UNIVERSITÄT
LEIPZIG

Forschungsseminar CSS – OCR and Transcription

GWZ H2 1.15, 06.11.2025

Felix Lennert, M.Sc.

OUTLINE

- Optical Character Recognition
  - Brief history
  - Problems and pitfalls
  - *tesseract*
- Automated transcription
  - Transcription in Sociology
  - Challenges
  - *OpenAI Whisper*
  - Diarization
- Where are we right now?

WHAT ARE WE TALKING ABOUT TODAY?

- Last week: how to acquire digital trace data
  - usually: in text format
- Today: data that come in different shape and need to be transformed
  - PDFs or images containing text
  - interviews/videos that contain spoken text

⟹ End goal: text

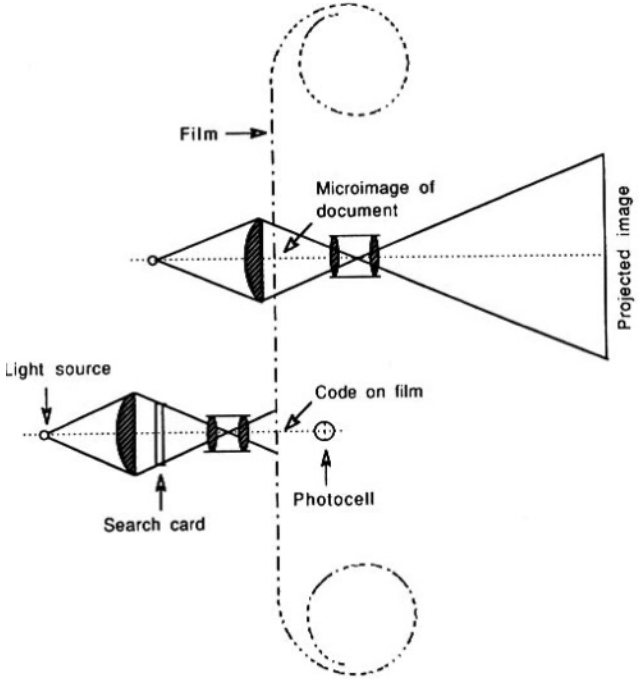
OPTICAL CHARACTER RECOGNITION

- Some documents might not be machine-encoded (yet)
  - Machine-encoded: text that can be edited, searched, displayed online, etc.
- Examples:
  - PDFs downloaded from newspaper archives
  - Book scans (“old” books)
  - Academic papers (if you’re doing “science of science”)

OPTICAL CHARACTER RECOGNITION

Early pioneer: Emanuel Goldberg

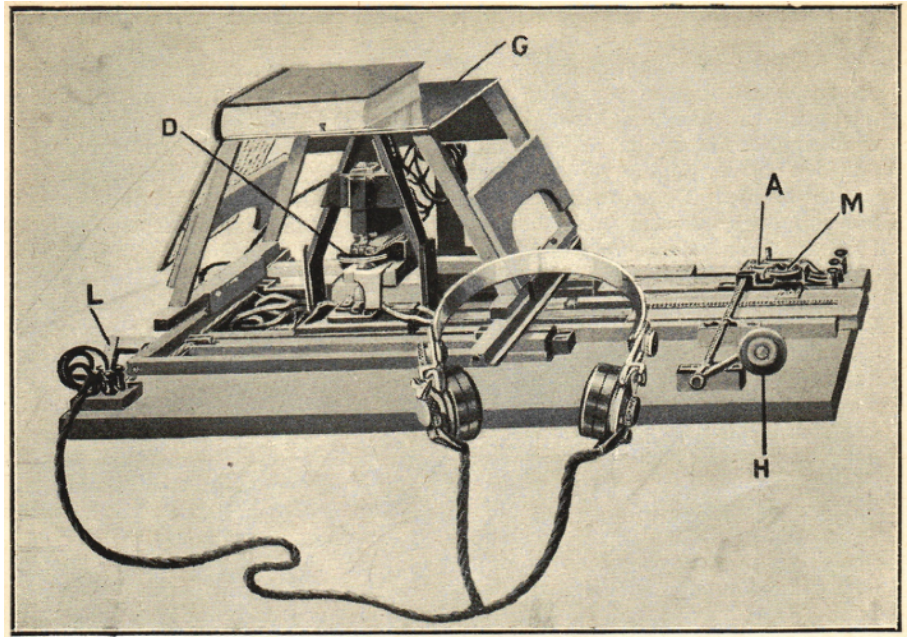
- “Statistical Machine” for document retrieval
- One basically punched letters into a search card
- Light is sent through the card to a photocell
- When pattern in punchcard matches particular pattern in the film, no light reaches the photocell anymore; alarm is triggered



