ASR4Memory

Ein KI-gestütztes Transkriptionsangebot für historische audiovisuelle Forschungsdaten

Gefördert durch die Deutsche Forschungsgemeinschaft (DFG); Projektnummer 501609550

Dr. Tobias Kilgus, Peter Kompiel | Freie Universität Berlin, 4Memory Incubator Funds, 12.03.2024

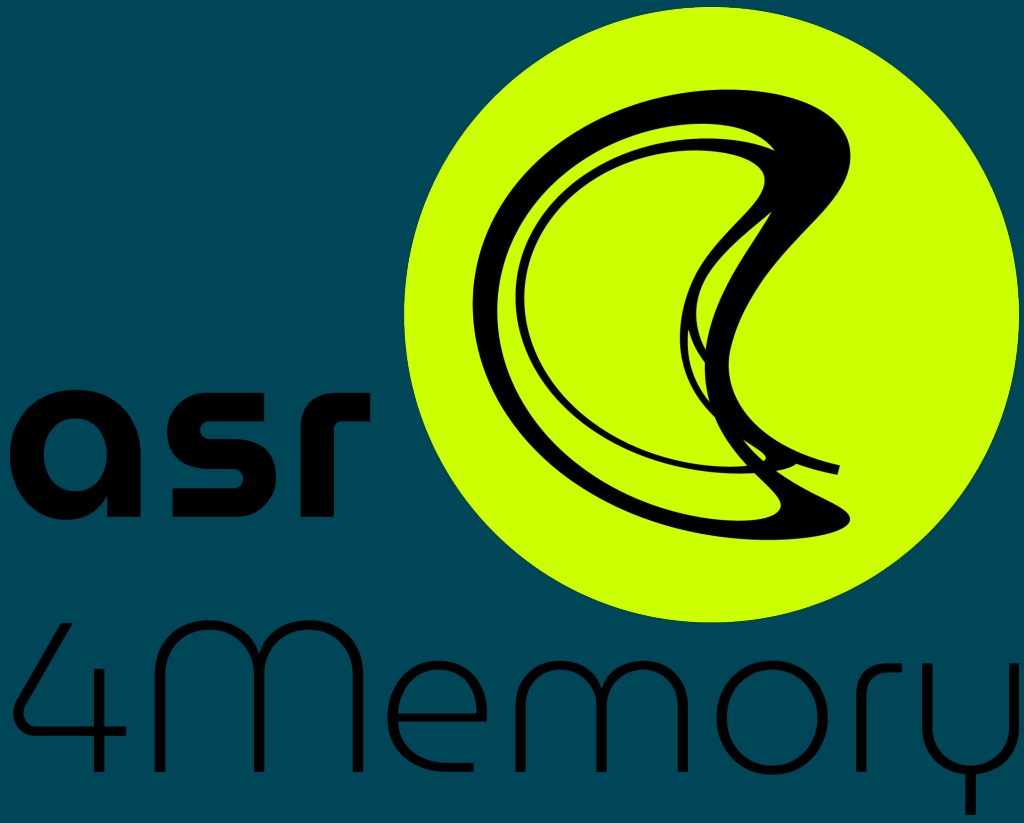
Rahmendaten

- **Projektlaufzeit**: Januar bis Dezember 2024
- **Mitarbeiter**: Peter Kompiel, Marc Altmann, Tobias Kilgus
- **4Memory Task Area 1**: Data Quality
- **Umsetzung**: Universitätsbibliothek der FU: Abt. Forschungs- & Publikationsservices mit Universitätsarchiv
- **Webseite**: https://www.fu-berlin.de/asr4memory
- **Github-Repository**: https://github.com/asr4memory

Ausgangslage

- **Umfangreiche Bestände/Sammlungen** von historischen audiovisuellen Ressourcen, die inhaltlich/wissenschaftlich zu erschließen sind (eine Grundlage ist Verschriftung)
- **Bislang Transkription** von AV-Ressourcen durch:
  - Manuelle Verschriftung
    ➔ sehr zeitaufwändig und kostenintensiv
  - Kommerzielle, proprietäre Transkriptionsdienste
    ➔ (z.T.) datenschutzproblematisch und kostenintensiv
    ➔ Ergebnisse oft nicht zufriedenstellend mit Blick auf Transkriptionsgenauigkeit und Exportformate

Was ist unser Ziel?



