Evaluating Automatic Speech Recognition (ASR) for Social Science Research

A Comparison of Semantic Metrics

Peter Kannewitz (presenting), Nicolas Ruth, Andreas Niekler, Stephan Poppe, Leonie Steinbrinker

aisicresearch.github.io

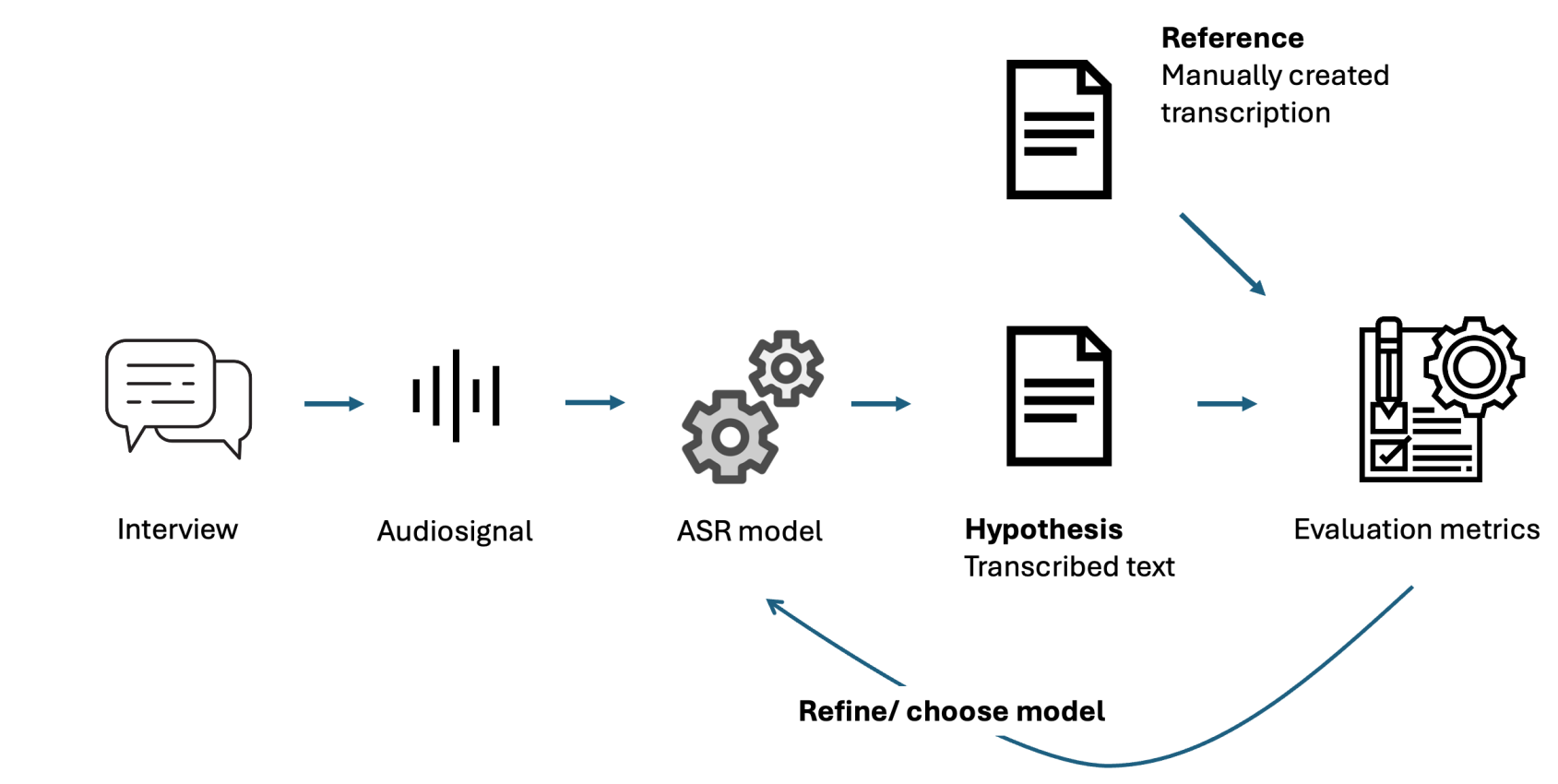
18.07.2025

ASR as key technology in the field

- Our project: AI-SIC, AI Enhanced Validation of Survey Instruments[1]
  - Goal: (Semi-) automated transcription and labeling of interviews with german speaking children
- ASR is a promising technology to open up new data spaces