OPTICAL CHARACTER RECOGNITION

Early pioneer: Edmund Fournier d'Albe

- "Optophone"
- Device for blind people:
  - Scans page and makes different sounds based on the letters it "sees"



OPTICAL CHARACTER RECOGNITION

Problem: devices only work with particular fonts

1974: Ray Kurzweil – “omni-font OCR”

- Scanner scanned text
- Mapped characters to sounds, reads out text

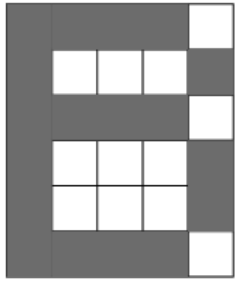
OPTICAL CHARACTER RECOGNITION

Potential problems and strategies to mitigate them:

- Document might be a bit tilted (scans for instance) – deskewing
- There might be some noise (speckles, black dots, etc.) – despeckling
- Text might be in different colors – binarization, make it black and white
- Words and characters need to be separated

⟹ We will use *magick* for this

Input sample character



Digitized character

1	1	1	1	0
1	0	0	0	1
1	1	1	1	0
1	0	0	0	1
1	0	0	0	1
1	1	1	1	0

POST-PROCESSING

- Matching words with lexicons
- Near-neighbor analysis to correct errors – Washington DoC. ⟹ Washington D.C.
  ⟹ Using a pre-trained language model

```
## # A tibble: 60 × 3
##    word  confidence bbox
##    <chr>      <dbl> <chr>
##  1 This        96.8 36,92,96,116
##  2 is          96.9 109,92,129,116
##  3 a           95.7 141,98,156,116
##  4 lot         95.7 169,92,201,116
##  5 of          96.5 212,92,240,116
##  6 12          96.5 251,92,282,116
##  7 point       96.4 296,92,364,122
##  8 text        96.2 374,93,427,116
##  9 to          96.9 437,93,463,116
## 10 test        97.0 474,93,526,116
## # i 50 more rows
```

TESSERACT

- we will use *tesseract*
- Probably the most commonly used software for OCR
- Open source
- Invented by HP (as part of their flatbed scanners)
- Open-sourced in 2005, maintained by Google (2005-2018)
- Needs to be installed via command line, can then be used in R (⟹ Thursday)

TESSERACT

tesseract uses a two-step approach:

- Preprocessing: break up text into components, single out words
- First step: predict words (not characters) using a neural network; recognized words are fed into the classifier and used to predict later text
- Second step: after first scan, the full text is again passed on to the “finished” model and words are predicted again (since it might have just learned some rare words)

TRANSCRIPTION

Another (sort of) common format for text to be stored in: spoken language

For instance:

- interviews
- speeches
- videos

⟹ we *could* use dedicated classifiers for audio data, but…

TRANSCRIPTION

⟹ We *could* use dedicated classifiers for audio data, but…

- Unified format (i.e., text)
- Easier annotation
- Text models work better (less *noise* – literally)
- Easier to interpret
- Transcription “renders the big booming confusion of raw conversation into forms that support inquiry” (Vanover 2022: 64)

WHEN I WAS YOUR AGE…I TOOK A COURSE ON QUALITATIVE SOCIOLOGY

- Transcription by hand – classic approach in qualitative sociology
- Assumption: you immerse yourself in the data while transcribing
  ⟹ “transcription slow[s] the process down and create[s] time for reflexivity and theory-building” (Vanover 2022: 64)
- Research happens *while* you are transcribing
- Strategies: naturalized transcription practices (i.e., adding context to what has been said, remodel the text) vs. atheoretical (word by word)

Version 1: An Atheoretical Transcription

And the reading special came in and just said you can't be in here. You have to go to your class. And that it was like she started sort of getting real direct with him. Telling him no, you are going to your class. You cannot stay in here. And I said and I sort of pulled her aside and I said you know his dad just passed away this morning. He's really calm. I don't mind him being in here. And she was like No, you are being very nice. But that's ridiculous. He has to go to his other class. Whipped him into a frenzy. Just you're going to this class. I'm not going to that class. You're going to this class. I'm not going to that class I mean just really and I stepped back because honestly I did not know what to do.

