IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Jan Gorisch, Department of Pragmatics; Thomas Schmidt, linguisticbits.de

EVALUATING WORKFLOWS FOR CREATING ORTHOGRAPHIC TRANSCRIPTS FOR ORAL CORPORA BY TRANSCRIBING FROM SCRATCH OR CORRECTING ASR-OUTPUT

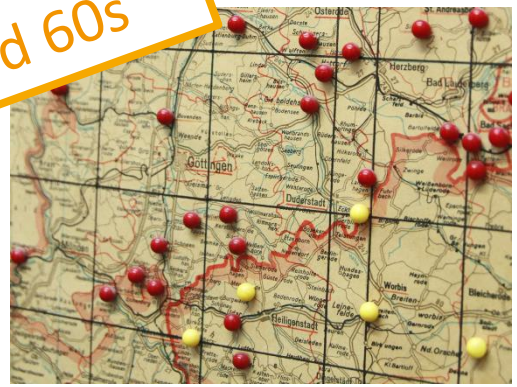
LREC-COLING 2024, Turin

Leibniz Gemeinschaft

INTRODUCTION

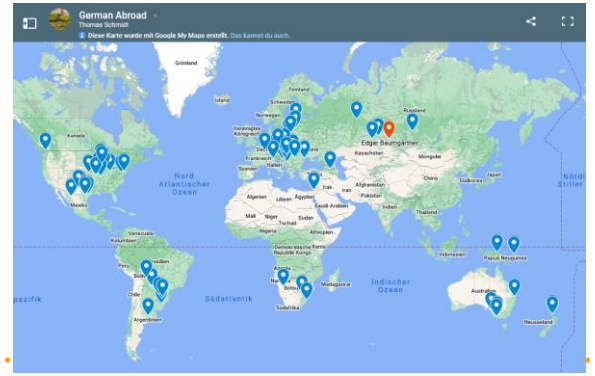
IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

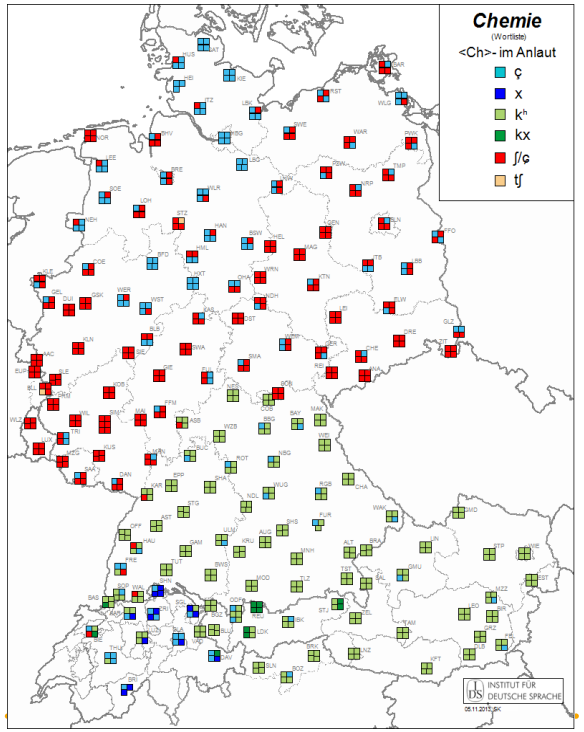
also historic data from the 1950s and 60s



The AGD* hosts…

- **Conversation Corpora**
- **Variation Corpora**
- **Extra-territorial Varieties**