[1] Funded by the DFG as part of the "New Data Spaces for the Social Sciences" Program

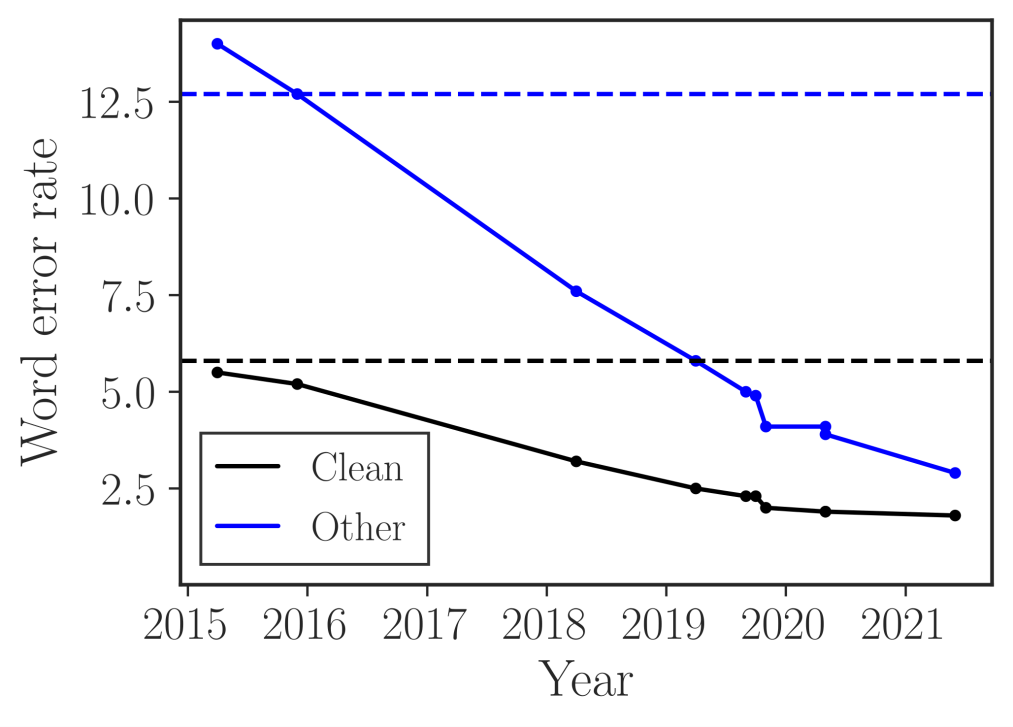
Our ASR Pipeline



Own Illustration.

Where we are

Word Error Rate (WER) for LibreSpeech (Read English Speech)



Source: `https://awni.github.io/future-speech/`.

Current Limitations

WER (%) on VoxPopuli corpus for selected languages.

(Radford et al. 2022, p. 23)

Model	**Czech**	**German**	**English**	*en_accented*	**...**
Whisper tiny	73.5	27.4	11.6	18.8	...
Whisper base	54.7	20.6	9.5	17.5	...
Whisper small	28.8	14.8	8.2	19.2	...
Whisper medium	18.4	12.4	7.6	19.1	...
Whisper large	15.9	11.9	7.2	20.8	...
Whisper large-v2	12.6	11.2	7.0	18.6	...

Applying ASR in the Social Sciences

AIS|C

How can we know if ASR errors are acceptable for our task?

Word Error Rate

$$WER = \frac{(Substitution + Deletion + Insertion)}{\text{Total Number of Words}}$$

System	Transcription	WER (%)
Reference	Find me flights to London	0.0
ASR 1	Find the flights to London	20.0
ASR 2	Find me flights to Lisbon	20.0

WER calculated by equal weighting of word errors.

