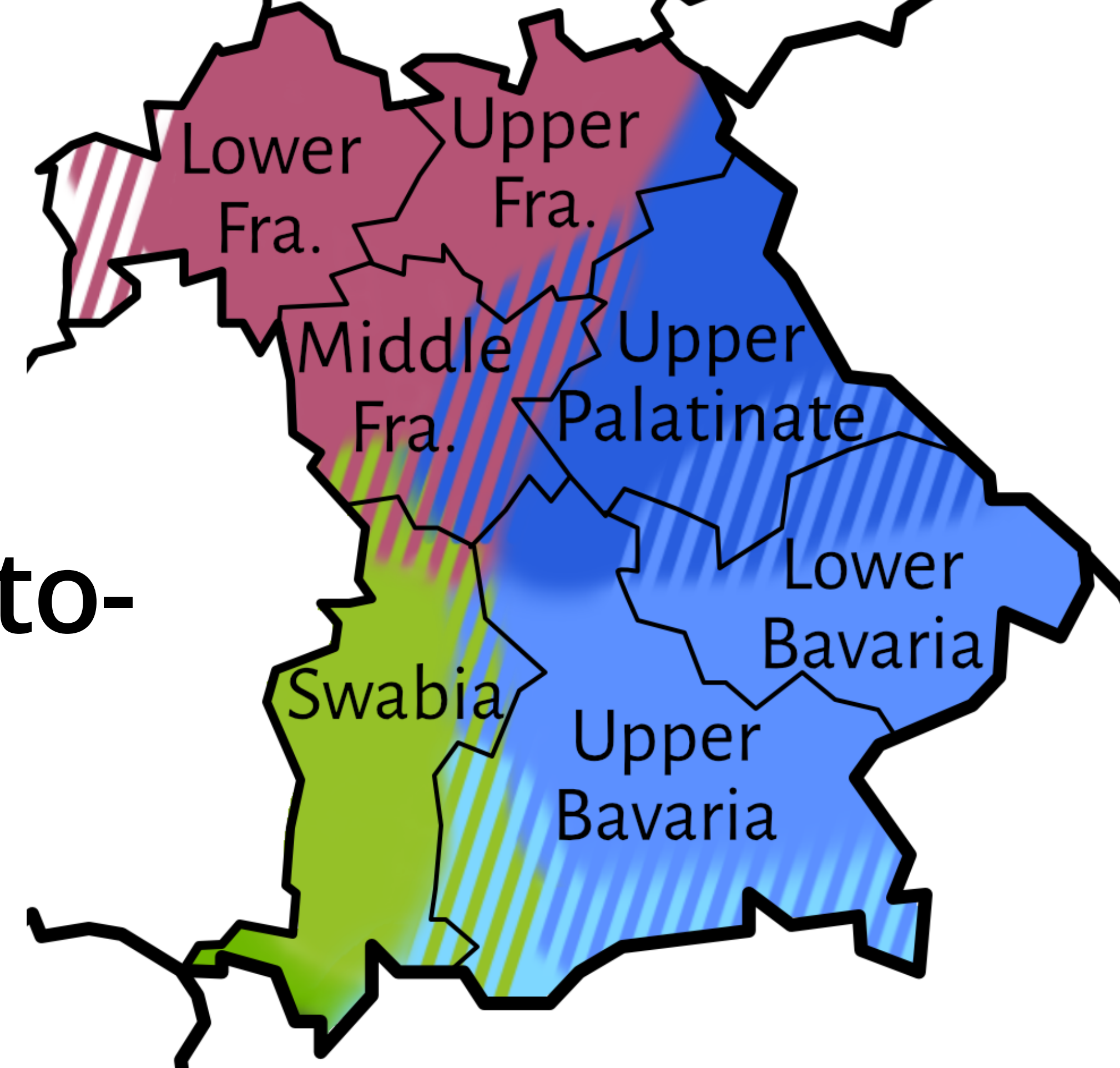


# A multi-dialectal dataset for German dialect ASR & dialect-to-standard speech translation

Verena Blaschke, Miriam Winkler, Constantin Förster,  
Gabriele Wenger-Glemser & Barbara Plank

LMU Munich, MCML & Bayerischer Rundfunk  
Interspeech | August 18, 2025



# Introduction

Automatic **speech** recognition



typically: high-resource languages

 standardized varieties of

# Introduction

Automatic speech recognition



typically: high-resource languages

standardized varieties of

**Can We Better Understand Swiss German? An Automatic, Quantitative Human Evaluation**

**State-of-the-Art ASR Models to Swiss German Dialects**

Victor Gillioz

...towards dialect-inclusive recognition in a low-resource

are balanced corpora the

...ergan, <sup>2</sup>Mengjie Qian, <sup>1</sup>Neasa Ní Chiaráin, <sup>1</sup>

**Dialectal Coverage And Generalization in Arabic Speech Recognition**

Amirbek Djanibekov<sup>1\*</sup>, Hawau Olamide Toyin<sup>1\*</sup>  
Raghad Alshalan<sup>2</sup> Abdullah Alitr<sup>2</sup> Hanan Aldarmaki<sup>1</sup>

**A Corpus of Read and Spontaneous Upper Saxon German for ASR Evaluation**

**Voices Unheard: NLP Resources and Models for Yorùbá Regional Dialects**

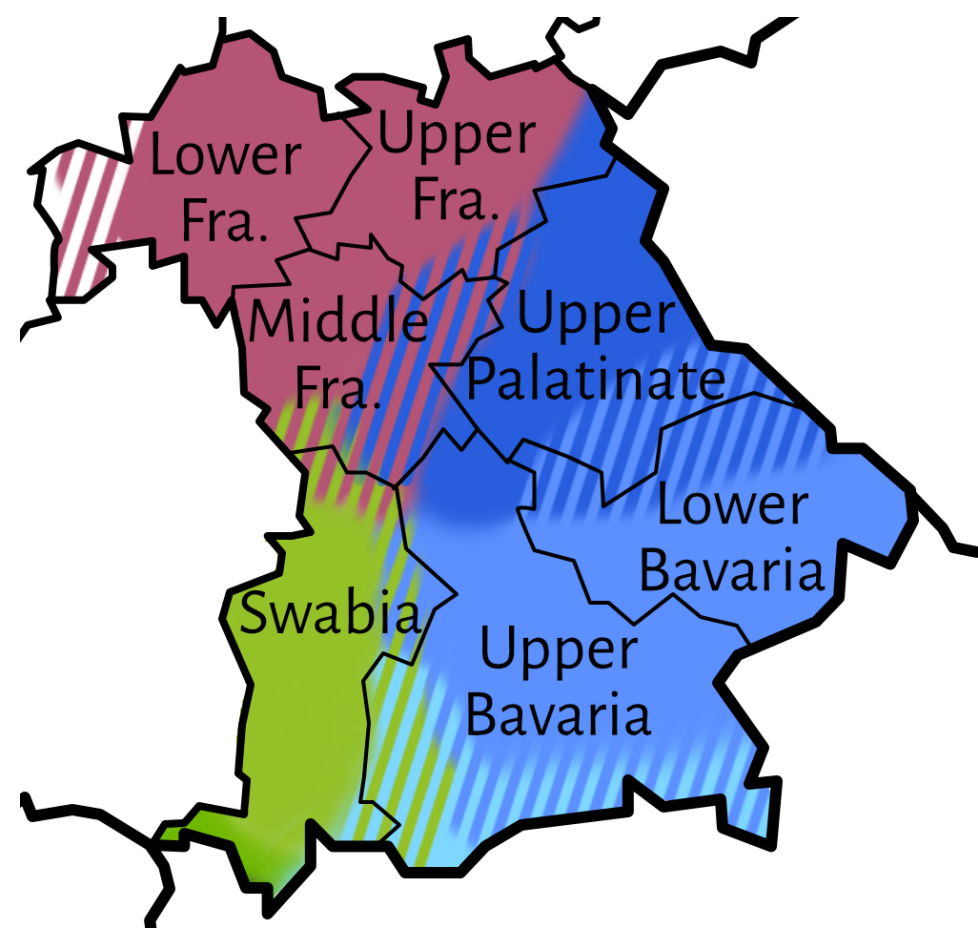
Orevaoghene Ahia<sup>1,5</sup> Anuoluwapo Aremu<sup>5,6</sup> Diana Abagyan<sup>1</sup> Hila Gonen<sup>1</sup>  
Adelani<sup>3,4,6</sup> Daud Abolade<sup>6</sup> Noah A. Smith<sup>1,2</sup> Yulia Tsvetkov<sup>1</sup>

