Es soll ein Korpus mit verschiedenen Texten mehrerer Autoren erstellt werden, welches zu Zwecken der Autorenanalyse aufbereitet/annotiert wird. Hauptsächlich wird es sich um WhatsApp-Daten (.txt) handeln.

Im Groben sollte es wie folgt aussehen:

- Das System sollte im weiteren Verlauf erweiterbar und anpassbar sein, um auf spätere Differenzierungen etc. eingehen zu können.
- Es dürfen weder die Daten ins Internet geladen noch direkter Internetzugriff benötigt werden, also rein lokal gearbeitet werden, bspw. über eine lokale Browseranwendung.
- Vergleichbar mit ANNIS (https://korpling.github.io/ANNIS/4/user-guide/index.html) mit zusätzlichen Annotationsebenen (s. unten), also:
  - Open-source oder einfach/modular erweiterbar
  - Datenbank getrennt von der Anwendung
  - Datensatz nicht nur in entwickelter Anwendung nutzbar, sondern auch exportierbar sein, um mit bestehender Korpus-Software bearbeitbar zu sein
  - multi-layer tagging/annotation
  - für einzelne Bestandteile gibt es bereits funktionierende open-Source-Lösungen, welche kombiniert und erweitert werden können.

Jeder Textteil muss klar dem ursprünglichen Autor zugewiesen werden können und mit folgenden Informationen verknüpft sein (idealerweise automatisch aus Fragebogen extrahiert):

- Autor
  - ID
  - Geschlecht
  - Muttersprache
  - Zweit-/Fremdsprachen
  - regionale Zugehörigkeit
  - Alter
  - Beruf
  - …
- Textsorte: WhatsApp / E-Mail / Brief / …
  - Textproduktion: Textverarbeitung / mobiles Endgerät (Touchscreen) / haptische Tastatur / …
- Autokorrektur: eingeschaltet / ausgeschaltet / Swype-Tastatur
- Textfunktion
- Gesprächspartner
  - Familie/Freunde/…
  - Grad der Vertrautheit
  - …
- Zeitpunkt der Textproduktion
- …

Die Annotation jedes Textes erfolgt auf mehreren Ebenen (* automatisch mit manueller Kontrolle):

- Texterfassung
  - Originaltext
  - korrigierter Text (an deutsche Rechtschreibung & Grammatikregeln angepasst)
  - Wortform*
  - Lexem*
    - einfaches Lexem (Buch)
    - Kompositum (Buchreihe)

    - Grammatik*
    - Satzstruktur*
    - => daraus resultierende Verknüpfungsmöglichkeit von bspw. „habe“ und „hab“ als Varianten von „haben - 1. Prs SG als Hilfsverb“ oder „haben - 1. Prs SG als Vollverb“
- Befunderhebung
    - Orthografiefehler
        - Graphemauslassung
            - a
            - b
        - Graphemhinzufügung
            - a
            - b
        - Groß-/Kleinschreibung
            - Groß- statt Kleinschreibung
            - Klein- statt Großschreibung
        - Getrennt-/Zusammenschreibung
            - Getrennt- statt Zusammenschreibung (bspw. bei *Buch Reihe* statt *Buchreihe*)
            - Zusammen - statt Getrenntschreibung
    - Interpunktion
    - …
- Stilistische Merkmale
    - Emoji-Nutzung
    - Wortwahl
    - …

Die verschiedenen Ebenen sollten übergreifend auswertbar sein, bspw.:

- Frequenzen/Varianz innerhalb
    - eines Autors
        - für einen Text
        - für mehrere anhand der anderen Merkmale definierter Texte
        - alle Texte
    - bestimmter gebündelter Autoren, bspw. alle Autoren einer bestimmten Altersgruppe/Region
- Statistiken zu einzelnen Kombinationen, bspw.:
    - Häufigkeit der Graphemauslassung bei…
        - Hilfsverben
        - bestimmten Lexemen
        - bestimmten zugrundeliegenden Wortarten oder -formen
    - Häufigkeit der Bündelung bestimmter Merkmale innerhalb eines Textes / Autors
    - Varianz innerhalb eines Autors, also Häufigkeit der verschiedenen Varianten (bspw. *hab/habe* als Hilfsverb) innerhalb eines/aller Texte
    - Häufigkeit der Befunde für bestimmte Lexeme/Wortarten/etc. (Verhältnis korrekter und inkorrekter Varianten), bspw.
        - Getrennt- statt Zusammenschreibung bei Komposita
    - des gesamten Korpus