Ways to construct more meaningful evaluations

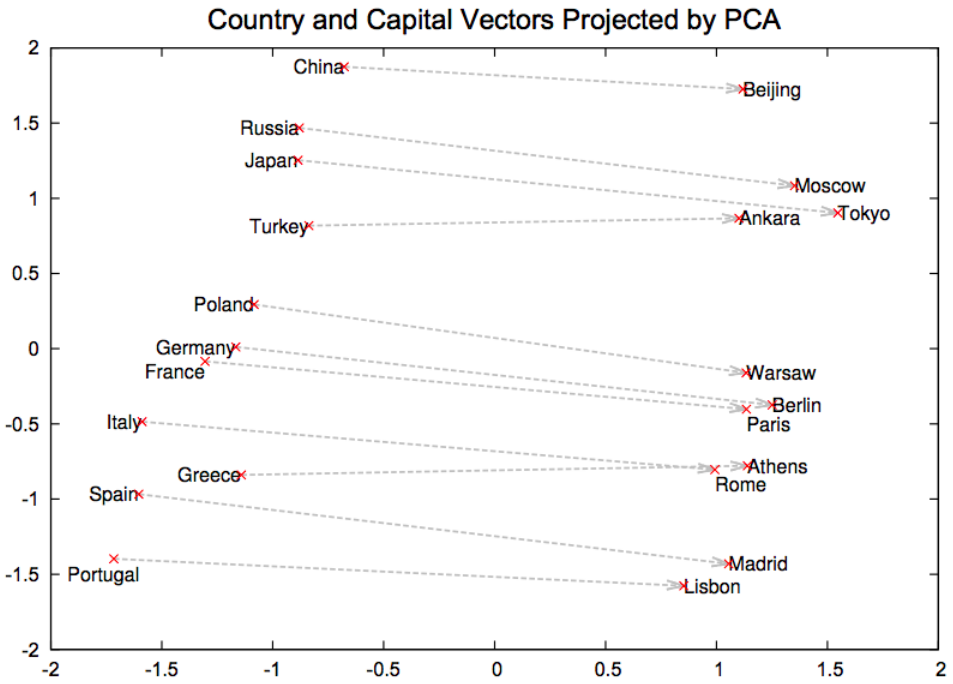
1. Weighted WER
2. Semantically based Error Rates
3. Window based evaluation

Strategy 1: Weighted WER

- **Semantic-WER** (Roy 2021)
  - rule based weights for S, D, I
  - example rules for substitution:
    - assign high error if reference is a named entity (e.g. London)
    - assign low error if words are similiar to each other
- **EmbER** Embedding Error Rate (Roux et al. 2022)

Strategy 2: Semantically based Error Rates

- **BERTScore** (Zhang et al. 2020)
- **SemDist** (Kim et al. 2021)
- **Aligned Semantic Distance (ASD)** (Rugayan et al. 2023)
- **SeMaScore** (Sasindran et al. 2024)

