Computerlinguistik

E03: Linguistische Annotation

INSTITUT FÜR
DIGITAL HUMANITIES
UNIVERSITÄT ZU KÖLN

Sprachdaten – Grundbegriffe

- **Korpus** (neutrum, plur: Korpora):
  - Sammlung sprachlicher Äußerungen (gesprochen oder geschrieben)
  - Bigitalisiert und maschinenlesbar
  - Besteht aus a) Daten selbst, b) Metadaten, c) Annotationen

- **Metadaten** (Daten über Daten):
  - Informationen über Text(korpus) als Ganzes
  - z.B. Zeit & und Ort der Entstehung, Autor:in, Genre, Lizenz, Identifier etc.

- **Annotationen**:
  - Informationen über definierte Regionen des Text(korpus) – Zeichen, Wörter, Absätze, Seiten etc.
  - Manuelle vs. automatische Annotation

Linguistische Annotation – Grundlagen

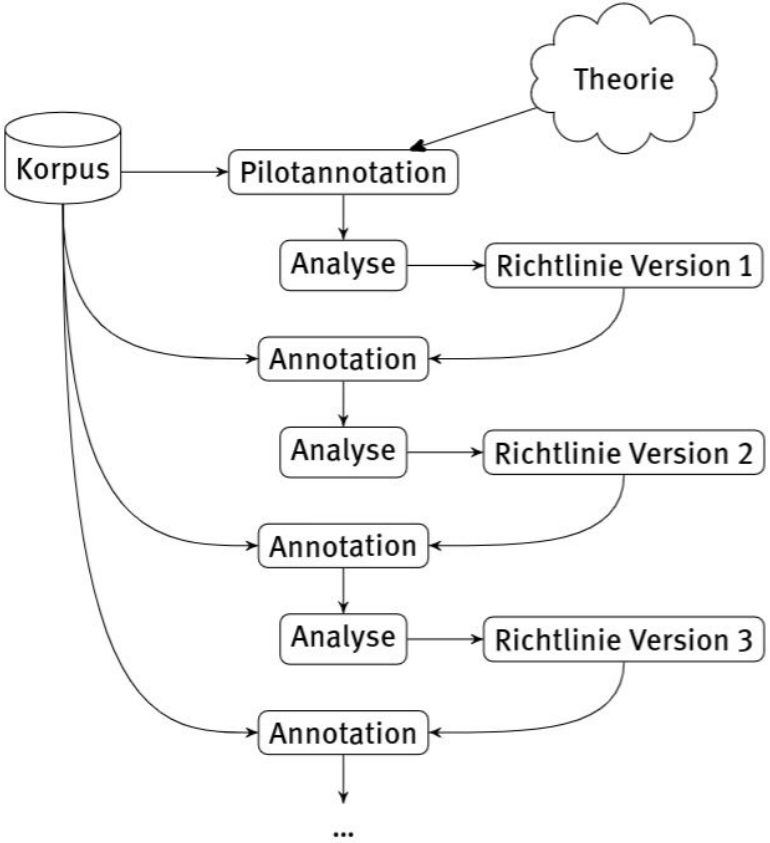
- **Annotation: Auszeichnung** von (Text)Einheiten mit Labels / Verknüpfung von Sprachdaten mit deskriptiven oder analytischen Notationen. Beispiel.
- **Nutzen im (computer)linguistischen Anwendungsbereich:**
  - Überprüfung linguistischer Theorien
  - Training von sprachverarbeitenden Anwendungen
- **Annotationsschema**:
  - Definiert gültige Auszeichnungen / Annotationen (**Labels**), die sprechend gewählt werden sollen, um auffindbar zu sein
  - Definiert mit welchen Annotationseinheiten (**Markables**) die Labels assoziiert werden sollen
  - Beispiel PoS-Tags.

Linguistische Annotation –

- **Richtlinien zur Tokenisierung:** Legen fest, wie Einheiten (Markables, wie z.B. Wörter, Phrasen) aus den Rohdaten (Text, Audioaufzeichnung etc.) gewonnen werden.
- **Richtlinien zur Annotation**: Legen fest, wie bei der Zuweisung von Labels zu Einheiten vorgegangen werden soll (ggfs. iterativ entwickelt).
- **Inter-Annotator-Agreement**: Übereinstimmung von Annotationen zwischen verschiedenen (meist menschlichen) Annotatoren. Wir über unterschiedliche Metriken ermittelt (z.B. Cohen's Kappa, Krippendorf's Alpha)
- **Annotation Tools**: Software zu manueller (Beispiel: INCEPTION) oder automatischer (Beispiel: Stanza) Annotation oder Mischformen daraus.

Annotationsrichtlinien

- Sollten sowohl **generisch**, als auch **präzise** sein.
- Sicherstellung von zuverlässiger, intersubjektiver Annotation.
- Identifikation von Lücken und Mehrdeutigkeiten für Zuweisung und Lokalisation.
- Iterativer Prozess: Schrittweise Verbesserung der Richtlinien



(aus Reiter 2020:194)

Beispiel-Theorie: Zeichentypen

Piktogramme (Bildzeichen): Bildlich erkennbare, festgelegte Inhaltseite, aber keiner Konvention für die Ausdrucksseite. Sprachübergreifend verständlich.

Bsp: 

Ideogramme (Begriffszeichen): Wie Piktogramme, aber nicht bildlich erkennbar

Bsp: 

Logogramme (Wortzeichen): Festgelegte Inhalts- und Ausdrucksseite

Bsp: @ € 3

Phonogramme (Lautzeichen): Festgelegte Ausdrucksseite, inhaltsunabhängig

Bsp: a N ə ç

Aufgabe: Annotation der Zeichen

1. Annotieren Sie im Beispieltext (zunächst ausschließlich auf Seite 1) alle Logogramme, Ideogramme und Piktogramme.
2. Setzen Sie sich in Dreiergruppen zusammen und ermitteln Sie, wo Sie unterschiedlich annotiert haben. Dokumentieren Sie diese Fälle (bspw. in einer Tabelle).
3. Erstellen Sie eine erste Version für Annotationsrichtlinien (vgl. Reiter 2020 Abb. 2), gerne auch mit Beispielen versehen.
4. Annotieren Sie Seite 2 des Beispieltextes in der Gruppe. Dokumentieren Sie, ob die Annotationsrichtlinien geändert werden müssen was der Grund dafür war. Reichen Sie Ihre Richtlinien bei ILIAS ein (siehe nächste Seite).

Literatur / Hausaufgabe

- **Schriftliche Aufgabe [obligatorisch!]**: Reichen Sie die ausgearbeiteten Annotationsrichtlinien Ihrer Gruppe bis zum 11.11. in ILIAS ein. Bitte spezifizieren Sie zu Anfang der Datei, wer bei den Richtlinien mitgewirkt hat (Namen und Matrikelnummer). Es muss nur ein:e Teilnehmer:in pro Gruppe einreichen!
- **Lesen zur Nachbereitung**:
  - [03A] Andresen (2024): Kapitel 1.3 (S. 16-20) und 10 (S. 143-156)
  - [03B] Reiter (2020): Anleitung (S. 193-201)
  - bei Interesse gerne auch [04B] Dürscheid 2016 (S. 65-69)