**Speech Recognition for Greek Dialects: A Challenging Benchmark**

Socrates Vakirtzian<sup>\*1</sup>, Chara Tsoukala<sup>\*2,3</sup>, Stavros Bompolas<sup>3</sup>, Katerina Mouzou<sup>1</sup>, Vivian Stam



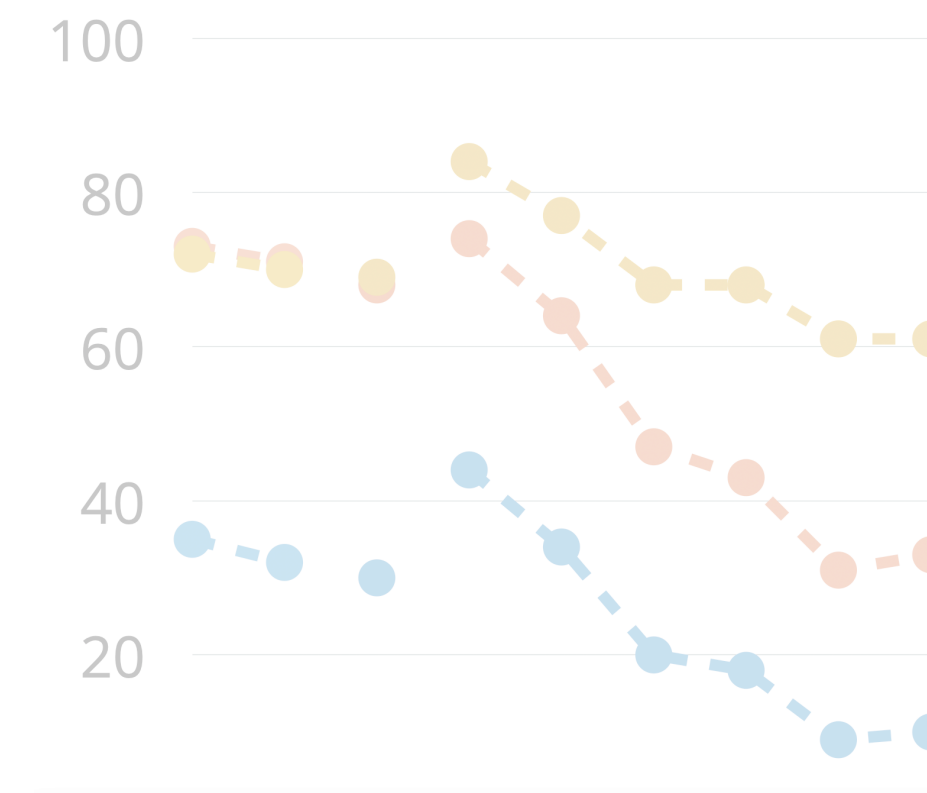
# Overview



1. Dialects in Bavaria



2. *Betthupferl* dataset



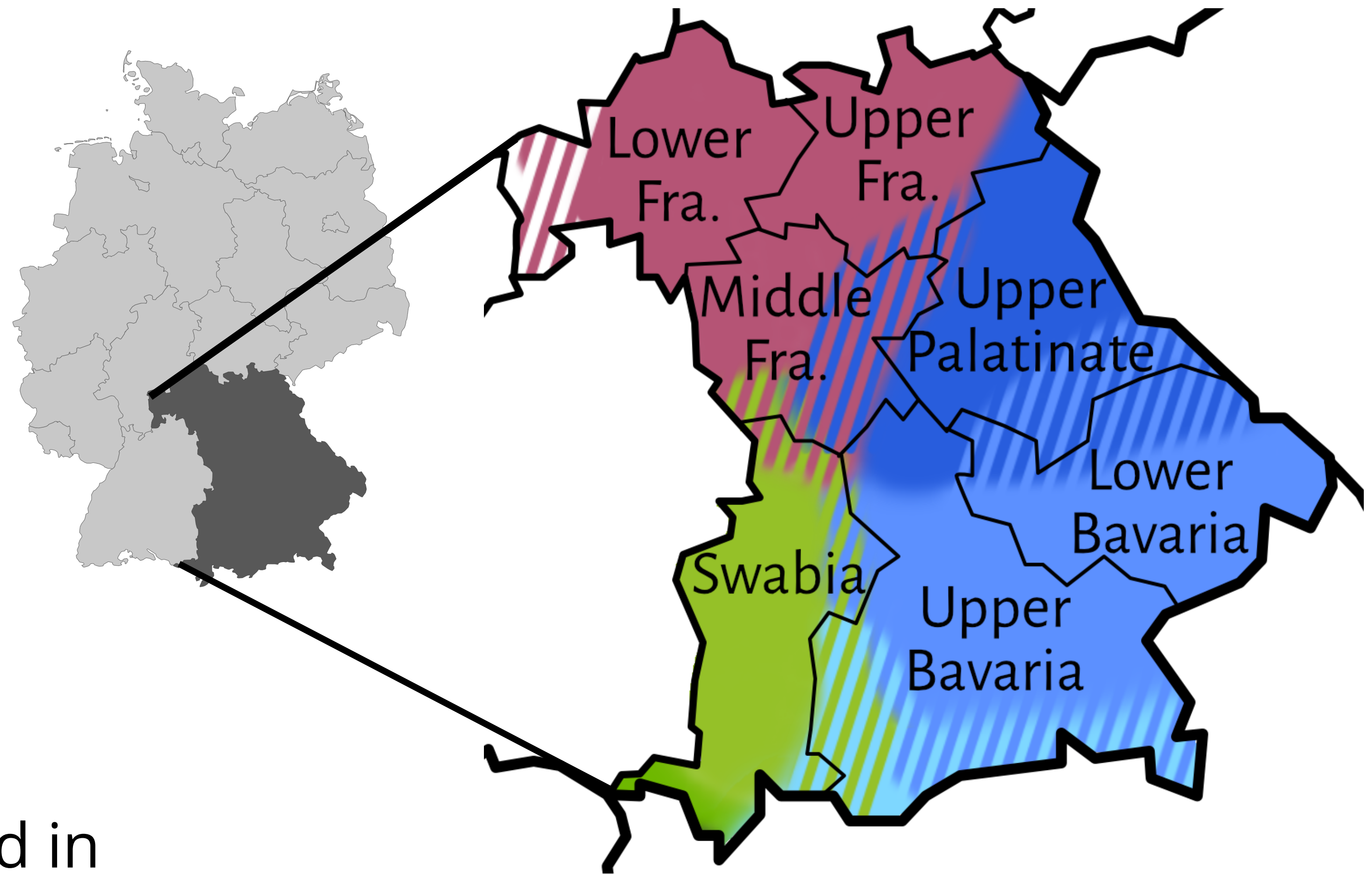
3. Benchmarking ASR models



4. Qualitative analysis

# Dialects in Bavaria

- 3 dialect groups
- Mostly spoken, occasionally written (no orthography)
- Dialect speakers are interested in ASR systems with dialectal and especially with German output



Franconian

● East Franconian

Alemannic

● Swabian

Bavarian

● North Bavarian

● Central Bavarian

● South Bavarian

# Differences between German + dialects

German & dialectal transcriptions of a Franconian sentence:

