →Zum Schneiden der Seiten stehen aber auch Tools zur Verfügung (hierzu siehe Anleitung **„02-Anleitung Abbyy FineReader_OCR von Frakturtexten“**) • Wenn möglich Ausgaben scannen, die keine Zeilenummerierung enthalten, da es im Nachhinein schwer ist diese wegzuschneiden. Außerdem werden Bindestriche beim OCRn nicht verbunden, wenn zwischen drin eine Zahl steht.
2. Buch/ Text in OCR umwandeln	• separate Anleitung für das Programm „Abbyy“ vorhanden (inkl. Anleitung wie man Doppelseiten mit Hilfe eines Tools trennen kann → s. PDF „Anleitung Abbyy FineReader_OCR von Frakturtexten“) • zum Vergleich, ob sich die Umwandlung mit „Abbyy“ für ein bestimmtes Buch/Text lohnt, kann ein Vergleich mit „Double-Keying“ aufschlussreich sein (wenn der Text in einem guten Zustand ist, sollte die Umwandlung mit „Abbyy“ wesentlich schneller gehen als mit der „Double-Keying“-Methode).

	• evntl. lohnt es sich auch die Umwandlung mit „Adobe“ durchzuführen. Eine Dokumentation des Vergleichs „Abbyy vs. Adobe“ siehe im Ordner „Abbyy vs. Adobe“.
3. Texte sammeln **(Alternative zu Punkt 1 – 2)**	• hierzu stehen verschiedene Quellen zur Verfügung (TextGrid, Das Deutsche Textarchiv, Gutenberg,…) • **Textgrid** → Texte können nicht in .txt-Format runtergeladen werden • **Gutenberg** → Texte können nicht runtergeladen werden, sondern müssen im Browser markiert + in einen Editor kopiert werden • **Kindle** (Amazon) Florian Barth lädt hier Texte runter
4. Korpus sortieren	• für eine einheitliche Benennung der Dateien sorgen → Hierzu gibt es Mehrfach-umbenennen-Tools (separate Anleitung vorhanden → PDF: **„03-Anleitung Total Commander_mehrfach-umbenennen-Tool“**) • hilfreich ist auch die Sortierung der Dateien in Ordnern, die jeweils beispielsweise 50-Jahre umfassen • Datenbank mit Excel erstellen [in Bearbeitung (Stand April 2016)]
5. Korpus normalisieren	• für eine einheitliche Kodierung sorgen (z.B. UTF-8) • für eine einheitliche Rechtschreibung sorgen • Programme zur Normalisierung (Rechtschreibung) von Texten: → CAB (Anjas Übersicht) →TICCL → VARD2 → CorA → selbst geschrieben von Computerlinguistin (Sarah Schulz/ Uni Stuttgart) Dokumentation der Probleme und Vorteile s. in den Dateien „**S.Schulz (Normalisierungstool) – Dokumentation**“ und „**15.16.17.18.Jhd._was wurde normalisiert _Sarah Schulz' Tool –Dilan**“ und auch „**PROBLEME_Schulz' tool (Tabelle Endversion)Dilan**“)

Universität Stuttgart • Institut für Literaturwissenschaft • Neuere Deutsche Literatur I Peggy Bockwinkel, M.A. • Keplerstr. 17 • Stockwerk 2b, Zi. 2.055 • 70174 Stuttgart Tel.: +49 (0)711-685-82279 • peggy.bockwinkel@ilw.uni-stuttgart.de

Dokumentation: Dilan Çakır

Finanzierung: www.uni-stuttgart.de/dda + QSM

	• weitere Punkte folgen ….