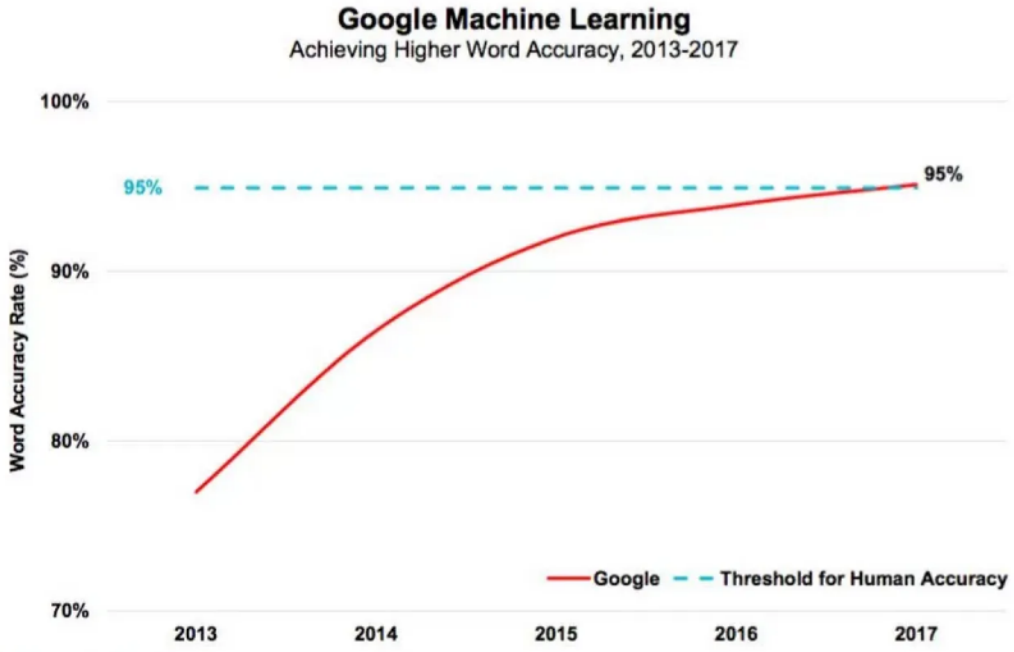
Version 2: A Narrativized Transcription of the Same Interview Segment

And the reading special came in and just said
You can't be in here. You have to go to your class.
And that, it was like, she started sort of getting real direct with him. Telling him,
No, you are going to your class. You cannot stay in here.
And I said—and I sort of pulled her aside and I said
You know, his dad just passed away this morning. He's really calm. I don't mind him being in here.
And she was like
No, you are being very nice. But that's ridiculous. He has to go to his other class
Whipped him into a frenzy. Just
You're going to this class.
I'm not going to that class
You're going to this class
I'm not going to that class
I mean, just really. And I stepped back because honestly I did not know what to do.

WHEN I WAS YOUR AGE…I TOOK A COURSE ON QUALITATIVE SOCIOLOGY

⟹ Time-consuming AF, not suited for “big data” and, thus, hypothesis testing with reasonable sample sizes

…Voice-Based Platform *Back-Ends* = Voice Recognition Accuracy Continues to Improve