*"Immediately, search for Mathilda's coin or I'll show you what's what!"*

[German]	Sofort	Mathildas	Geldstück	suchen,...
	<i>Immediately</i>	<i>Mathilda's</i>	<i>coin</i>	<i>search</i>
[Dialect]	Sofort	da Mathilda ihr	Geldstückle	sung, ...
		<i>the Mathilda her</i>		

Small word-level differences (morphology and/or pronunciation)

Different words/phrases

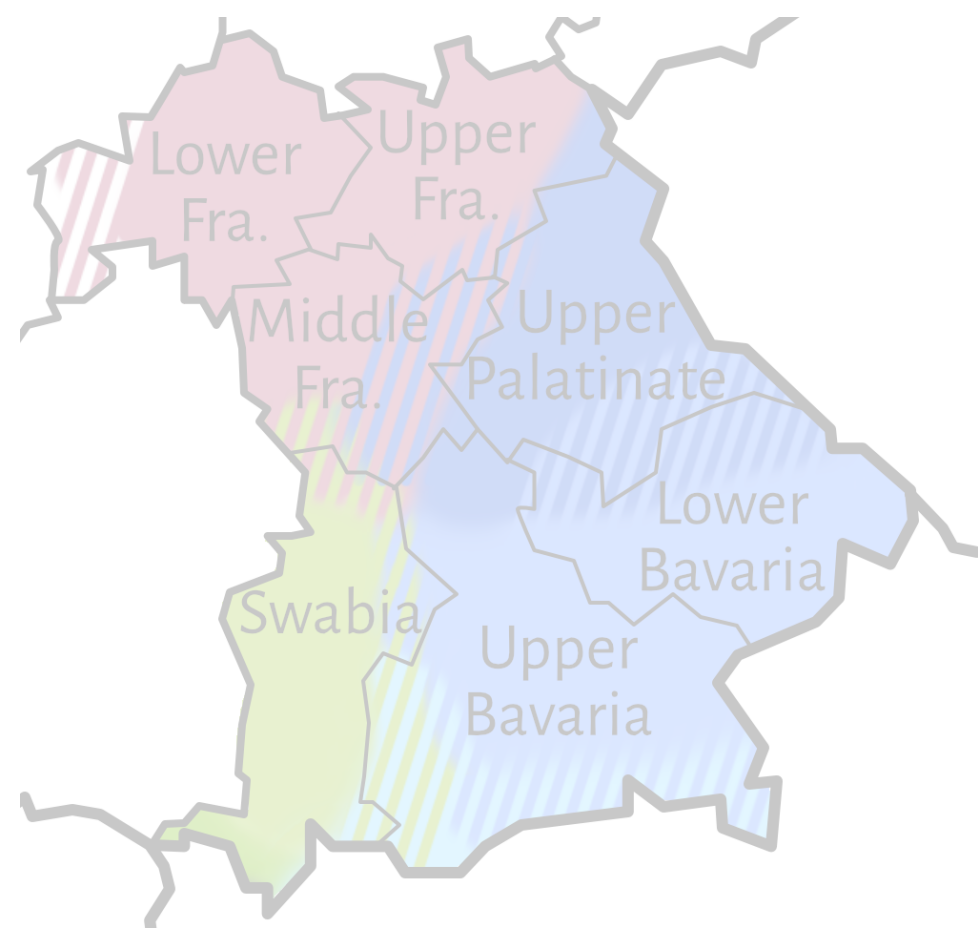
Syntactic differences

- determiner + name
- possessives
- ...

... sonst zeige ich euch,	wo's	langgeht
<i>else show I you</i>	<i>where it runs along.</i>	
... sonst zach ich eich,	wo	da Bartl an Most hoid.
	<i>where the Barthel the cider fetches.</i>	