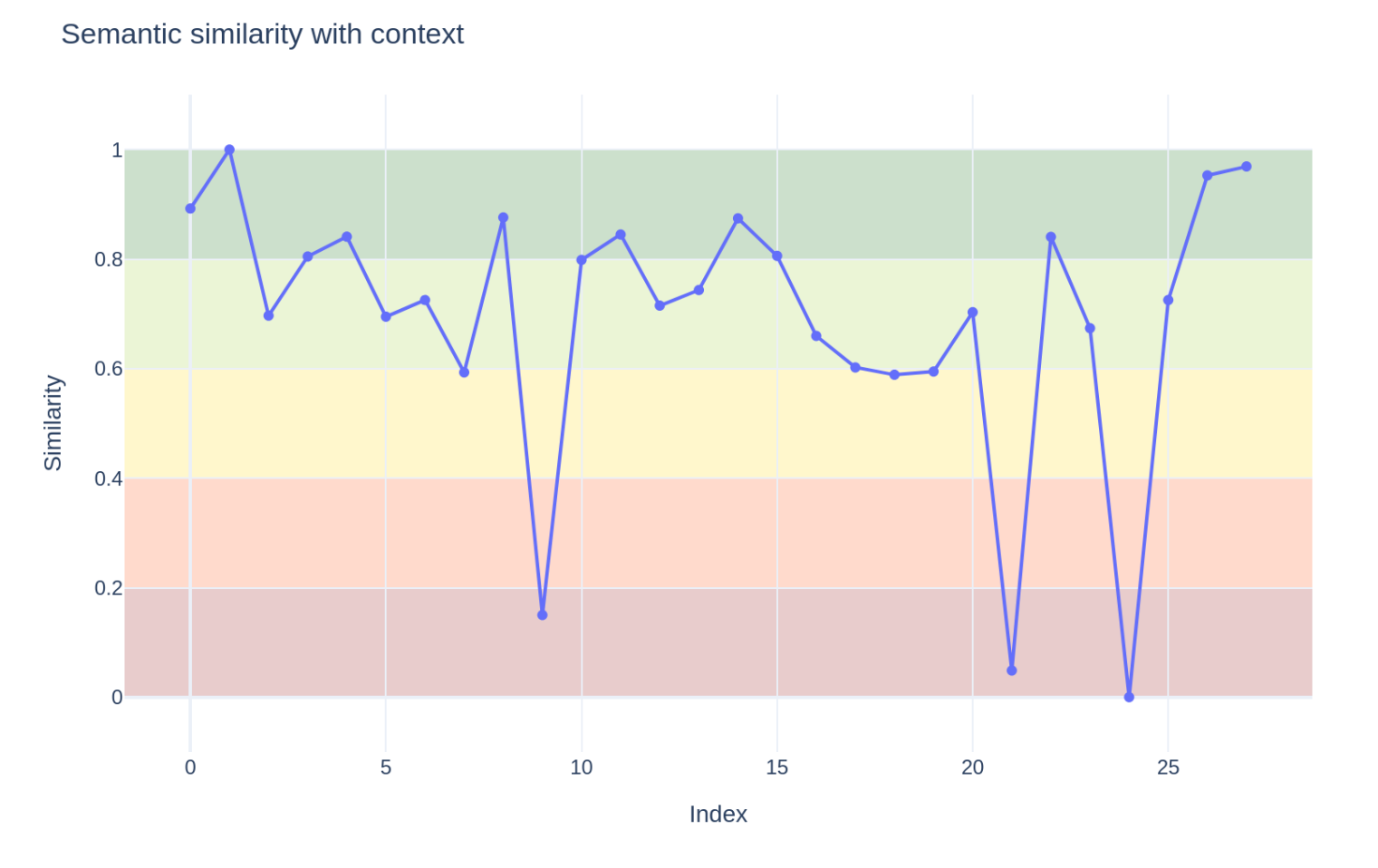


Source: `https://lovit.github.io/assets/figures/word2vec_country_capital.png`.