Zielsetzungen

- **Angebot** für die Forschungscommunity zur automatisierten Transkription von audiovisuellen Forschungsdaten in geschichtswissenschaftlichen Kontexten

  ➔ Angebot: prototypisch, (noch) nicht skalierend, bislang limitierte Rechenkapazitäten
  ➔ Projekt: Entwicklung von Knowhow und Programmcode (erweiterbar, anwendbar und skalierbar)

Zielsetzungen

- Nutzung von Algorithmen der **Künstlichen Intelligenz** (KI) bei der **Automatisierten Spracherkennung** (ASR)
  ➔ Deep-Learning-Modelle (Transformer) ermöglichen:
    → **Erkennung** komplexer Muster in großen Datensätzen
    → Bessere **Verarbeitung** versch. Sprecher und Akzente
    ➔ Höhere **Transkriptionsqualität** im Vergleich zu früheren ASR-Ansätzen
- **Open-Source-basierte ASR**
  ➔ lizenzkostenfrei
  ➔ Code flexibel nachnutzbar und anpassbar

Zielsetzungen

- **Bestmögliche Datenqualität**: Niedrige Wortfehlerrate im Prozess der automatisierten Transkription (➔ WER)
- **Zeiteffizient**: Automatisierte Prozesse mit hoher Performance bei großen Datenbeständen (➔ RTF)
- **Bedarfsorientiert**: Bereitstellung standardisierter Transkript- und Metadatenformate für die Community
- **Datenschutzkonform**: Verarbeitung der Daten auf lokalen Servern (keine Clouds oder externe Dienste)

Zielsetzungen

- **Für heterogene Quellen:**
  - Zeitzeugeninterviews, Dokumentarfilme, Tonaufzeichnungen, Radioübertragungen, Fernsehsendungen, akademische/politische Veranstaltungen/Vorträge, etc.
- **Unterstützung verschiedener Sprachen**:
  - deutsch, englisch, spanisch, ukrainisch, mandarin, usw.
- **Einsatz in unterschiedlichen Forschungs-, Nachnutzungs- und Archivierungsszenarien** (GLAM):
  - Oral-History-Projekte, Archive, Gedenkstätten, Museen, Bibliotheken

Zielsetzungen

- **Synchronisierung** zwischen Audio/Video und Text (mit Zeitcodierung/Timecodes):

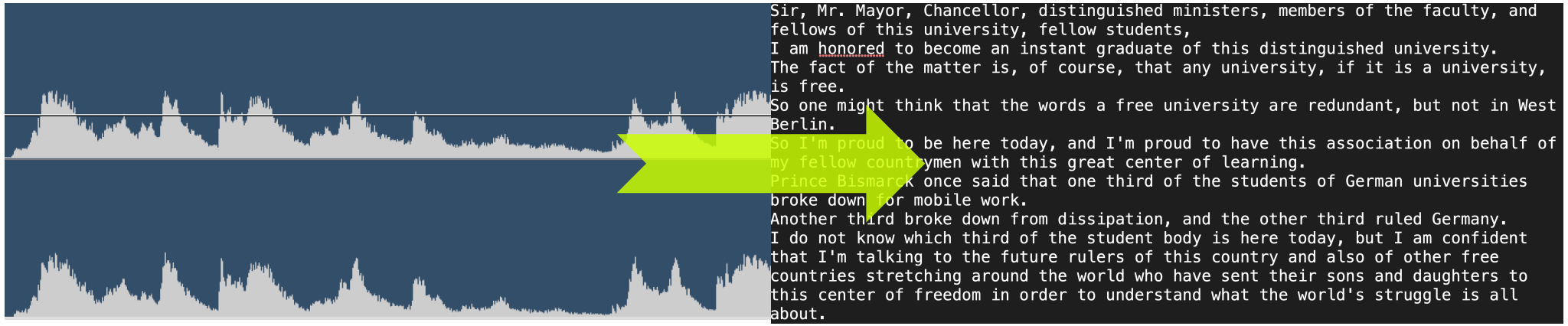
➔ verbesserte Auffindbarkeit, Zugänglichkeit, Interoperabilität und Nachnutzbarkeit von AV-Ressourcen (FAIR)

➔ weitere Nutzungsmöglichkeiten, u.a. im Bereich der Textanalyse (Topic Modelling, NER) oder Anonymisierung/Pseudonymisierung von AV-Ressourcen