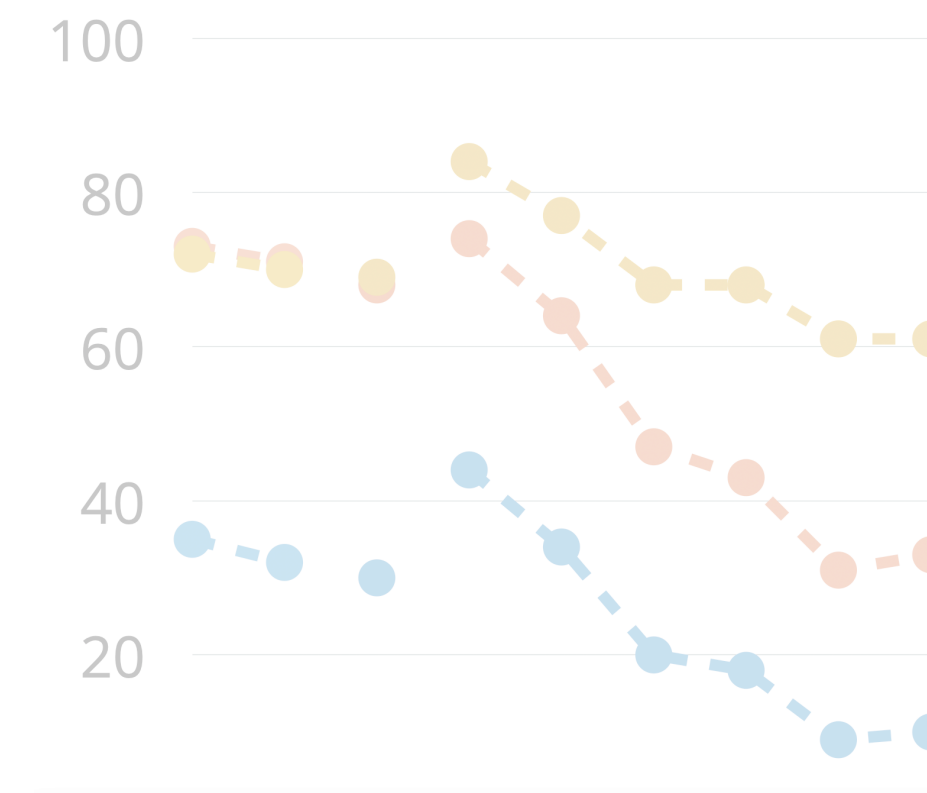
# Overview



1. Dialects in Bavaria



2. *Betthupferl* dataset



3. Benchmarking ASR models

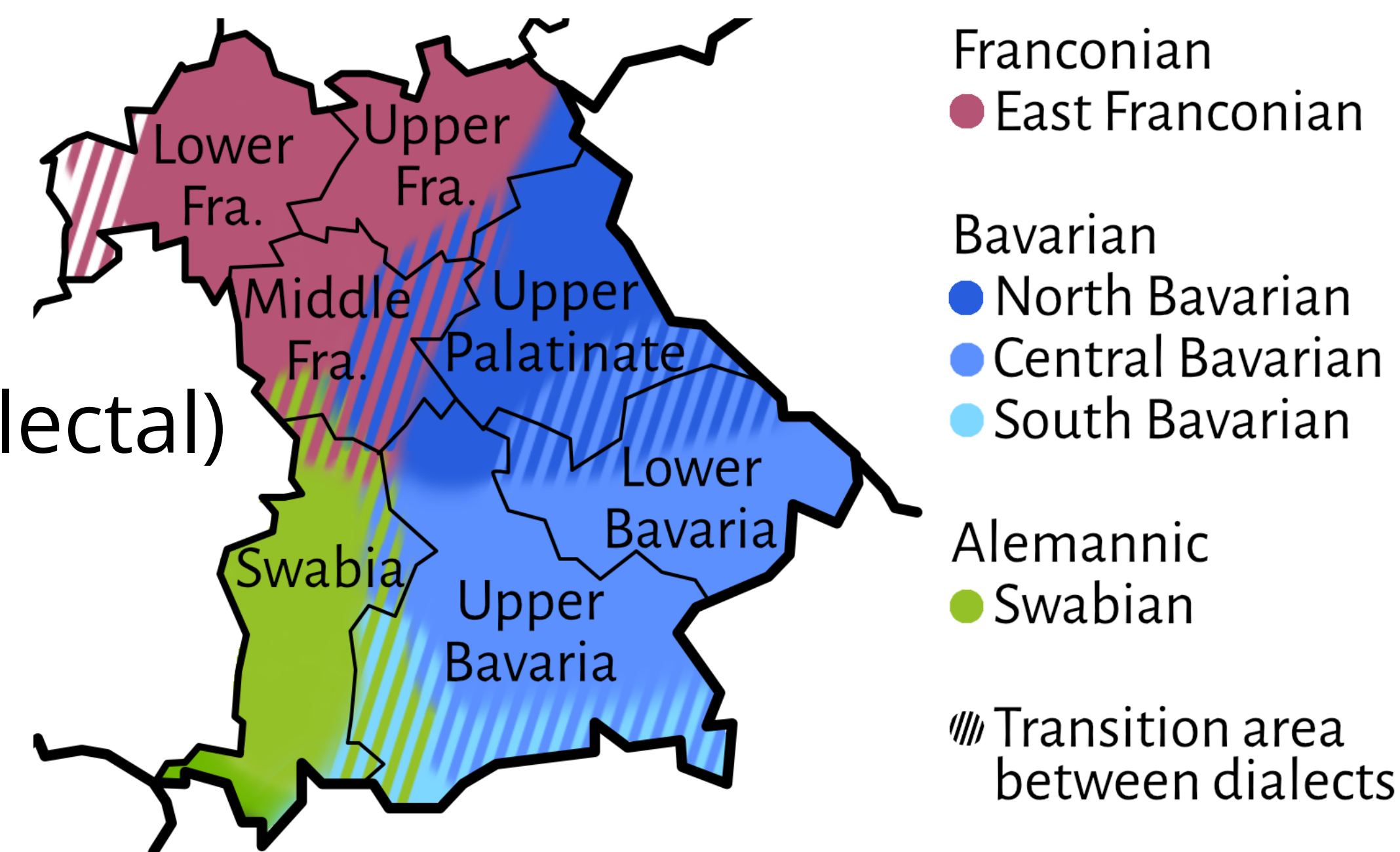


4. Qualitative analysis

# *Betthupferl* dataset

## Data

- Good-night stories for children in German or dialects
- Read speech; professionally written & recorded
- 32–37 mins per administrative region (dialectal)
- 32 mins (Standard) German audio
- Total: 4.5 h



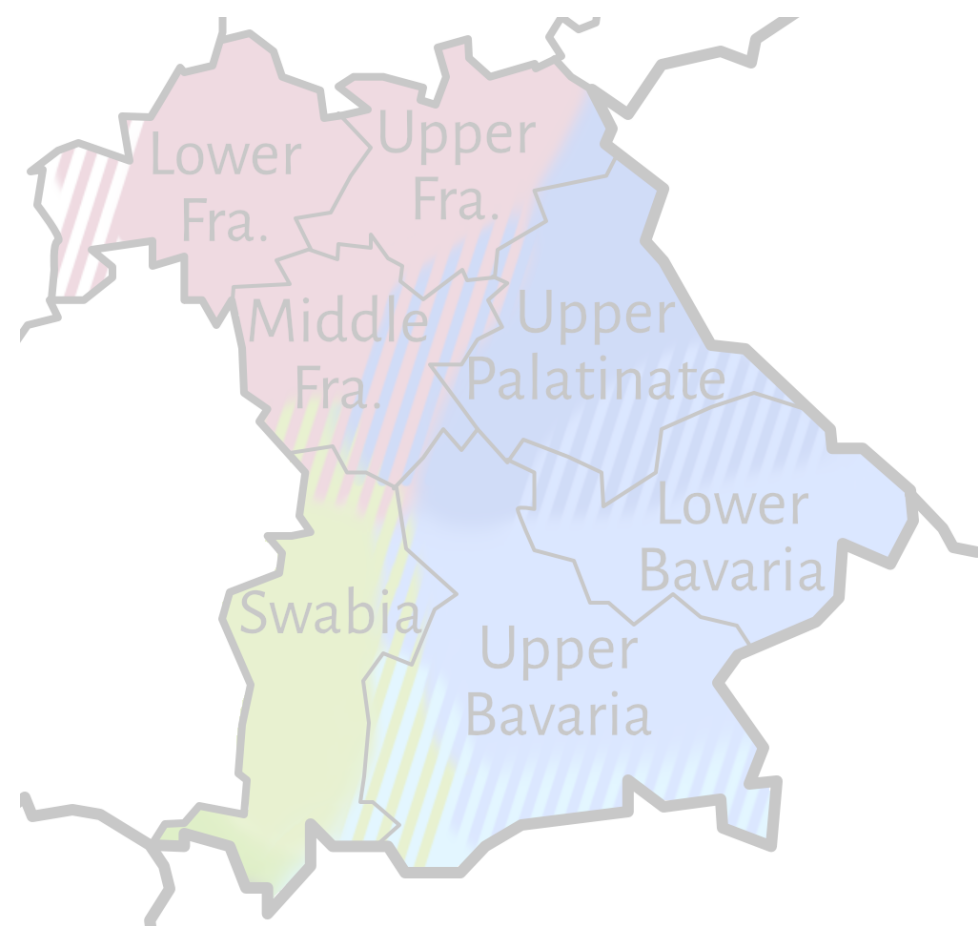


# *Betthupferl* dataset

## Transcriptions

- Sentence level (~4.3 s; 11–12 words)
- 1 dialectal & 1 German transcription per sentence
- Transcriber = native speaker of a Bavarian dialect & German