Sprecher Ortsdaten Koordinaten Geocode Geografische_Länge

```
33    <Sigle_in_Transkripten>AAC1</Sigle_in_Transkripten>
34    <Anmerkungen/>
35  </Basisdaten>
36  <Ortsdaten Typ="Geburtsort">
37    <Land Kürzel="DE">Deutschland</Land>
38    <Region>Nordrhein-Westfalen</Region>
39    <Kreis>Nicht vorhanden</Kreis>
40    <Ortsname>Aachen</Ortsname>
41    <Koordinaten>
42      <Geocode>
43        <Geografische_Breite>50.776207</Geografische_Breite>
44        <Geografische_Länge>6.083788</Geografische_Länge>
45      </Geocode>
46      <Planquadrat>Nicht vorhanden</Planquadrat>
47      <Anmerkungen>Nicht dokumentiert</Anmerkungen>
48    </Koordinaten>
49    <Ortsteil>Nicht dokumentiert</Ortsteil>
50    <Ortsbeschreibung>Nicht vorhanden</Ortsbeschreibung>
51    <Aufenthaltsdauer>1988-2006</Aufenthaltsdauer>
52    <KB-Link>
57    <Anmerkungen>18 Jahre</Anmerkungen>
58  </Ortsdaten>
59  <Ortsdaten Typ="Schulort">
82  <Ortsdaten Typ="Wohnort">
105 <Sprachdaten>
106   <Sprachkenntnisse Sprachname="Englisch">
119   <Sprachkenntnisse Sprachname="Französisch">
132   <Sprachkenntnisse Sprachname="Griechisch">
145   <Sprachproduktion>
176   <Sprachgebrauch Domäne="In der Familie">
184   <Sprachgebrauch Domäne="Mit Freunden">
192   <Sprachgebrauch Domäne="Mit Mitschülern">
200   <Sprachgebrauch Domäne="Im Schulunterricht">
208   <Sprachgebrauch Domäne="In formellen Situationen (Vortrag/Rede)"
216 </Sprachdaten>
217 <Beziehung_zu_anderem_Sprecher Kennung_anderer_Sprecher="DH--_S_00
218   <Typ_der_Bezugsperson>Klassenkamerad</Typ_der_Bezugsperson>
219   <Dauer der Beziehung>Nicht dokumentiert</Dauer der Beziehung>
```

*AGD: Archive for Spoken German (https://agd.ids-mannheim.de)

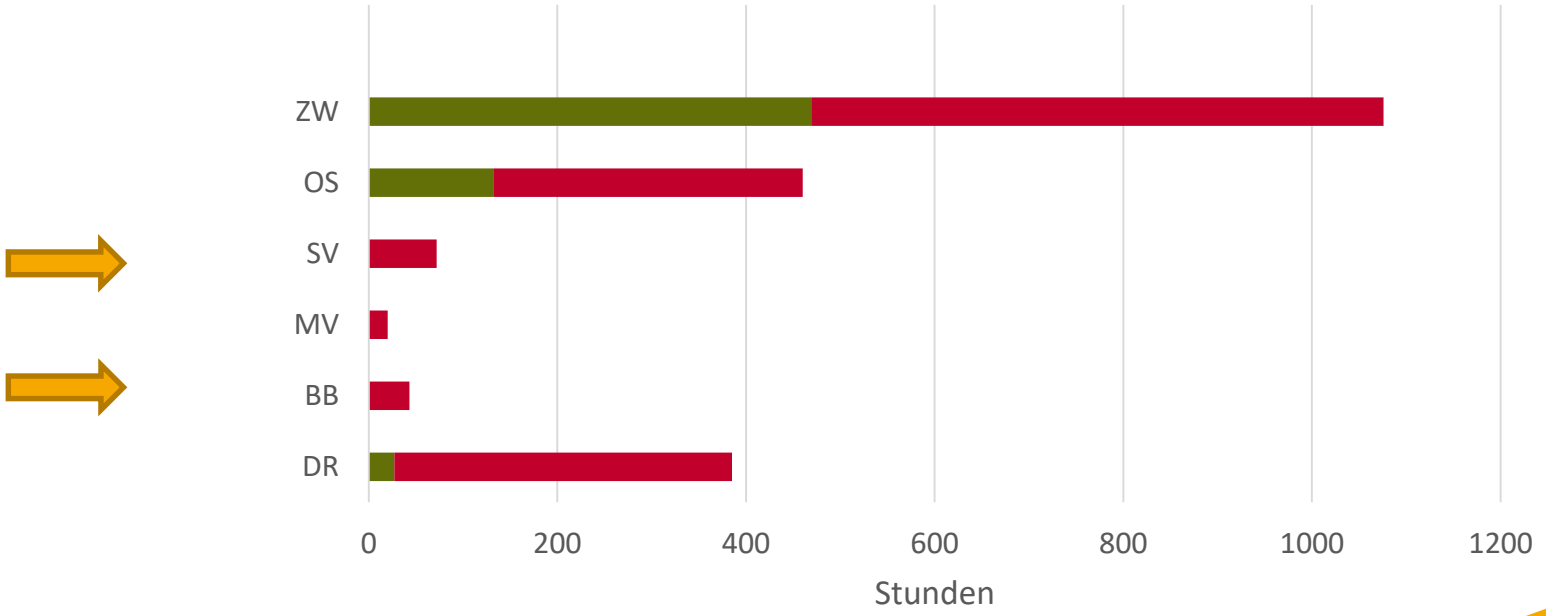
TRANSCRIPTS (STATE 2018)