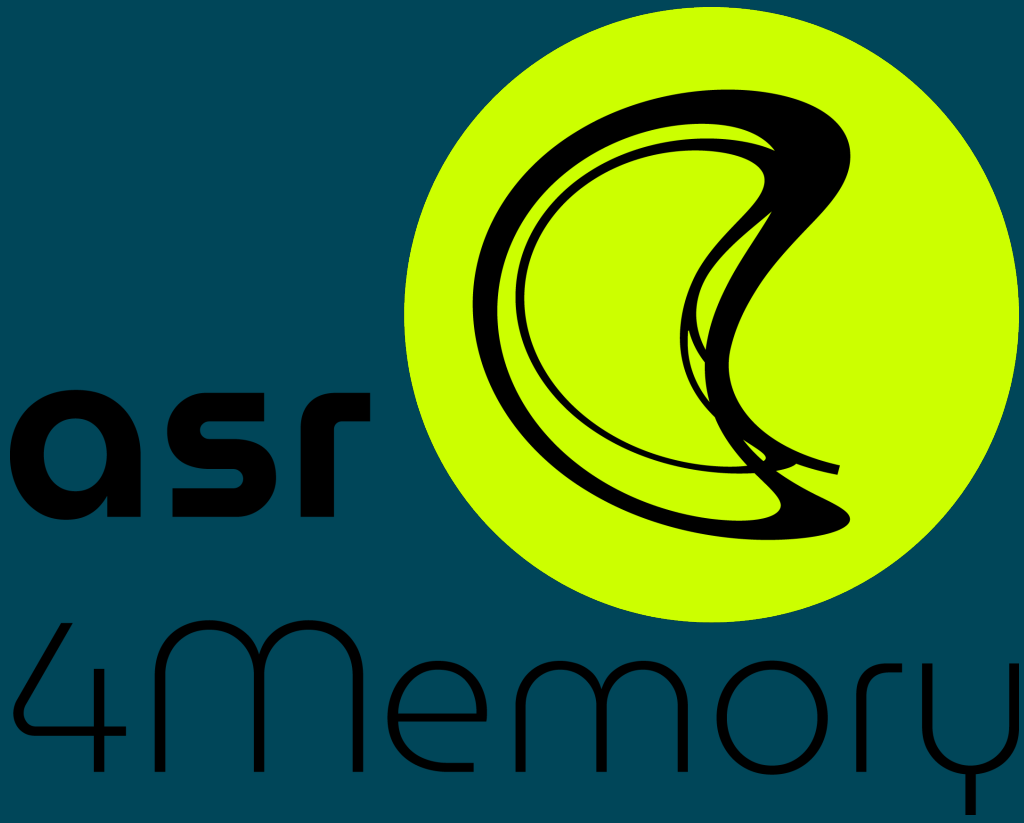
Zielsetzungen

Transkriptions-Pipeline: Speech-to-Text-Wandlung

Upload der AV-Ressourcen → Optimierung der Tonspur → Automatische Transkription → Generierung der Exportformate → Bereitstellung der Exportformate



Wie gehen wir vor?



Vorgehensweise

Phase 1 (Januar bis Mai 2024)

- **„Grundlagenforschung“**: Verstehen der ASR-„Blackbox“ → Erklärbarkeit der Ergebnisse?
- **Bedarfsermittlung** in der Forschungscommunity (Workshop): Pilotnutzende: → Input/Quellen; → Output/Anforderungen
- **Autom. Aufbereitung** d. AV-Quellen in höchstmöglicher Audio-/Videoqualität für optimale ASR (Input)
- **Konvertierung** der ASR-Ergebnisse in standardisierte, zeitkodierte Transkript- und Metadatenformate (Output)

Vorgehensweise

Phase 2 (Juni bis September 2024)

- **Evaluation** der Open-Source-Spracherkenner: OpenAI Whisper, Nvidia NeMo etc. ➔ Vergleich mit ASR-Providern
- **Automatisierung** der Workflows und **Optimierung** der Rechenperformance für Batch Processing
- **Anpassung und Reflektion** der Sprachmodelle: Training und Feintuning ➔ domänen-spezifisches ASR-Modell?

Vorgehensweise

Phase 3 (Oktober bis Dezember 2024)

- **Schnittstellen-Entwicklung**:
  - API für Anbindung externer Umgebungen an ASR-Pipeline
  - Entwicklung eines Web-Services (GUI) für ASR-Pipeline
- **Rückkopplung** mit den Pilotnutzenden (kontinuierlich)
  ➔ Nutzungscommunity, Dokumentation, Best-Practice
- **Veröffentlichung** der Projektergebnisse im 4Memory-Konsortium und auf Digital-Humanities-Veranstaltungen

Zwischenstand I

- **Beispielergebnisse (Projektstand)**:
  - siehe Projektwebseite

- Beispiele für
  - **Genauigkeit** der Transkription und Timecodes
  - **Schwächen** der ASR (z.B. Dialekt)

Zwischenstand II

- Exportformate: TXT VTT

```
Sir, Mr. Mayor, Chancellor, distinguished ministers, members of
the faculty, and fellows of this university, fellow students,
I am honored to become an instant graduate of this distinguished
university.
The fact of the matter is, of course, that any university, if it
is a university, is free.
So one might think that the words a free university are
redundant, but not in West Berlin.
So I'm proud to be here today, and I'm proud to have this
association on behalf of my fellow countrymen with this great
center of learning.
Prince Bismarck once said that one third of the students of
German universities broke down for mobile work.
Another third broke down from dissipation, and the other third
ruled Germany.
I do not know which third of the student body is here today, but
I am confident that I'm talking to the future rulers of this
country and also of other free countries stretching around the
world who have sent their sons and daughters to this center of
```

```
WEBVTT

1
00:00:01.164 --> 00:00:14.503
Sir, Mr. Mayor, Chancellor, distinguished ministers, members of the
faculty, and fellows of this university, fellow students,

2
00:00:14.503 --> 00:00:22.721
I am honored to become an instant graduate of this distinguished
university.

3
00:00:24.082 --> 00:00:30.727
The fact of the matter is, of course, that any university, if it is a
university, is free.

4
00:00:32.560 --> 00:00:39.806
So one might think that the words a free university are redundant,
```