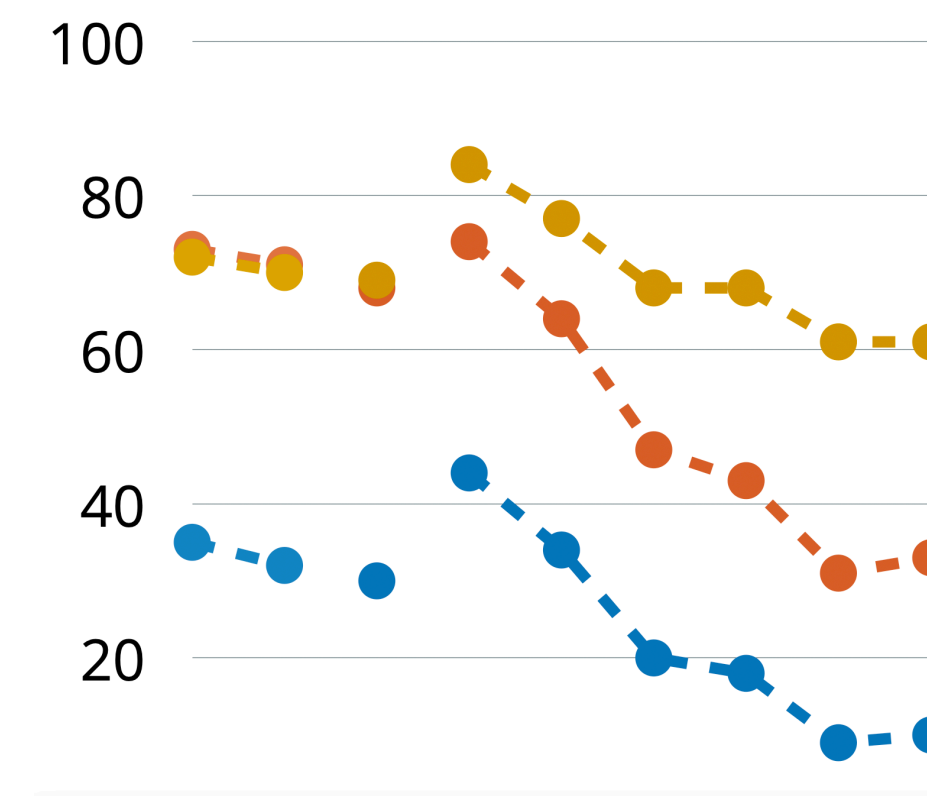
# Overview



1. Dialects in Bavaria



2. *Betthupferl* dataset



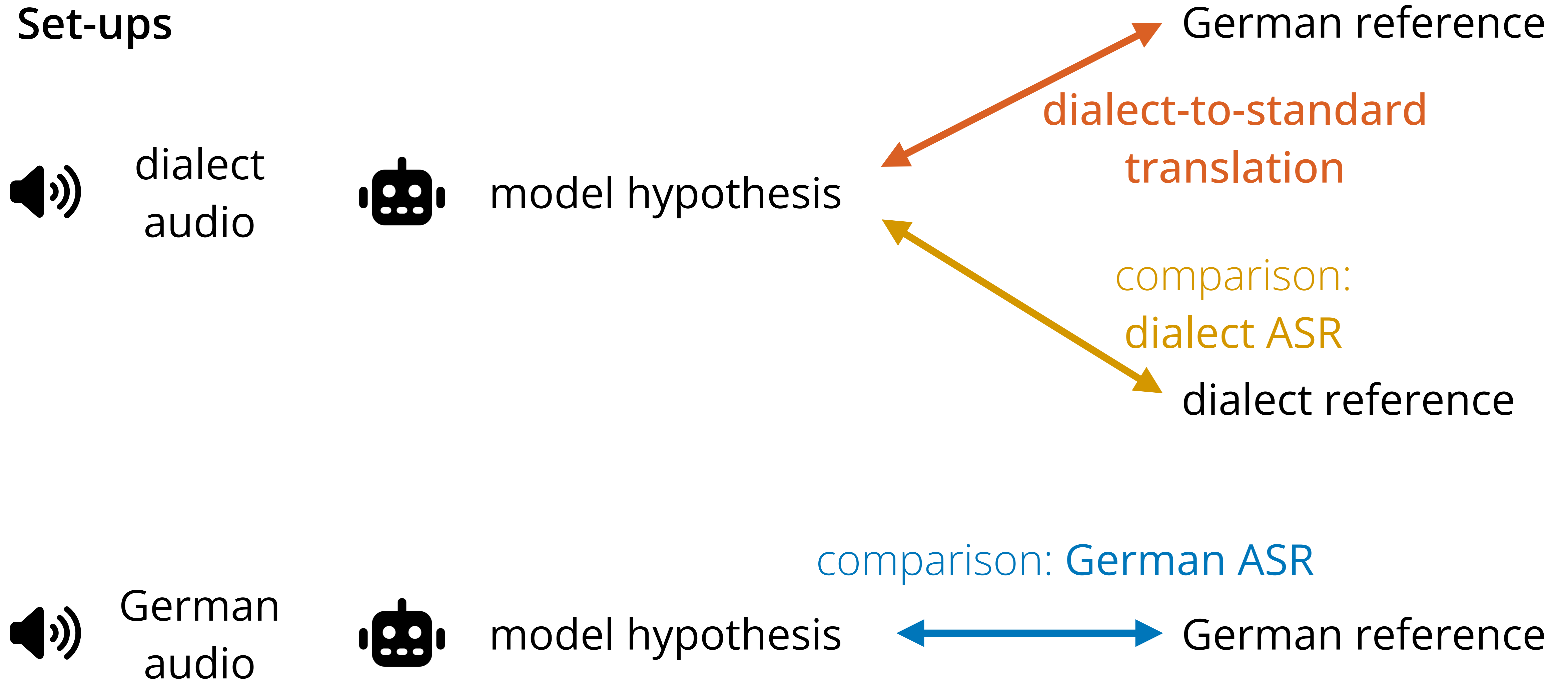
3. Benchmarking ASR models



4. Qualitative analysis

# Experiments

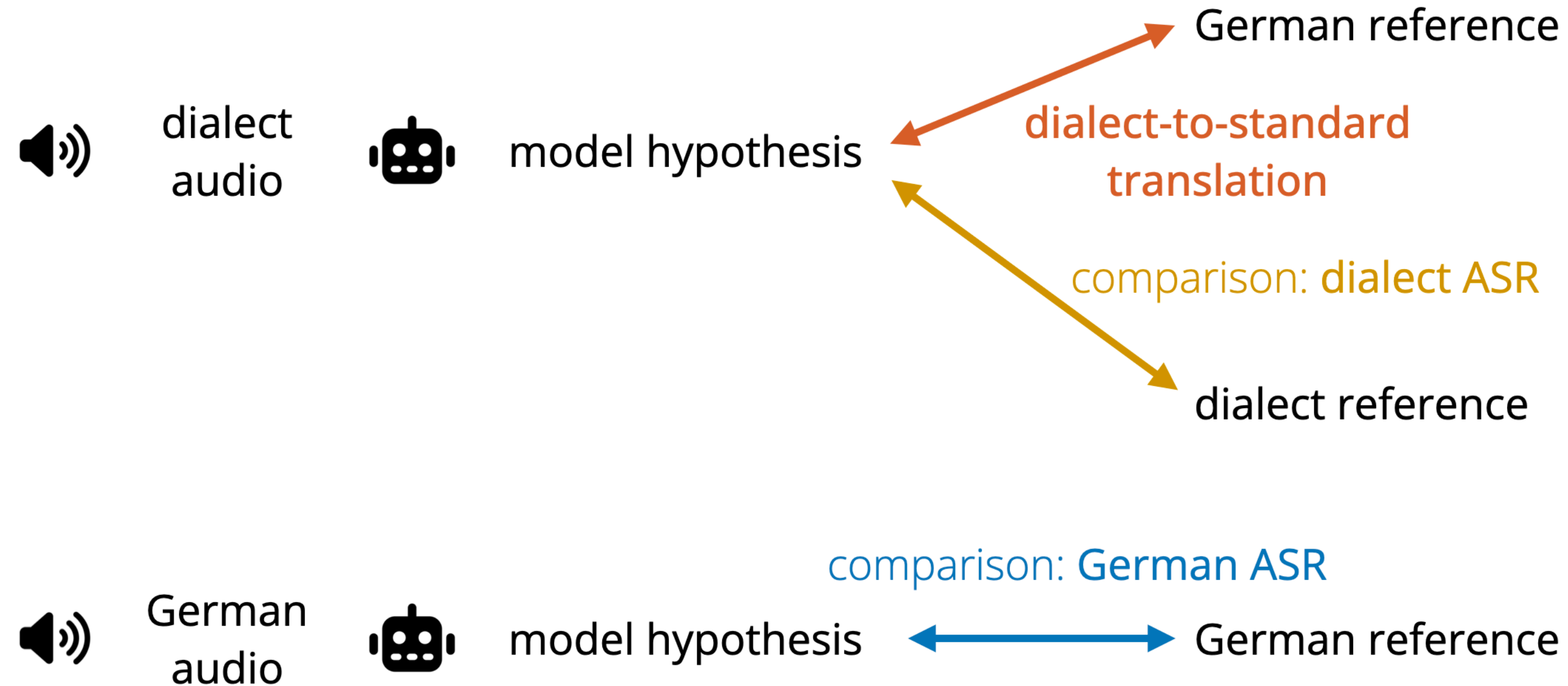
## Set-ups





# Experiments

## Metrics



- CER – spelling differences between standard & dialect
- WER, BLEU – lexically/structurally similar outputs desired,  
also for translation!  
*(BLEU only in paper)*