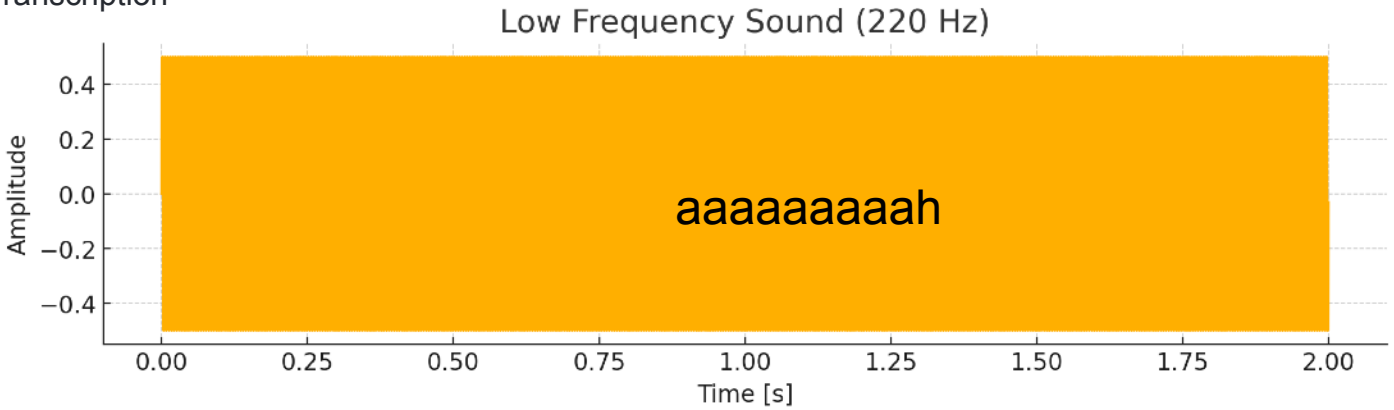


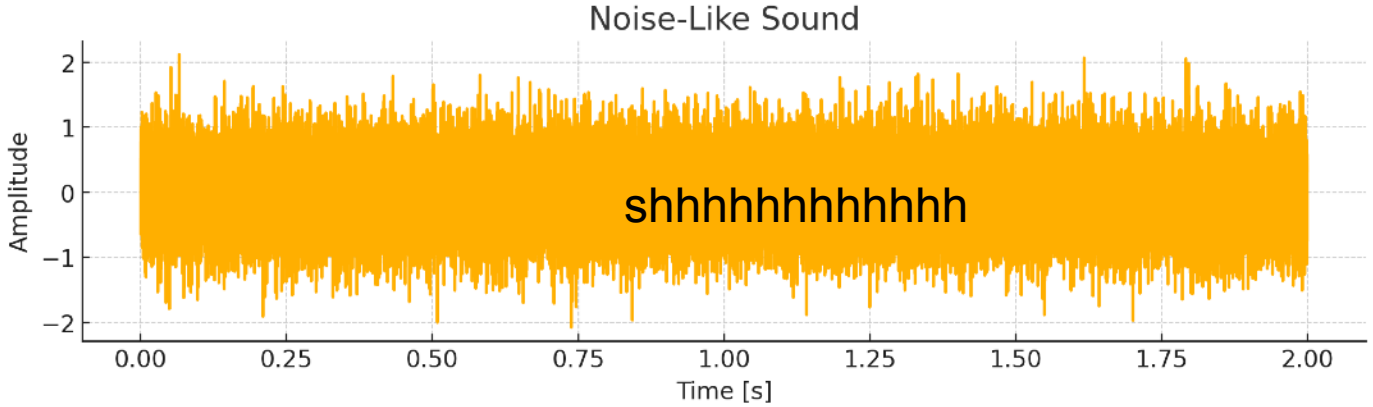
KLEINER PERKINS

Source: Google (5/17)
Note: Data as of 5/17/17 and refers to recognition accuracy for English language. Word error rate is evaluated using real world search data which is extremely diverse and more error prone than typical human dialogue.

HOW TRANSC

- Audio informat
- We sample “sr
- Features are extracted from our raw data
- Goal: *Phonem*
  - “th” sounc /θ/.
  - “ee” soun
- Based on phor
- Post-processin





HOW TRANSCRIPTION WORKS

- Basic prediction task: words ~ phonemes
- Trained on labels (words) ~ speech
- Language is very diverse
- Solution: BIG BIG model that has seen lots of examples

⟹ OpenAI Whisper

WHISPER

- Trained on labels (words) ~ speech
  - 680,000 hours of labeled speech data from the internet
  - 563,000 hours English, 117,000 hours in 96 languages
- Goal: zero-shot model with highest possible accuracy