Zwischenstand III

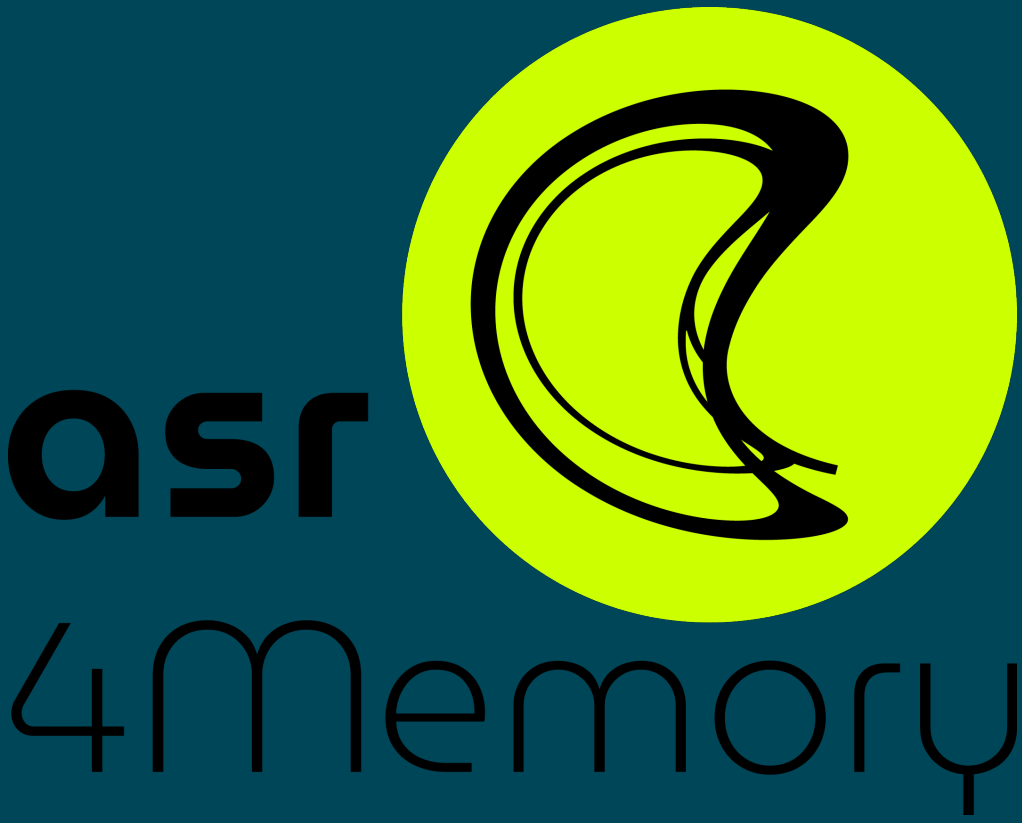
- Exportformate: JSON CSV


```
{
    "start": 1.164,
    "end": 14.503727722772277,
    "sentence": "Sir, Mr. Mayor, Chancellor, distinguished ministers, members of
},
{
    "start": 14.503727722772277,
    "end": 22.721,
    "sentence": "I am honored to become an instant graduate of this distinguishe
},
{
    "start": 24.082,
    "end": 30.727,
    "sentence": "The fact of the matter is, of course, that any university, if it
},
{
    "start": 32.56,
    "end": 39.806,
    "sentence": "So one might think that the words a free university are redundar
},
{
    "start": 41.067,
    "end": 50.454,
    "sentence": "So I'm proud to be here today, and I'm proud to have this associ
```


IN	TRANSCRIPT
00:00:01.164	Sir, Mr. Mayor, Chancellor, distinguished ministers, members of the faculty, and fellows of this university, fe
00:00:14.504	I am honored to become an instant graduate of this distinguished university.
00:00:24.082	The fact of the matter is, of course, that any university, if it is a university, is free.
00:00:32.560	So one might think that the words a free university are redundant, but not in West Berlin.
00:00:41.067	So I'm proud to be here today, and I'm proud to have this association on behalf of my fellow countrymen w
00:00:52.256	Prince Bismarck once said that one third of the students of German universities broke down for mobile wo
00:01:01.570	Another third broke down from dissipation, and the other third ruled Germany.
00:01:08.713	I do not know which third of the student body is here today, but I am confident that I'm talking to the future
00:01:35.037	I know that when you leave this school, you will not imagine that this institution was founded by citizens of
00:01:48.189	and was developed by citizens of West Berlin, that you will not imagine that these men who teach you hav
00:02:10.448	This school is not interested in turning out merely corporation lawyers or skilled accountants.
00:02:18.774	What it is interested in, and this must be true of every university, It must be interested in turning out citizen
00:02:28.925	men who comprehend the difficult, sensitive tasks that lie before us as free men and women, and men wh
00:02:44.141	That's why you're here, and that's why this school was founded, and all of us benefit from it.
00:03:08.126	It is a fact that in my own country, in the American Revolution, that revolution and the society developed th