# Experiments

## Models

### Architectures

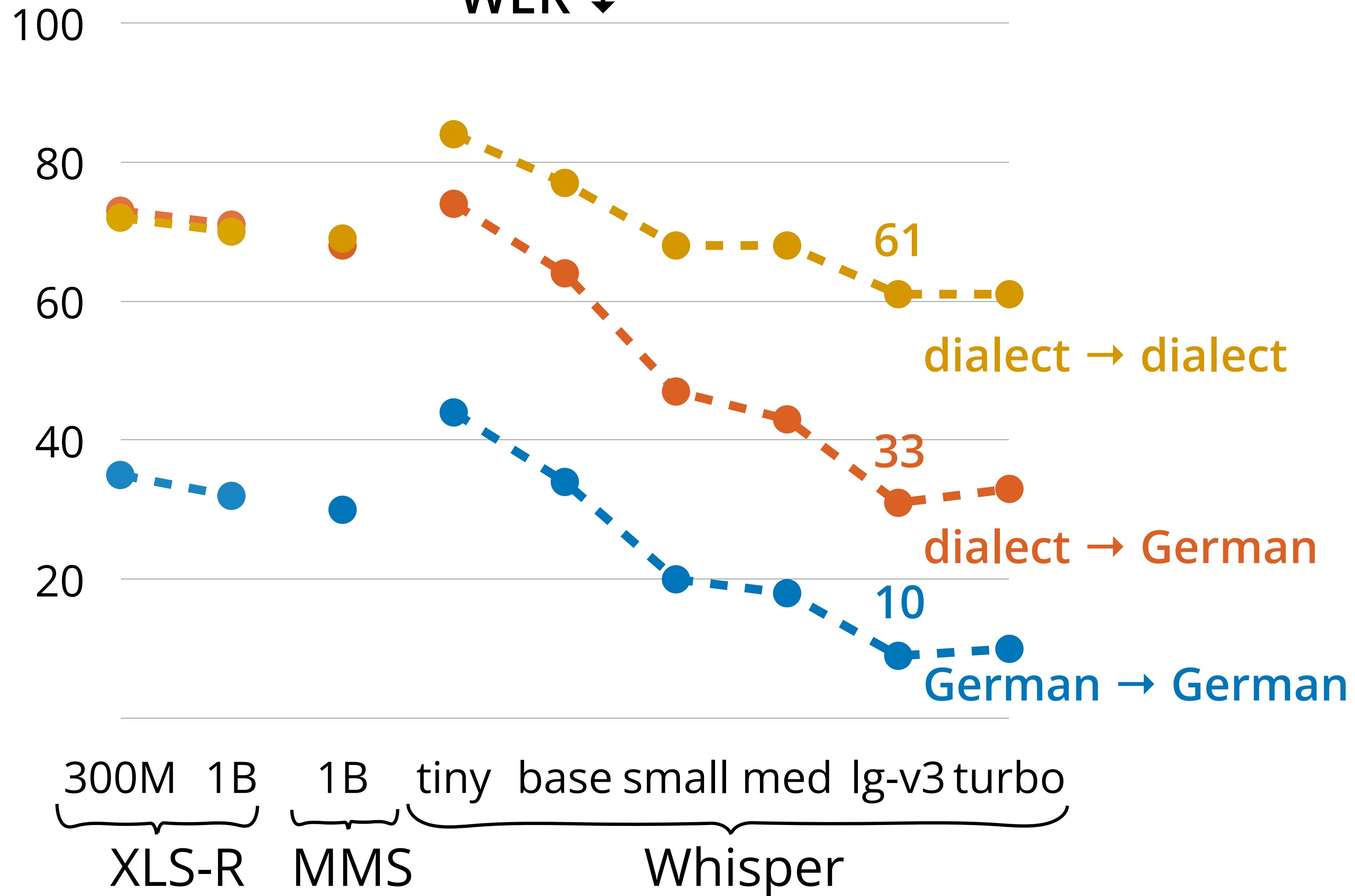
- Whisper – language model decoding
- MMS – connectionist temporal classification (CTC)
- XLS-R (fine-tuned for German ASR) – CTC

Multiple sizes *(more sizes & fine-tuned versions in paper)*

Output language setting: German (no dialects available)

# Quantitative results

WER ↓



## Performance gap

German vs. dialectal audio  
(but no systematic differences across regions)

## Larger models = better

- Distilled *turbo* also good

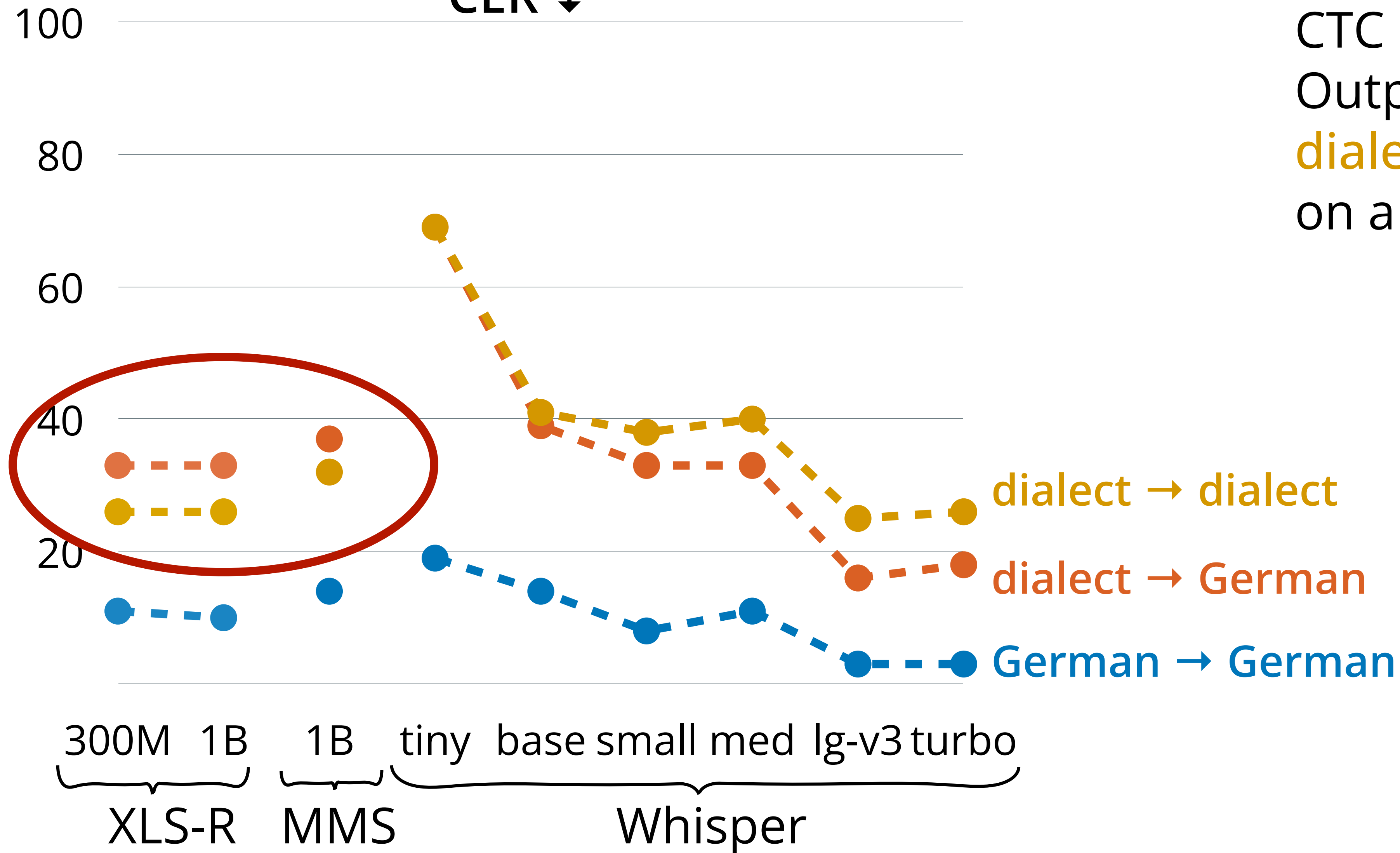
## Dialect audio & decoder types

- Whisper outputs: closer to **German**
- XLS-R & MMS (CTC): similarly distant to both **German** & **dialect**