⟹ zero-shot: no further human input necessary to refine the model

WHISPER

```
# Load Whisper model
whisper_model = whisper.load_model("base")
```

```
transcribe_audio("mlk_ihaveadream.wav")
```

```
transcribe_audio("mlk_ihaveadream.wav")
```

```
 I have the pleasure to present to you Dr. Martin Luther King. Y
eah! I am happy to join with you today in what will go down in h
istory as the greatest demonstration for freedom in the history
of our nation. Five score years ago, a great American in whose s
ymbolic shadow we stand today, signed the emancipation proclamat
ion. This momentous decree came as a great beacon light of hope
to millions of Negro slaves who had been seared in the flames of
withering injustice. It came as a joyous daybreak to end the lon
g night of their captivity. But one hundred years later, the Neg
ro still is not free. One hundred years later, the life of the N
egro is still sadly crippled by the manacles of segregation and
the chains of discrimination one hundred years later. The Negro
lives on a lonely island of poverty in the midst of a vast ocean
```

```
transcribe_audio("mlk_ihaveadream.wav")
```


```
'segments': [{'id': 0,
  'seek': 0,
  'start': 0.0,
  'end': 4.32,
  'text': ' I have the pleasure to present to you Dr. Martin Luther King.',
  'tokens': [50364,
   286,
   362,
   264,
   6834,
   281,
   1974,
   281,
   291,
   2491,
   13,
   9184,
   20693,
   3819,
   13,
   50580],
  'temperature': 0.0,
  'avg_logprob': -0.45925567263648626,
  'compression_ratio': 1.2521008403361344,
  'no_speech_prob': 0.10073500126600266},
```

```
transcribe_audio("mlk_ihaveadream.wav")
```

```
 I have the pleasure to present to you Dr. Martin Luther King. Y
eah! I am happy to join with you today in what will go down in h
istory as the greatest demonstration for freedom in the history
of our nation. Five score years ago, a great American in whose s
ymbolic shadow we stand today, signed the emancipation proclamat
ion. This momentous decree came as a great beacon light of hope
to millions of Negro slaves who had been seared in the flames of
withering injustice. It came as a joyous daybreak to end the lon
g night of their captivity. But one hundred years later, the Neg
ro still is not free. One hundred years later, the life of the N
egro is still sadly crippled by the manacles of segregation and
the chains of discrimination one hundred years later. The Negro
lives on a lonely island of poverty in the midst of a vast ocean
```

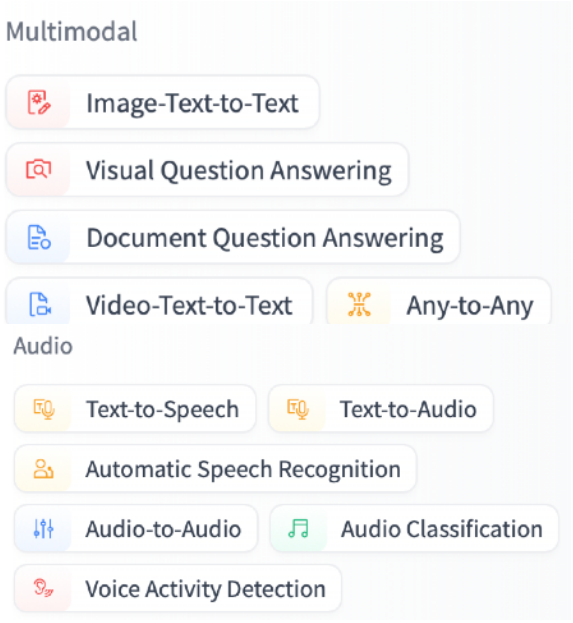
DIARIZATION

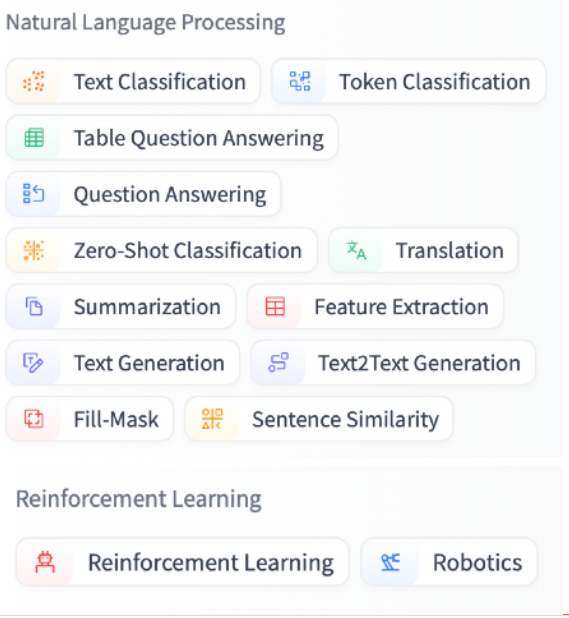
- Often in audio files there might be different speakers
  - e.g., interviews, talk shows, etc.
- We would like to distinguish these speakers, attribute their speech to them
- This process is called *diarization*
- In Python, we can use the *pyannote.audio* package