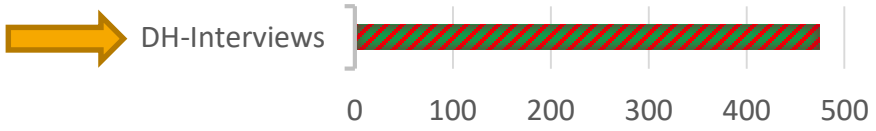
IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

CORPUS (selection)	# record. total	# record. transcribed	% transcribed	hours transcribed	hours untranscribed
ZW – Zwirner	5796	2495	43%	470	606
OS – ehem. Dt. Ostgebiete	981	280	29%	132	328
SV – Südwest. u. Vorarlb.	242	0	0%	0	72
MV – Varia	72	0	0%	0	20
BB – Böblingen	73	2	3%	1	42
DR – dt. Mundart. DDR	444	33	7%	27	358
total	**7608**	**2810**	**37%**	**630**	**1426**
DH-Interviews – Deutsch Heute	688	688 / 2	~50%	(interviewee) 237	(interviewer) 237

TRANSCRIPTS (STATE 2018)

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE





also FOLK:
a yearly growing reference corpus of interactional/multimodal Talk

TRANSCRIPTION BOTTLENECK

IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

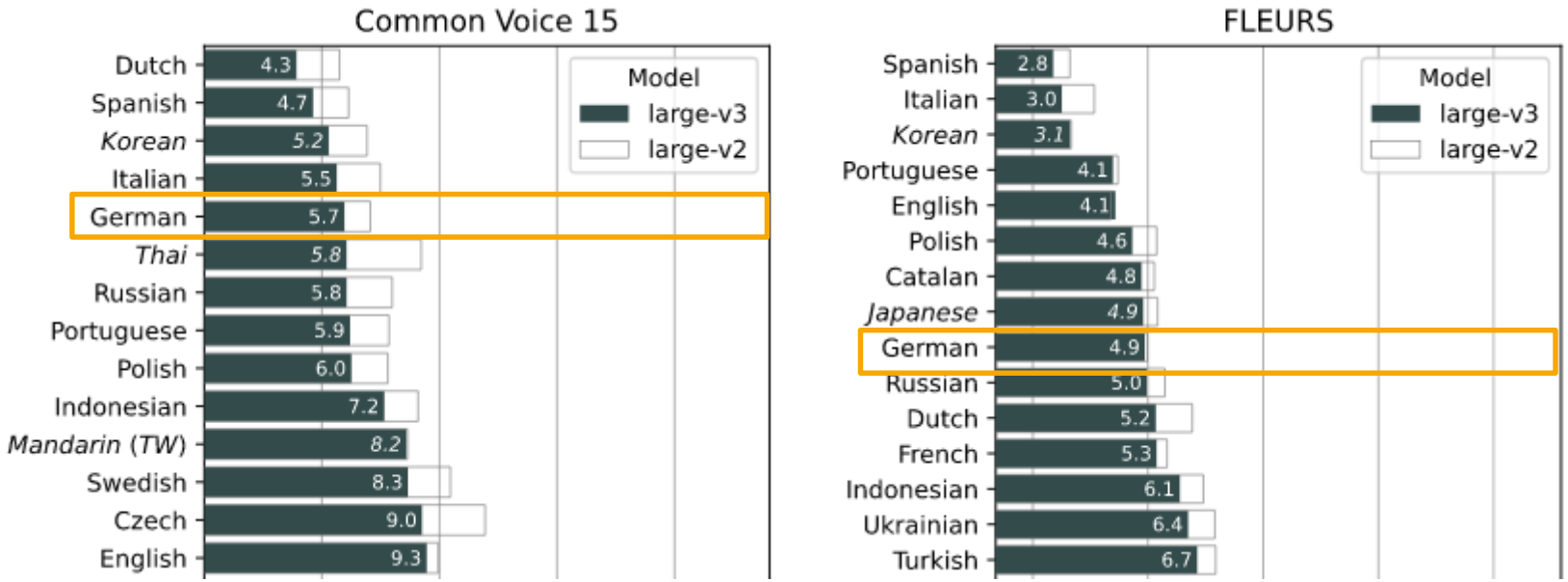
- Estimated need for transcription: 75.000h
  - e.g. 15 students for 10 years
  - + Overhead for management, quality control, technical supervision, documentation
  - Advanced dialect-competence(s) necessary – Alsatian, Low-German, Silesian, (data also contain Frisian, Sorbian, Dutch)
- Central, manual transcription in principle too expensive, not organisable
- Alternatives (cf. Brinckmann 2009)
  - Outsourcing
  - „Crowd“-Sourcing
  - (partial) automation