Word Embeddings Concept

Visually aided Semantic Evaluation

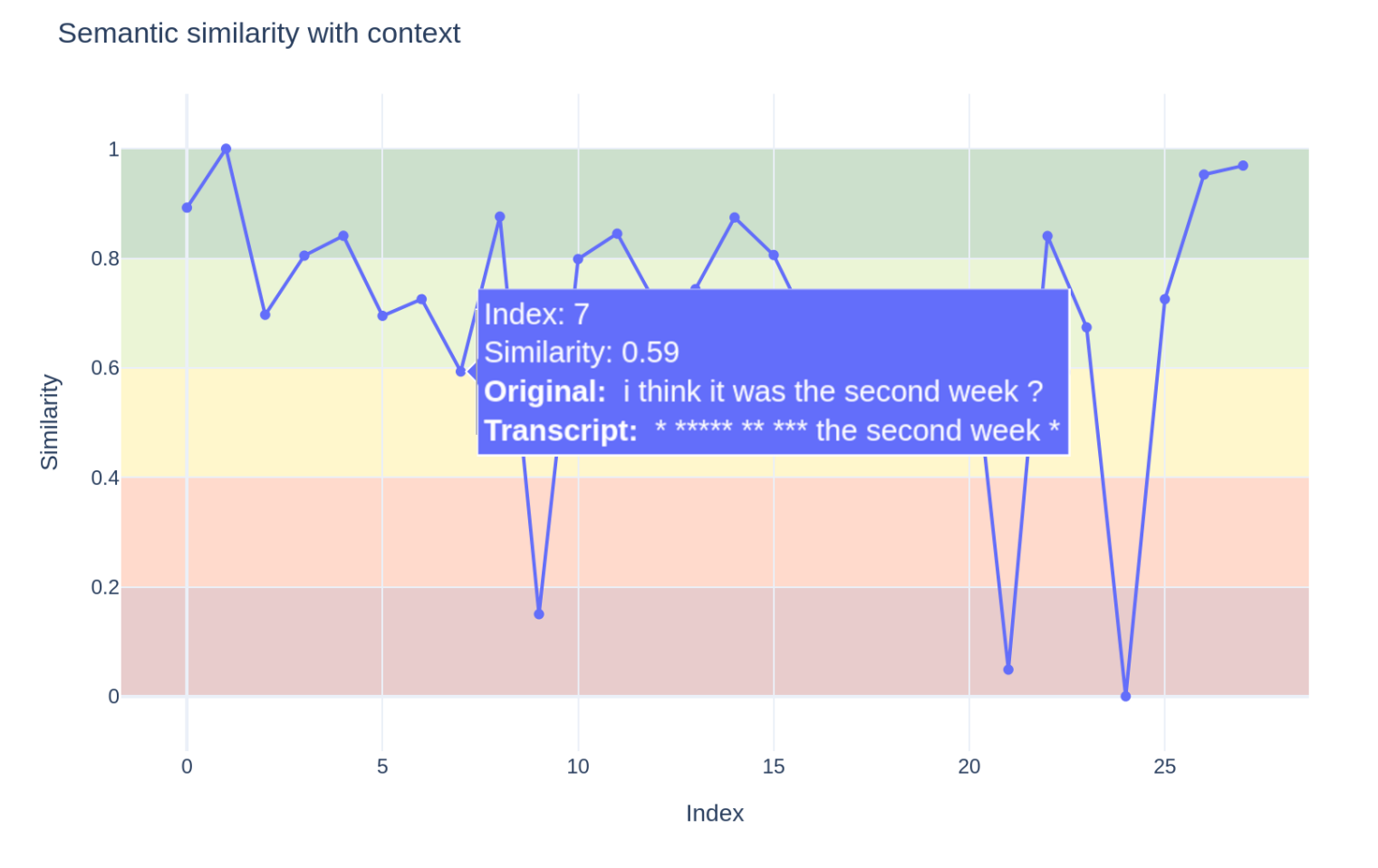
- Our proposal: Interactive analysis of errors[2]



[2] https://github.com/aisicresearch/semantic-asr-evaluation

Visually aided Semantic Evaluation

- Our proposal: Interactive analysis of errors[2]



[2] https://github.com/aisicresearch/semantic-asr-evaluation

Lessons Learned

AIS|C

There is no one-size-fits-all method/criteria/metric!

What we propose for evaluating ASR in the Social Sciences:

1. Evaluate according to your task
2. Combine different error metrics
3. Interactively analyse classes of errors

References

Kim, Suyoun, Abhinav Arora, Duc Le, Ching-Feng Yeh, Christian Fuegen, Ozlem Kalinli, and Michael L. Seltzer (2021). **"Semantic Distance: A New Metric for ASR Performance Analysis Towards Spoken Language Understanding".** In: *Interspeech 2021*. ISCA, pp. 1977–1981.

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever (2022). **Robust Speech Recognition via Large-Scale Weak Supervision.** arXiv:2212.04356 [eess].

Roux, Thibault Bañeras, Mickael Rouvier, Jane Wottawa, and Richard Dufour (2022). **"Qualitative Evaluation of Language Model Rescoring in Automatic Speech Recognition".** In: *Interspeech 2022*. ISCA, pp. 3968–3972.

Roy, Somnath (2021). **Semantic-WER: A Unified Metric for the Evaluation of ASR Transcript for End Usability.** arXiv:2106.02016.

Rugayan, Janine, Giampiero Salvi, and Torbjørn Svendsen (2023). **“Perceptual and Task-Oriented Assessment of a Semantic Metric for ASR EvaluatiPerceptual and Task-Oriented Assessment of a Semantic Metric for ASR Evaluationon”.** In: *INTERSPEECH 2023*. ISCA, pp. 2158–2162.

Sasindran, Zitha, Harsha Yelchuri, and T. V. Prabhakar (2024). **“SeMaScore: A new evaluation metric for automatic speech recognition tasks”.** In: *Interspeech 2024*. ISCA, pp. 4558–4562.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi (2020). **BERTScore: Evaluating Text Generation with BERT.** arXiv:1904.09675.