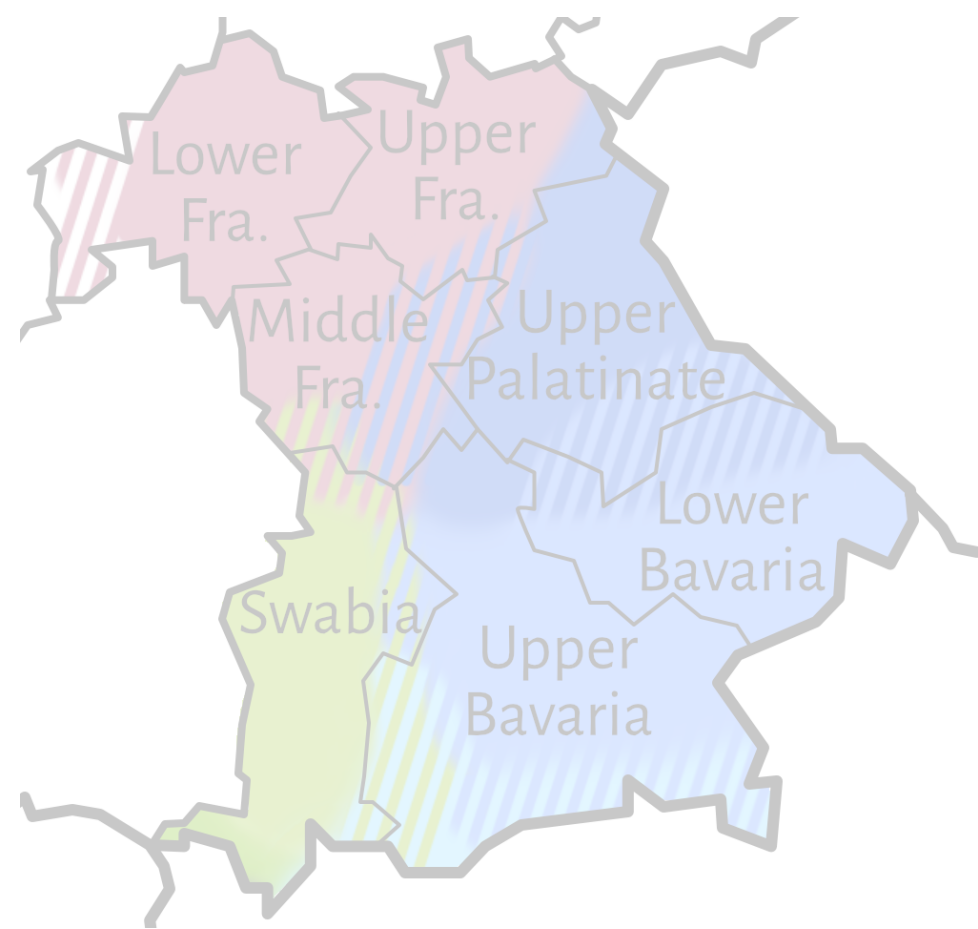
# Quantitative results

CER ↓



CTC models:  
Output is closer to  
**dialect** than **German**  
on a character level

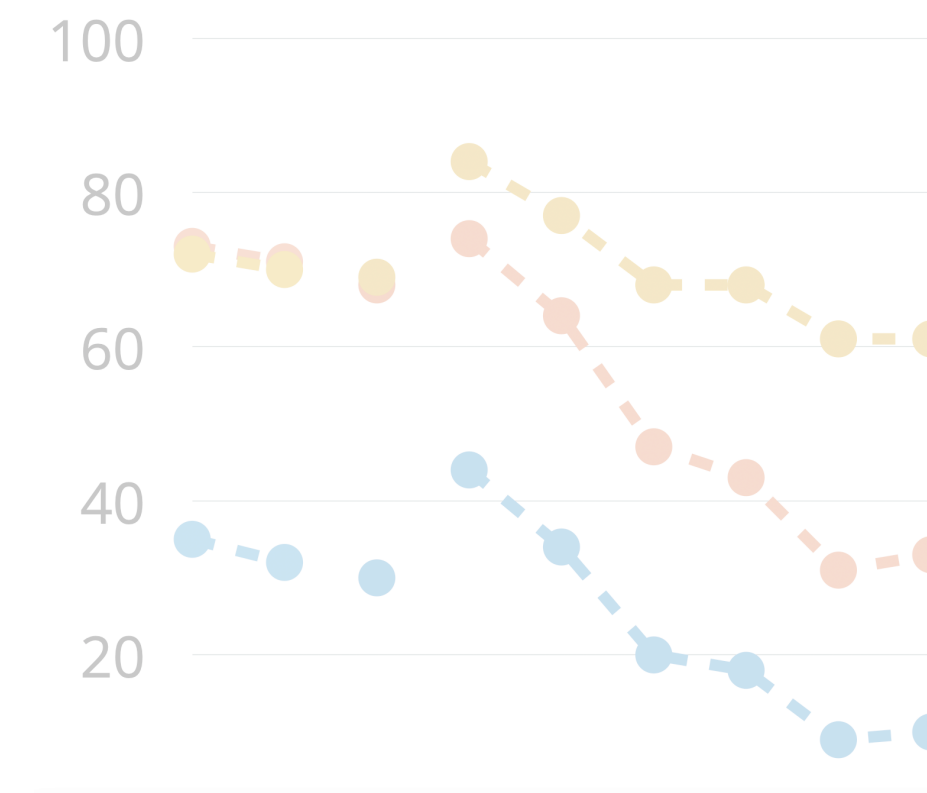
# Overview



1. Dialects in Bavaria



2. *Betthupferl* dataset



3. Benchmarking ASR models



4. Qualitative analysis

# Qualitative analyses – Human evaluation

Comparing ~600 of the best model's hypotheses (Whisper large-v3) to the German references:

- **Meaning:** Is the meaning fully preserved?  $\rightarrow \mu = 3.9 \pm 1.1$
- **Fluency:** Does the output sound like fluent German?  $\rightarrow \mu = 3.7 \pm 1.1$

Likert scale: 1 = worst, 5 = best; 2–3 annotators / sentence


Moderately correlated w/ automatic metrics:  $0.48 \leq |\rho| \leq 0.59$

- Higher when taking the mean of *meaning* and *fluency*:  $0.53 \leq |\rho| \leq 0.63$   
 $\rightarrow$ interplay



# Qualitative analyses – Error analysis

Same ~600 sentences:  identical to German reference  
 different, but acceptable  
 different, and wrong






[German]	Sofort	Mathildas	Geldstück	suchen,	...
	<i>Immediately</i>	<i>Mathilda's</i>	<i>coin</i>	<i>search</i>	
[Dialect]	Sofort	da Mathilda	ihr Geldstücke	sung,	...
		<i>the Mathilda</i>	<i>her</i>		
[ASR]	Sofort	der Mathilda	ihr Geldstück	lesung,	...
					
					

# Qualitative analyses – Error analysis

Same ~600 sentences:

-  identical to German reference
-  different, but acceptable
-  different, and wrong

Words/constructions that...