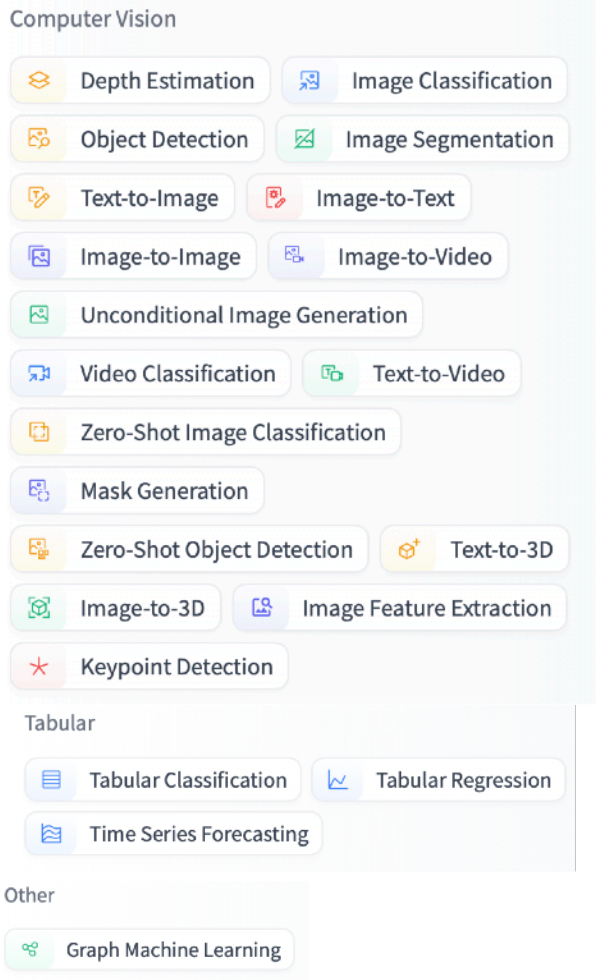

ohne_gleitgel_2
Und ~~DVD~~ Diarization kann der auch?
~~Blu-ray~~ pyannote, zero-shot
Ach hör doch auf
@ohne_gleitgel
348
3

- In Python, we can use the *pyannote.audio* package

⟹ like whisper, it uses a pre-trained model – which you can download from 🤗 **Hugging Face**







DIARIZATION

The pipeline then lo

- read in file
- separate speaker
- transcribe their te
- create one tibble

```
[{'speaker': 'SPEAKER_00',
  'start': 345.0,
  'end': 349.0,
  'text': ' 1963 is not an end, but a beginning.'},
 {'speaker': 'SPEAKER_00',
  'start': 137.0,
  'end': 139.0,
  'text': " so we've come here today."},
 {'speaker': 'SPEAKER_00',
  'start': 337.0,
  'end': 338.0,
  'text': ' will not pass.'},
 {'speaker': 'SPEAKER_00',
  'start': 623.0,
  'end': 624.0,
  'text': ' is redemption.'},
```

SOME THINGS TO BEAR IN MIND

- These things require plenty of (GPU) power
- So this week's script will also cover some Python/SC basics
- In particular:
  - General workflow (modules, environments, scripts)
  - *slurm* jobs

⟹ THIS DOES NOT MEAN THAT YOU HAVE TO WORK WITH THESE THINGS; HOWEVER, THEY EXIST IF YOU NEED THEM

DONE FOR TODAY

- How’s it going?
- Anything you need more information on (in terms of acquisition)?
- Has it been useful thus far?
- Overwhelming as hell?

UNIVERSITÄT LEIPZIG

MERCI!

Felix Lennert
Institut für Soziologie

felix.lennert@uni-leipzig.de
www.uni-leipzig.de