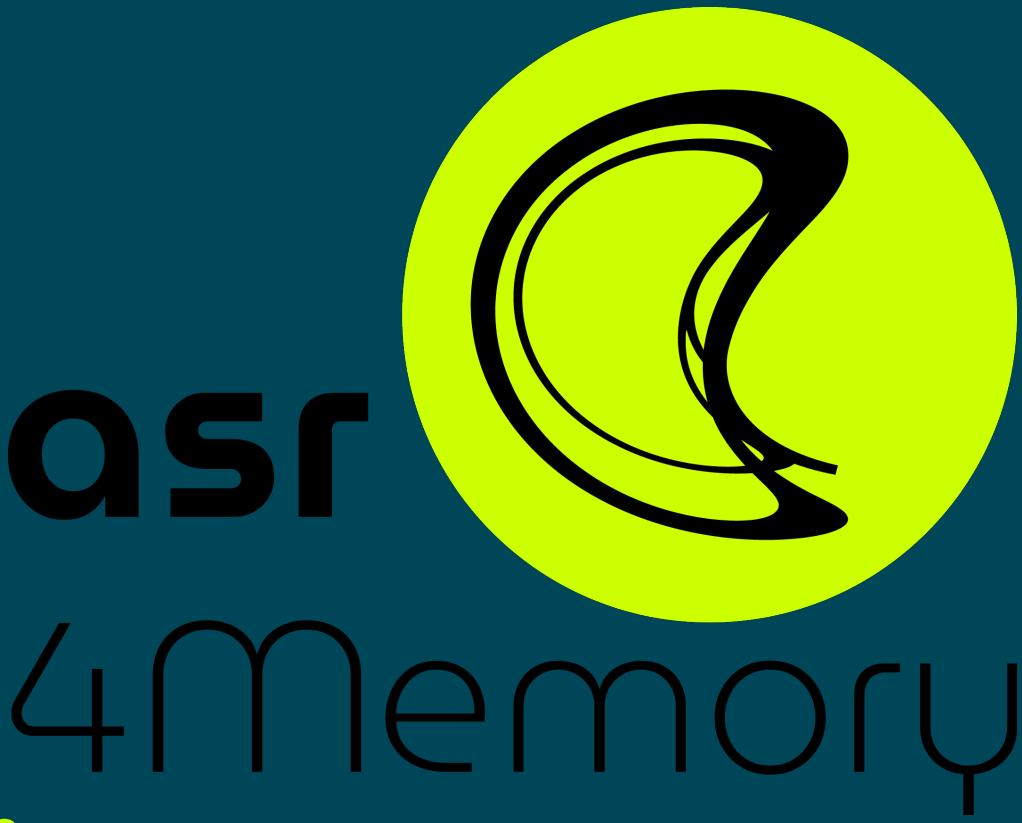
➔ Weitere Formate: IIIF-AV, TEI-XML, CMDI, EAD etc. ?

Wie knüpfen wir an 4Memory an?