ASR

OPENAI‘S WHISPER MODEL

IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

- Word-Error-Rate (WER)



As discussed in the accompanying paper, we see that performance on transcription in a given language is directly correlated with the amount of training data we employ in that language.

https://github.com/openai/whisper (25.03.2024) https://huggingface.co/openai/whisper-large-v2 (25.03.2024)

TODAY'S RESEARCH QUESTIONS

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Huge improvements in ASR-quality in recent years (month

- When is ASR-output good enough? (
- If ASR-output needs to be
- g oral corpora)
- ore efficient
- nscribing from scratch?

We might now be at
where corr
ing

Is ASR the solution to the problem?

Or do we need a different problem to the solution?

MATERIAL FOR THE EXPERIMENT

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

- Speech-biographic interviews of the DH-Korpus („deutsch heute“)
  - Hannover (speaker HAN1, HAN2, … HAN4 and interviewer NL)
  - Innsbruck (speaker IBK1, IBK2, …, IBK4 and interviewer MF)
- Later: Recordings from DH-Zurich, corpora: BB, SV,

ASR: OpenAI-Whisper

medium model, later: large-v1

Speaker diarization acc… …aset recordings), later: pyannote.audio

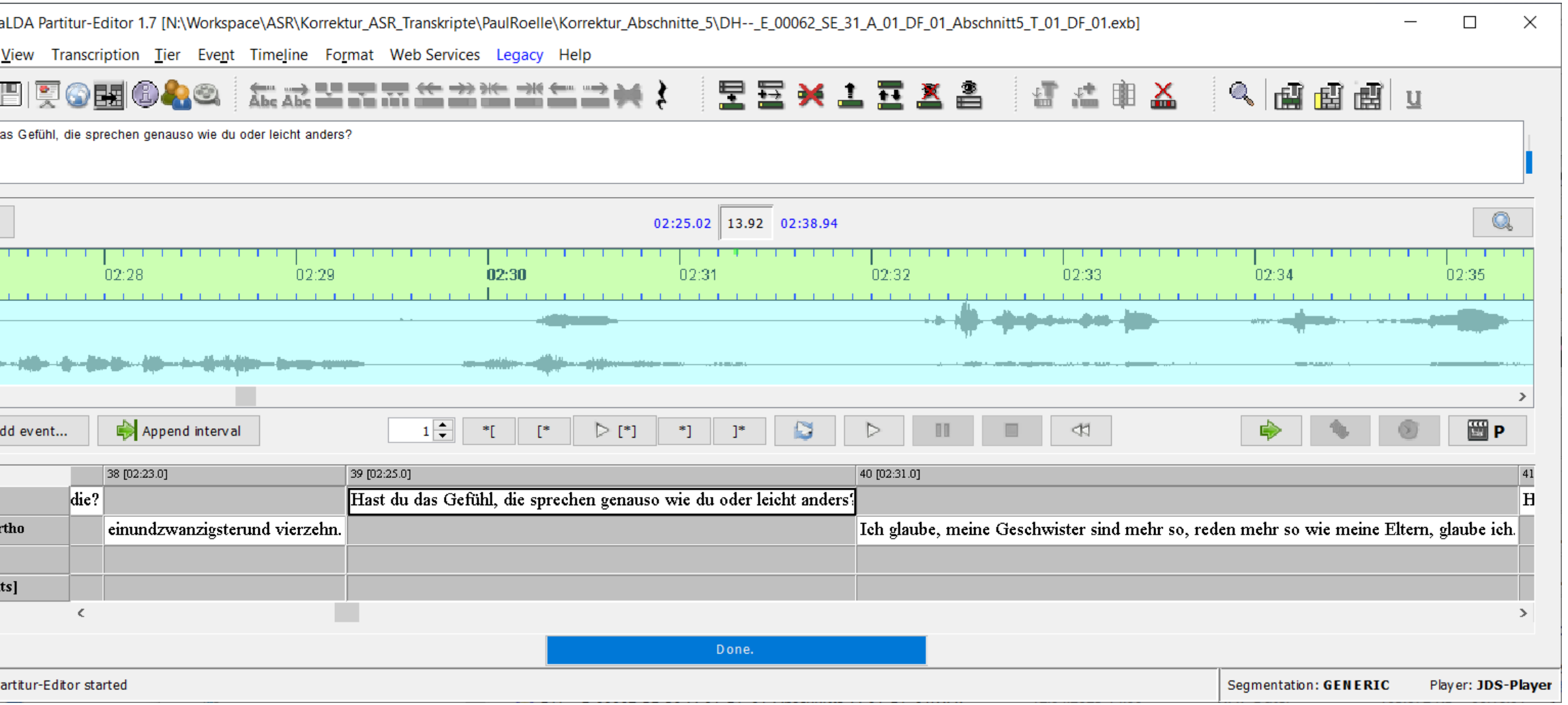
Conversion to EXM… …Editor format (*.exb)