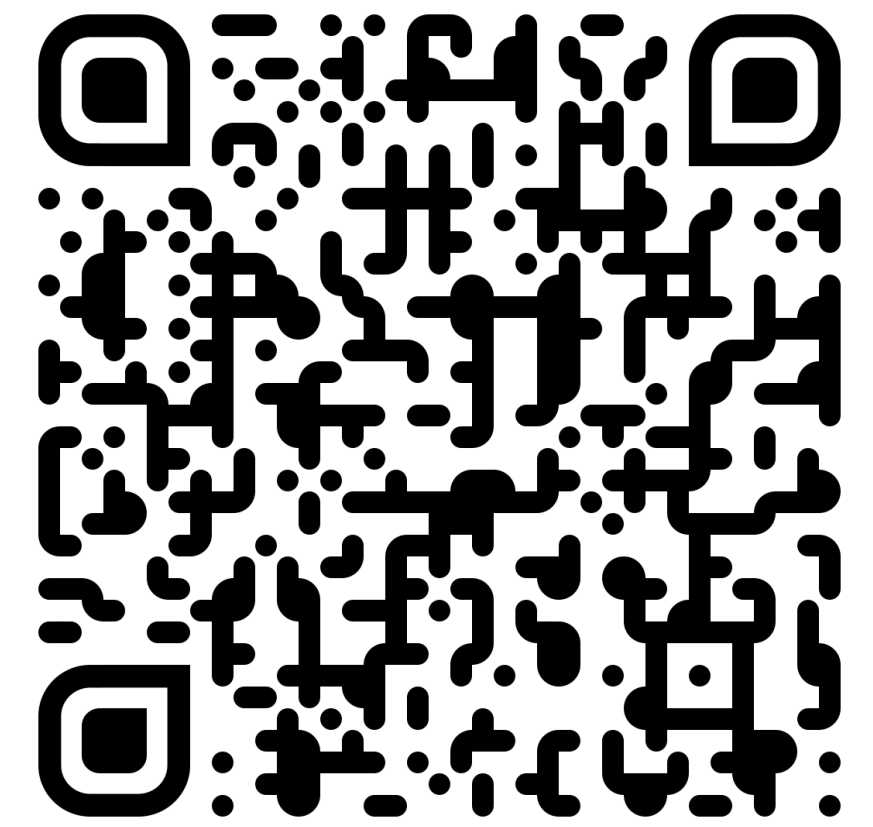
- are identical in German & the dialect: usually correct (86 %) 
- differ only in terms of pronunciation/morphology: usually correct (75 %) 
- lexically different: usually nonsense (63 %) 
- syntactically different: usually like the dialectal structure (acceptability in German varies)  

Common error source: incorrectly recognized word boundaries

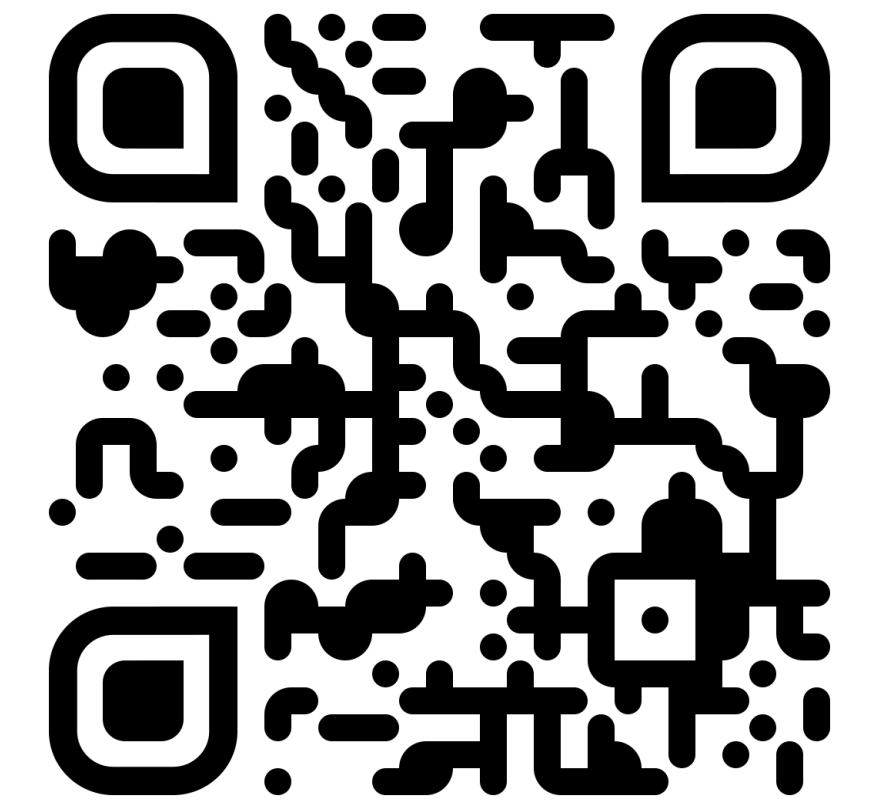
# Summary

- 4.5h of audio w/ dialectal & German transcriptions
- Help us close the performance gap between dialectal & German audio :)
- Lexical/syntactic differences between dialect & standard are a challenge, both for models and for automatic evaluation
- Repo with reference transcriptions, model hypotheses, annotation guidelines, annotations, code
  - Audio clips are shared by request
- Thank you!

Paper



Supplementary  
material



[github.com/mainlp/bettthupfer1](https://github.com/mainlp/bettthupfer1)



# Appendix

## Dataset stats

Region/split	Speakers	Sent	Min	Words/sent		Lev dist	
				Dial	Std	Word	Char
Lower Franconia	1F, 1M	403	33	12.6 <sub>7.6</sub>	12.5 <sub>7.6</sub>	46 <sub>21</sub>	19 <sub>11</sub>
Upper Franconia	3M	561	33	9.1 <sub>5.7</sub>	9.0 <sub>5.6</sub>	52 <sub>22</sub>	23 <sub>13</sub>
Middle Franconia	4M	371	36	15.1 <sub>8.9</sub>	15.2 <sub>8.8</sub>	57 <sub>20</sub>	23 <sub>11</sub>
Upper Palatinate	1F, 1M	394	34	14.0 <sub>9.0</sub>	13.9 <sub>8.9</sub>	58 <sub>19</sub>	24 <sub>11</sub>
Lower Bavaria	2F, 1M	488	32	10.8 <sub>7.3</sub>	11.1 <sub>7.4</sub>	68 <sub>21</sub>	30 <sub>12</sub>
Upper Bavaria	1F, 2M	465	37	11.8 <sub>8.0</sub>	12.1 <sub>8.2</sub>	57 <sub>21</sub>	23 <sub>11</sub>
Swabia	1F, 1M	575	37	10.5 <sub>6.6</sub>	10.7 <sub>6.7</sub>	57 <sub>22</sub>	22 <sub>12</sub>
All dialects	6F, 13M	3 257	241	11.7 <sub>7.7</sub>	11.8 <sub>7.8</sub>	57 <sub>22</sub>	24 <sub>12</sub>
Std. German	6F, 7M	531	32	—	8.9 <sub>5.6</sub>	—	—
Full dataset	8F, 14M	3 788	273	—	11.4 <sub>7.6</sub>	—	—

# Appendix

## Differences between references & Error analysis

Difference	Proportion w. type of diff. (%)		Hypothesis words		
	Sent	Word	✔	⊙	✗
— (identical word)	97	45	86	4	10
Phonetic/morphological	96	47	75	5	20
Word splitting	41	4	54	10	36
Determiner + name	29	3	10	77	13
Word choice	23	2	8	30	63
Verb phrase construction	7	1	13	23	63
Word order	6	1	0	82	18
Dropped/fused pronoun	5	0	40	0	60
Possessive	2	0	0	57	43
Other	8	1	27	47	27

# Appendix

## Human judgements

	Avg	IAA	Correlations ( $\rho$ , mean over annotators)			
			Fluency	WER	CER	BLEU
Meaning	3.9 <sub>1.1</sub>	0.76 <sub>0.05</sub>	0.73 <sub>0.05</sub>	-0.57 <sub>0.03</sub>	-0.56 <sub>0.03</sub>	0.48 <sub>0.02</sub>
Fluency	3.7 <sub>1.1</sub>	0.75 <sub>0.03</sub>	—	-0.59 <sub>0.04</sub>	-0.56 <sub>0.02</sub>	0.51 <sub>0.03</sub>
Both	3.8 <sub>1.0</sub>	0.83 <sub>0.03</sub>	—	-0.63 <sub>0.04</sub>	-0.61 <sub>0.03</sub>	0.53 <sub>0.03</sub>