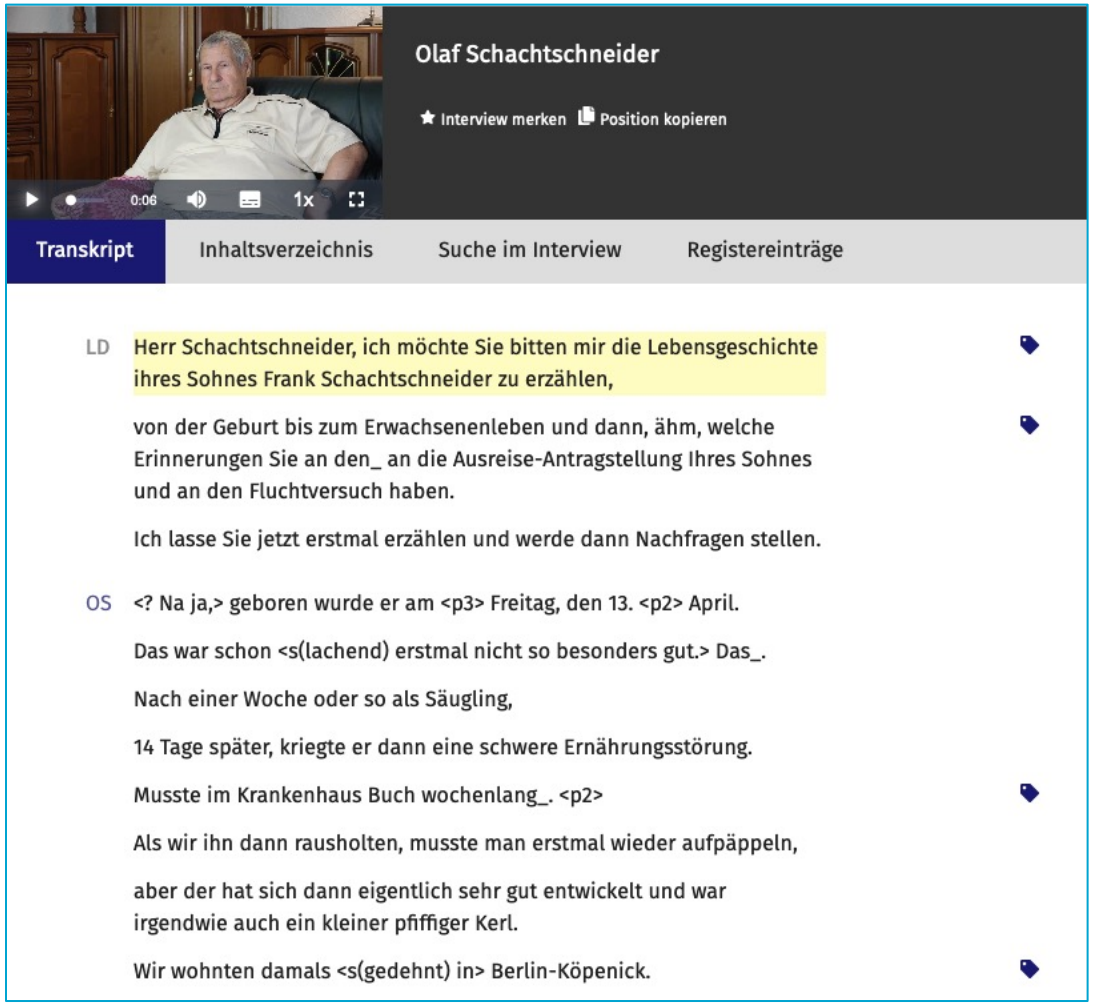


Wie knüpfen wir an 4Memory an?

➔ Datenqualität und automatisch erstellte Transkripte



Beispiel: Oral History-Interview



- Interview zum Thema „gescheiterte Fluchten aus der DDR“
- Erschlossen in der Recherche- und Erschließungsumgebung Oral-History.Digital im Rahmen des Projekts „Eiserner Vorhang“
- Manuelle Transkription und Alignment

Schachtschneider, Olaf , Interview ev001, 03.06.2019, Interview-Archiv "Eiserner Vorhang", https://archiv.eiserner-vorhang.de/de/interviews/ev001, 07.03.2024

ASR vs. manuelle Transkription

0) Sprechererkennung

1) Füllwörter

OT: „Der hat uns da, äh, äh, ein paar Sachen, äh, äh, Schriftstücke gezeigt, die also eigentlich vollkommen belanglos waren.“[1]

WhisperX: „Der hat uns da ein paar Sachen, Schriftstücke gezeigt, die also eigentlich vollkommen belanglos waren.“

[1] Schachtschneider, Olaf , Interview ev001, 03.06.2019, Band 1 – 0:06:07, Interview-Archiv "Eiserner Vorhang", https://archiv.eiserner-vorhang.de/de/interviews/ev001?tape=1&time=0h06m07s, 07.03.2024

ASR vs. manuelle Transkription

2) Wiederholungen / Satzabbrüche / Verzögerungen

OT: „Wir wohnten damals in Berlin-Köpenick. In_ in_. Äh, na, eine Hütte war es ja eigentlich gewesen.“[2]

WhisperX: „Wir wohnten damals in Berlin-Köpenick. In einer Hütte war es ja eigentlich.“

[2] Schachtschneider, Olaf , Interview ev001, 03.06.2019, Band 1 – 0:01:09, Interview-Archiv "Eiserner Vorhang", https://archiv.eiserner-vorhang.de/de/interviews/ev001?tape=1&time=0h01m09s, 07.03.2024

ASR vs. manuelle Transkription

3) Dialekte

Beispiele: „ick“ → „ich“ / „jewesen“ → „gewesen“ / „kriech‘ ta“ → „kriegte er“ / usw.

OT: „Nach einer Woche oder so als Säugling, 14 Tage später, kriegte er dann eine schwere Ernährungsstörung.“

WhisperX: „Nach einer Woche oder so als Säugling, 14 Tage später, kriegte ich dann eine schwere Ernährungsstörung.“

[3] Schachtschneider, Olaf , Interview ev001, 03.06.2019, Band 1 – 0:00:42, Interview-Archiv "Eiserner Vorhang", https://archiv.eiserner-vorhang.de/de/interviews/ev001?tape=1&time=0h00m42s, 07.03.2024

ASR vs. manuelle Transkription

4) Named Entities

OT: „Ja, das war Sommer 89, wo die dann alle da in Prag, da in der Botschaft gesessen haben und der Genscher hat sie dann da rausgekloppt.“[4]

WhisperX: „Ja, das war Sommer 89, wo die dann alle da in Prag da in der Botschaft gesessen haben und der Kenja hat sie dann da rausgeklappt.“