! Limitations !

The AGD is an operating Archive:
Besides evaluation, we try to produce transcripts and provide annotators woth the best input possible

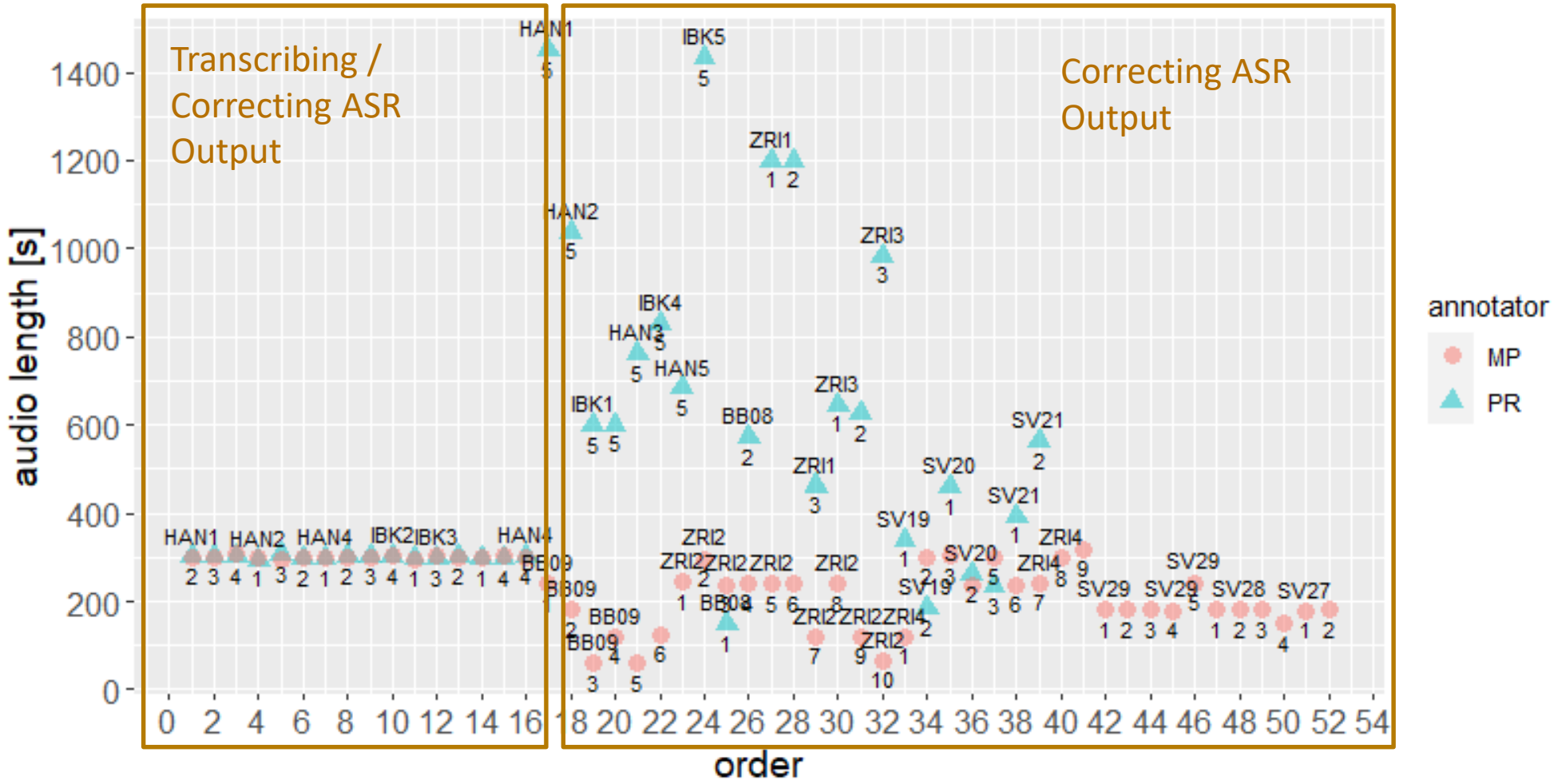
EXAMPLE & INTERFACE

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE



EXPERIMENTAL DESIGN / TASK(S)

