IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Erfahrungen mit der KI-gestützten Transkription und Annotation von gesprochenen Sprachdaten

Jan Gorisch & Mark-Christoph Müller

AUSGANGSLAGE

Die Anwendung von aktueller automatischer Spracherkennung (ASR), entwickelt für z.B. die Transkription von YouTube-Videos (Szenario A), wird momentan auch in wissenschaftlichen Szenarien wie z.B. Gesprächsanalyse oder Untersuchung von Lernersprache verwendet (Szenario B). Wegen der stark unterschiedlichen Anforderungen kann dies zu großem Korrekturaufwand führen. Es zeigt sich, dass es momentan noch keine Tools gibt, die auf Szenario B zielgenau passen würden. Gleichzeitig kann es Sinn machen, diese ASR-Tools – wegen Alternativlosigkeit (manuelle Transkription zu zeit- und kostenintensiv) – als ersten Schritt bei der Erschließung von Sprachaufnahmen einzusetzen, solange man den möglichen Nutzen (Ersttranskription ermöglicht inhaltliche Einblicke in die Aufnahmen) und die Risiken (Ersttranskription kann falsch und inhaltlich irreführend sein) bewusst abwägt.

I. ERFAHRUNGEN

Basierend auf Daten des AGD (Archiv für Gesprochenes Deutsch, agd.ids-mannheim.de)

ASR-Tools: Sie machen zuverlässig, worauf sie trainiert wurden, „versagen" aber (noch) bei

- herausfordernden akustischen Bedingungen (z.B. Raumhall) ⟶ schlechte Erkennungsrate und Sprecherseparierung
- historischen/regionalen Bedingungen (Aufnahmen aus den 50er/60er Jahren im Dialekt) ⟶ schlechte Erkennungsrate, unerwünschte Normalisierung (siehe Beispiel rechts)
- Überlappungspassagen bei einkanaligen Aufnahmen ⟶ mangelhafte Sprecherseparierung
- gesprächstypischen Phänomenen wie Häsitationen, Diskursmarkern ⟶ werden ignoriert
- Wiedergabe tatsächlich geäußerter Satzstrukturen (z.B. Abbrüche, Tempus) ⟶ unerwünschte Normalisierung und Veränderung (siehe Beispiel rechts)
- Code-Switching ⟶ Unerwünschtes Übersetzen von mehrsprachigen Äußerungen

Referenz (manuell, literarische Transkription)	*abr*	*des*	*isch*	*früher*	*net*	*g'si*
Referenz (manuell, normalisiert)	*aber*	*das*	*ist*	*früher*	*nicht*	*gewesen*
Hypothese (ASR)	*aber*	*das*	*war*	*früher*	*nicht*	*so*

Beispiel (1) aus Korpus SV (Südwestdeutschland und Vorarlberg, 1966): Aufnahmeereignis E_00019/04m34s; transkribiert mittels OpenAI-Whisper, Modell large (2024).

Diese Glättungen bzw. automatischen Bearbeitungen geschehen, weil sie in Anwendungsszenarien vom Typ A (s.o.) in der Regel erwünscht sind. Aus der Perspektive z.B. der Gesprächsforschung oder bei Lernersprachen (Typ B) sind diese Eingriffe (außer für eine einfache inhaltliche Ersterschließung sehr großer Datenmengen) aber inakzeptabel und die sich daraus ergebenden Transkripte daher nicht sinnvoll verwendbar.

II. MÖGLICHKEITEN

Anwendung von ASR zur Ersterschließung von Audio und Video-Aufnahmen

- Transkribiert gesprochene Sprache automatisch
- Separiert Gesprächsanteile nach Sprecher:innen
- Aligniert Text und Ton
- Kann lokal auf handelsüblicher Hardware installiert werden (DSGVO-kompatibel)
- Output kann interoperabel zwischen Analyseprogrammen transferiert werden

Praktischer Workflow mit ASR

- Erhebung von Primärdaten: Audio/Video und Metadaten
- Schneiden der Medien
- Transkribieren (mit Unterstützung von ASR)
- Importieren in Transkriptionseditor (EXMARaLDA: Partitur-Editor oder FOLKER, OCTRA, ...)
- Korrekturschritte (optional)
  - Prüfen der Sprecherzuordnung
  - Korrigieren der Transkription (von einzelnen Wörtern bis hin zu Glättung rückgängig machen)
  - Anpassen der Alignierung

ASR-Tools (Beispielauswahl)

- aTrain (Grafisches User-Interface auf MS-Windows)
- OpenAI Whisper bzw. WhisperX (auf Kommandozeile oder – sofern für die Daten zulässig – via BAS Web-Services) https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface)

Interoperabilität und Analyse in Transkriptions-Tools

- Import: ASR Output (z.B. *.json, *.vtt oder *.srt) kann in spezialisierten Transkriptionseditoren importiert werden
- Annotation: POS-Tagging, Lemmatisierung, Normalisierung bzw. literarische Umschrift (automatisch oder manuell, z.B. via OrthoNormal)
- Export: in weitere Formate, wie z.B. Praat (*.TextGrid) oder ELAN (*.eaf)
- Korpuserstellung: z.B. via Corpus-Manager (*.coma)
- Recherche: z.B. in EXAKT (auch verknüpft mit Metadaten und Annotationen)
  - Export: KWIC (Keywords in Context)
  - Detailanalyse je nach Forschungsfeld (Excel, R, ...)