[4] Schachtschneider, Olaf , Interview ev001, 03.06.2019, Band 1 – 0:20:33, Interview-Archiv "Eiserner Vorhang", https://archiv.eiserner-vorhang.de/de/interviews/ev001?tape=1&time=0h20m33s, 07.03.2024

ASR vs. manuelle Transkription

5) non-verbale Kommunikation / Pausen / direkte Rede

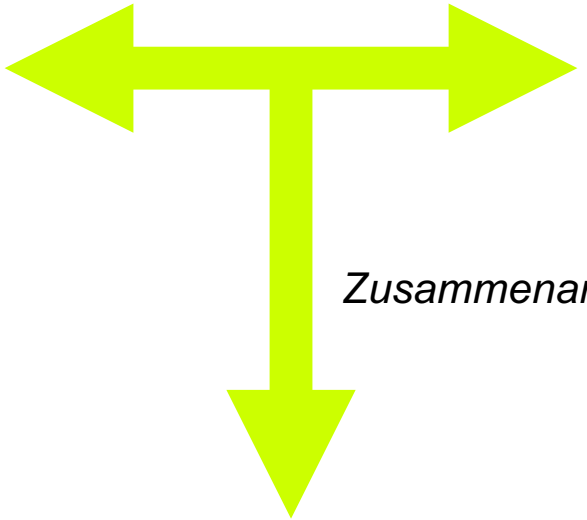
OT: „Ich sage: ‚Weißt du was, jetzt fahre ich mal schnell rüber, hole uns ein paar Zigaretten.‘ <s(lachend) Fahre die Schönhauser runter, bieg in die Brunnenstraße ein,> <g(darstellende Handbewegungen) da standen sie, einer nebeneinander.> <p4> Da bin ich nicht weitergekommen.“[5]

WhisperX: „Ich sage, weißt du was, jetzt fahre ich mal schnell rüber, hol uns ein paar Zigaretten. Ich fahre da schön runter und da kriege ich eine Brunnenstraße hin. Da standen sie jeder nebeneinander. Da bin ich nicht weitergekommen.“

[5] Schachtschneider, Olaf , Interview ev001, 03.06.2019, Band 1 – 0:11:23, Interview-Archiv "Eiserner Vorhang", https://archiv.eiserner-vorhang.de/de/interviews/ev001?tape=1&time=0h11m23s, 07.03.2024

Optimierung der Spracherkennung

„Blackbox“ KI-Sprachmodell



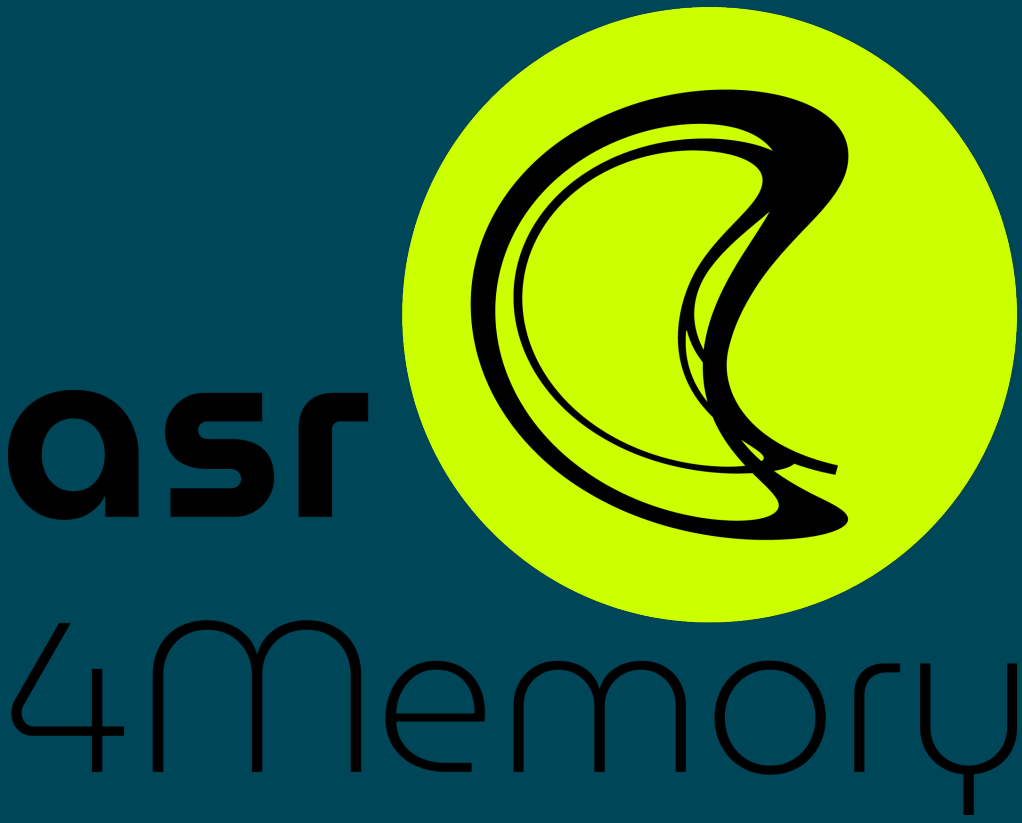
Geschichtswissenschaftliche Standards

Zusammenarbeit mit 4Memory-Community

Welche Anforderungen an qualitativ hochwertige Transkriptionen gibt es?