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE



INSTRUCTIONS AND METRIC

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

“Create orthographic transcripts”

- Alignment
- Mask tier

“Note the (working) time spent per stretch of recording”

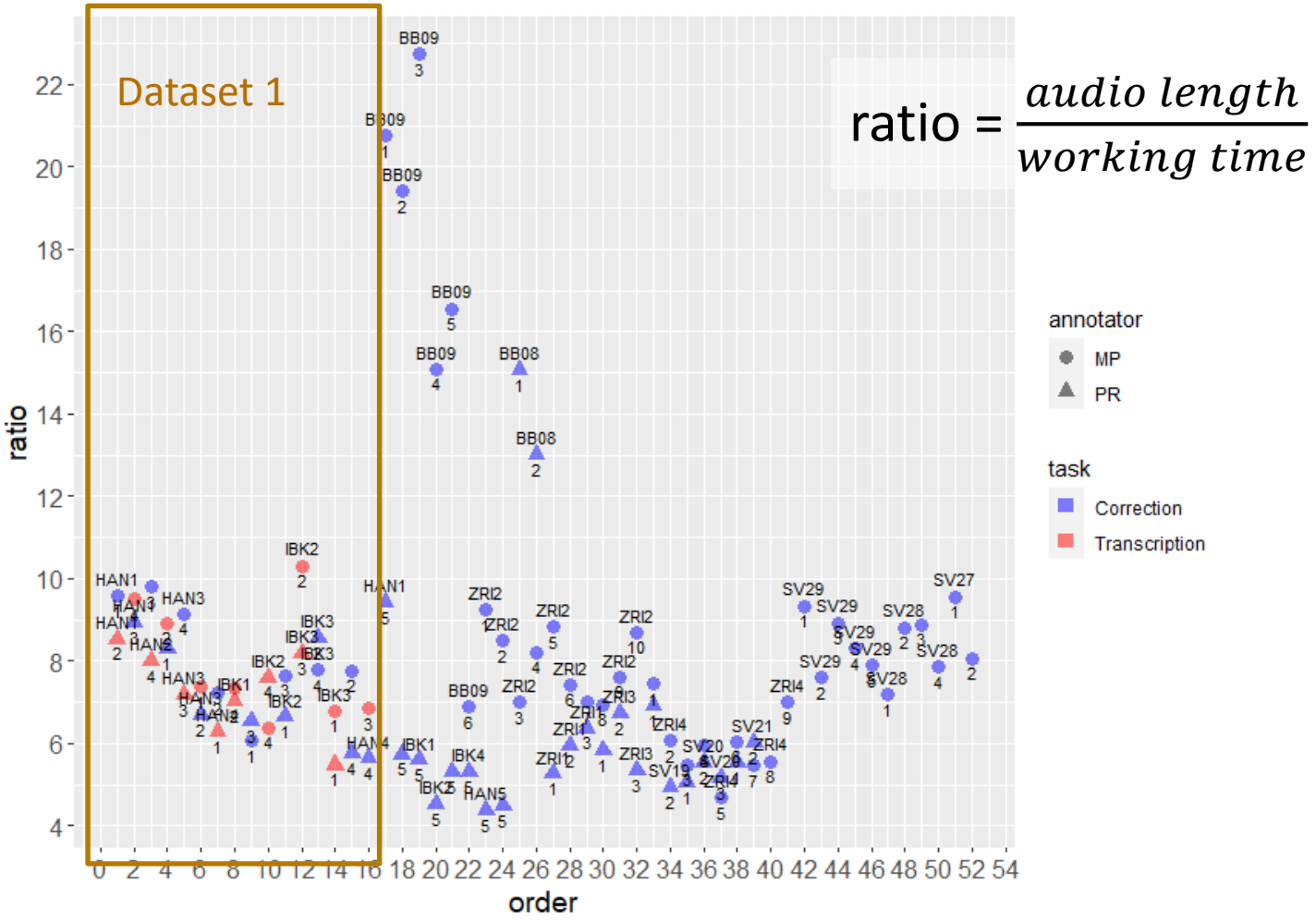
Metric: ratio of $\frac{audio\ length}{working\ time}$