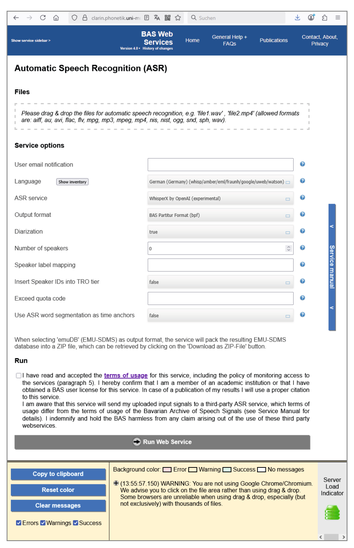
Wichtig: Gute Audio-Qualität

- z.B. mit Headset- oder Ansteckmikrofonen arbeiten; individuell je Sprecher:in (alternativ: KI-gestützte Kanalseparierung)
  - bessere Sprecherseparierung (auch bei Überlappungen)
  - höhere Qualität der Transkripte ⟶ weniger Korrekturaufwand

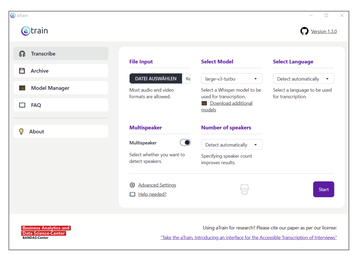
Datenerhebung und Fallstricke

„DSGVO-kompatibel" bedeutet nur, dass die Daten bei der lokalen ASR-Installation den Rechner nicht verlassen. Es bedeutet nicht, dass man mit den Daten alles machen darf. Es ist stets individuell zu prüfen, was mit den aufgenommenen Personen in der Einwilligungserklärung vereinbart wurde*.
Idealerweise wird der oder die **Datenschutzbeauftragte** und ein **Forschungsdatenzentrum** bzgl. der möglichen **Nachnutzung** der Daten bereits in der Aufnahmeplanung hinzugezogen.

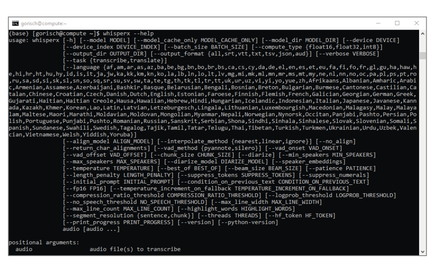
*Disclaimer: Wir machen hier keine Rechtsberatung.



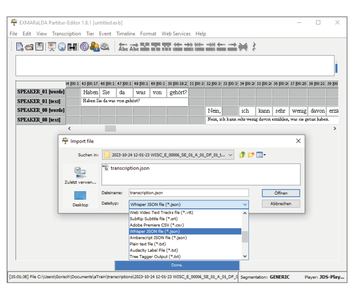
◂ BAS Web-Services ASR



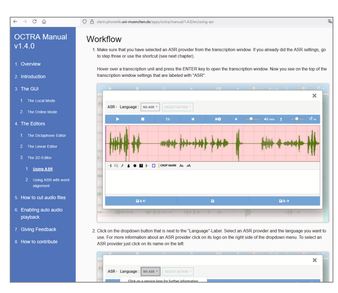
aTrain GUI ▸



◂ WhisperX



◂ EXMARaLDA



OCTRA ▸

III. AUSBLICK

Wann werden ASR-Systeme realistische, für die Linguistik verwendbare Transkripte produzieren? Es gibt erste prototypische Ansätze, z.B.

- GailBot
- CrisperWhisper
- noScribe

Die Entwicklung ist aber sehr langsam. Weitere Alternative:

- Finetuning existierender Modelle (experimentell, technisch anspruchsvoll)

Kontakt:
Dr. Jan Gorisch
Dr. Mark-Christoph Müller
Abteilung Pragmatik
Postfach 10 16 21
68016 Mannheim, Germany
{gorisch|mark-christoph.mueller}
@ids-mannheim.de

Hausanschrift:
Leibniz-Institut für Deutsche Sprache
R 5, 6-13
68161 Mannheim
Tel: +49 621 1581-0
Fax: +49 621 1581-200
info@ids-mannheim.de
www.ids-mannheim.de

Leibniz Gemeinschaft

Literaturhinweise

Gorisch, Jan/Schmidt, Thomas (2024): Evaluating Workflows for Creating Orthographic Transcripts for Oral Corpora by Transcribing from Scratch or Correcting ASR-Output. In: Proceedings of LREC-COLING 2024. Paris: ELRA, 2024. S. 6564–6574.
Haberl, Armin/Fleiß, Jürgen/Kowald, Dominik/Thalmann, Stefan (2024): Take the aTrain. Introducing an Interface for the Accessible Transcription of Interviews. Journal of Behavioral and Experimental Finance 41 (2024): 100891.
Radford, Alec/Kim, Jong Wook/Xu, Tao/Brockman, Greg/McLeavey, Christine/Sutskever, Ilya. Robust Speech Recognition via Large-scale Weak Supervision. In: International Conference on Machine Learning. PMLR, 2023. S. 28492-28518.
Schmidt, Thomas (2012): EXMARaLDA and the FOLK Tools. Two Toolsets for Transcribing and Annotating Spoken Language. In: Proceedings of LREC-12, Paris ELRA, 2012. S. 236-240.