Learning by doing:
Recording a speech corpus as seminar project

Beeke Muhlack

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Seminar

- BA Empirische Sprachwissenschaft
- Bachelorseminar „Sprachdeskription und -dokumentation“ (4 SWS)
- Teilnehmer:innen: 9 Studierende
- Lernziele:
  - Planung und Durchführung von Sprachaufnahmen
  - mit Aufnahmekabine und –technik vertraut werden
  - Datenaufbereitung & Analyse
- Prüfungsleistungen: Posterpräsentation + Hausarbeit zum eigenen Thema

Multilingual **Co**rpus of **S**pontaneous **S**peech

- 9 Sprecher:innen im Alter von 18 bis 50 Jahren (8 weiblich, 1 männlich)
- Nicht-deutsche Muttersprachler:innen mit Deutsch und Englisch als L2 (mind. B1-Niveau, LexTALE)
  - L1: Französisch, Italienisch, Spanisch, Portugiesisch, Schwedisch, Polnisch, Russisch, Thai
- Spontansprache (5 min, Interviewfragen), Wortlisten mit Vokalphonemen in allen drei Sprachen (L1, DE, EN)
- Veröffentlichung im Goethe University Data Repository (GUDe): **Public access**

Datenaufbereitung

- Ersetzen von identifizierendem Material mit Sinuston (z.B. Namen, Orte)
- Automatische Transkription mit der KI Software **WhisperAI** (Textdatei)
- **Manuelle Korrektur** von Whisper-Output (Fehlerkorrektur, Ergänzen von Unflüssigkeiten)
- Erstellung von TextGrids mit **WebMAUS**
- Grobe Korrektur von TextGrids
- Zusammenfassen von sprachbiographischen Daten

Seminarplan

Woche	Sitzung 1	Sitzung 2
1	Organisatorisches, Einführung	Phasenmodell (Draxler, 2008), Forschungsfragen
2	Korpusarbeit + Spezifikation Korpus	Phasenmodell: Vorbereitung, Declaration of Helsinki
3	Besprechung versch. Aufgabentypen	Erstellung des Promptmaterials, Fragebogen
4	Aufnahmetechnik (Mikrophone, Digitalisierung)	Aufnahmetest
5	Besprechung wissenschaftl. Poster	Exkurs Feldforschung
6	Gastvortrag zum Thema Feldforschung	Exkurs Feldforschung
7	Data cleaning	Arbeiten mit WhisperAI
8	Annotationsschemata, WebMAUS	Arbeiten mit WebMAUS
9	Scripting in Praat	Scripting in Praat
10	Statistik in R	Statistik in R
11	Bearbeitungszeit Poster	Bearbeitungszeit Poster
12	Bearbeitungszeit Poster	Bearbeitungszeit Poster
13	Hausarbeitsbesprechung	Posterpräsentationen

Feedback

„Die allgemeine Struktur [hat mir gefallen], da sehr viele Kleinigkeiten und Feinheiten besprochen wurden, an die man selber niemals denken würde.“

„Die Schritt-für-Schritt Anleitung war sehr angenehm, sehr praxisnah. Nichts davon war unnötig, man konnte es direkt selbst anwenden.“

„Ich empfand es als sehr angenehmes learning by doing, bei dem man von der Lehrperson bei Fragen und Problemen unterstützt wird.“

„Mich hat die Arbeit am Projekt motiviert zur Sitzung zu kommen. Es gab eine nützliche Verbindung zu Inhalten aus anderen Veranstaltungen.“

„Ich fand es gut, dass wir gleich in den Sitzungen an unseren Aufnahmen arbeiten konnten.“

„Die Methoden wurden umfangreich behandelt und das Korpus-Projekt macht die Arbeit relevanter.“

Literatur und Programme

- Seminarliteratur: Draxler, C. (2008). Korpusbasierte Sprachverarbeitung: Eine Einführung. Narr Studienbücher.
- LexTALE (Lemhöfer & Broersma, 2012): https://www.lextale.com/takethetest.html
- WhisperAI: https://openai.com/index/whisper/
- Tutorial zum WhisperAI Installationsprozess: https://www.youtube.com/watch?v=ABFqbY_rmEk
- WebMAUS: https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/WebMAUSBasic
- GUDe repository: https://gude.uni-frankfurt.de/home

b.muhlack@em.uni-frankfurt.de