➔ Verbesserung/Anpassung der Spracherkennung (Audio-Optimierung, Evaluation, Finetuning)

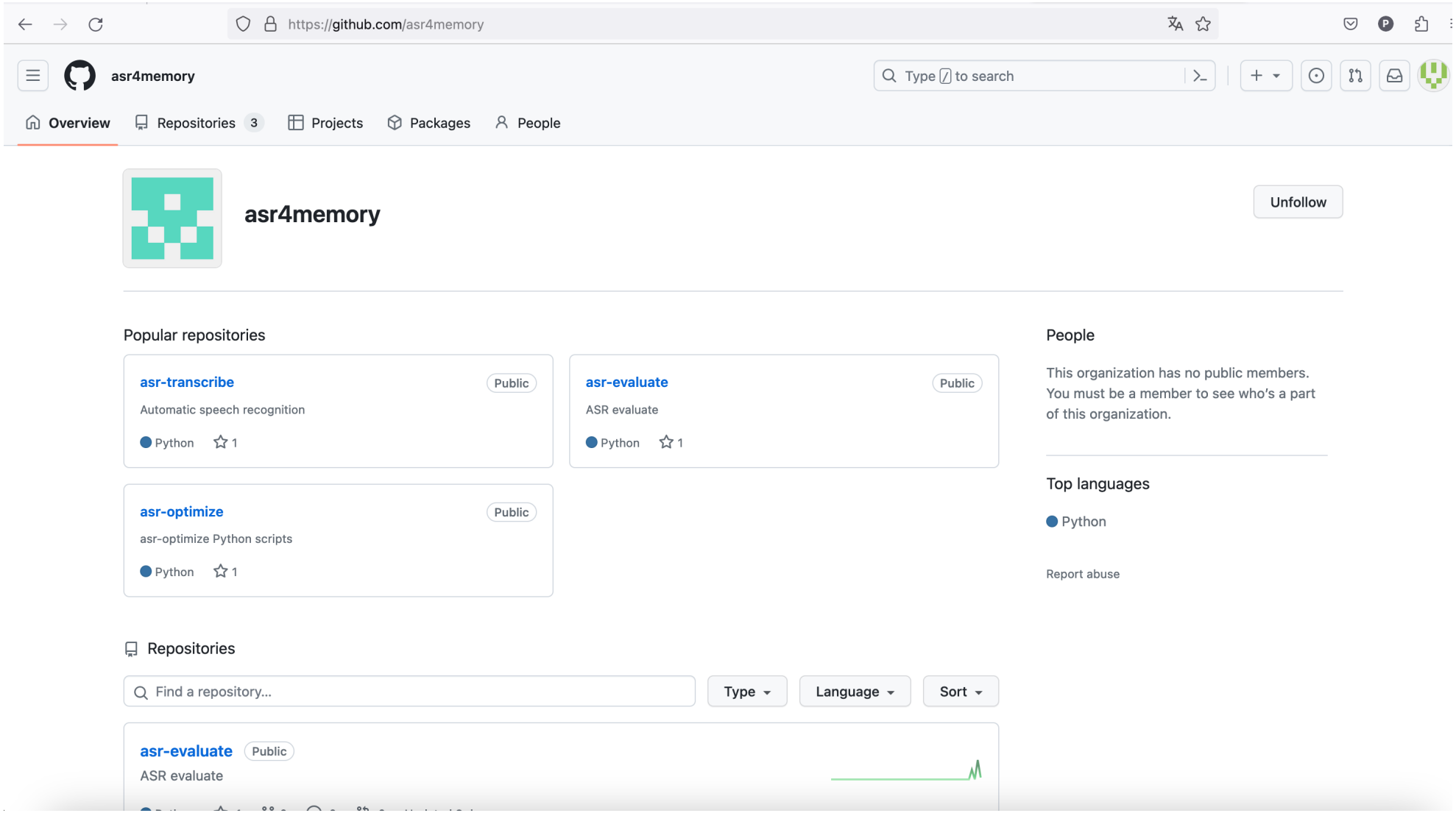
Wer kann dabei wie mitmachen?



Wer kann dabei wie mitmachen?

- Hands On-Workshop am 18.03.2024 mit bestehenden Pilotnutzenden
- Bereitstellung audiovisueller Datensätze durch neue Pilotnutzende
- Öffentlich zugängliches Github-Repositorium

Github-Repositorium



Vielen Dank!