RESULTS

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

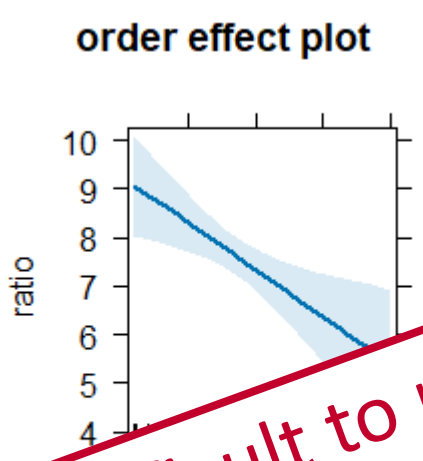
Linear regression model in R

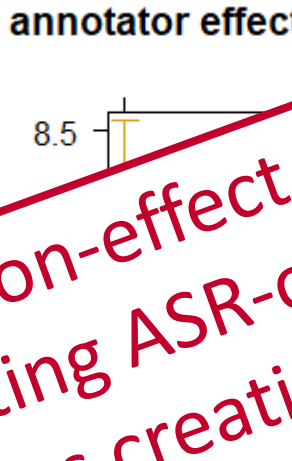
- Criterion variable:
  - ratio
- Predictors:
  - Task (transcribing vs. correcting)
  - Corpus
  - Recording place
  - Annotator
  - Order



TRANSCRIBING VS. CORRECTING

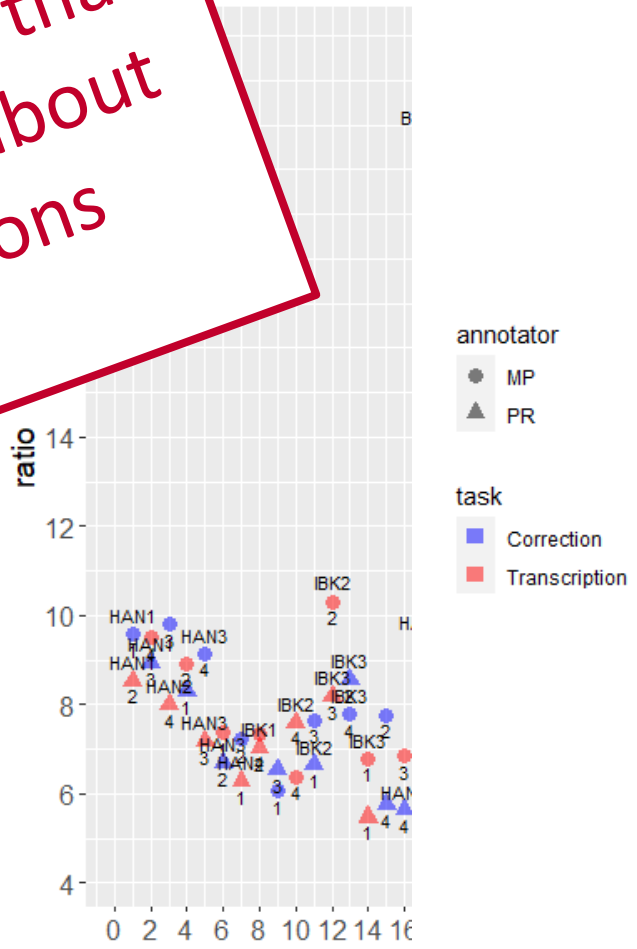
IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE





Difficult to report a non-effect, but it seems that: Correcting and editing ASR-output takes about the same effort as creating transcriptions “manually” from scratch.

Factor			z	p	
(Intercep			19.240	< 2e-16	***
taskT		0.3705	−0.298	0.76780	
order		0.0618	−3.115	0.0043	**
placeIBK	0.5344	0.5685	0.940	0.3555	
annotatorPR	−0.8195	0.3692	−2.220	0.035	*



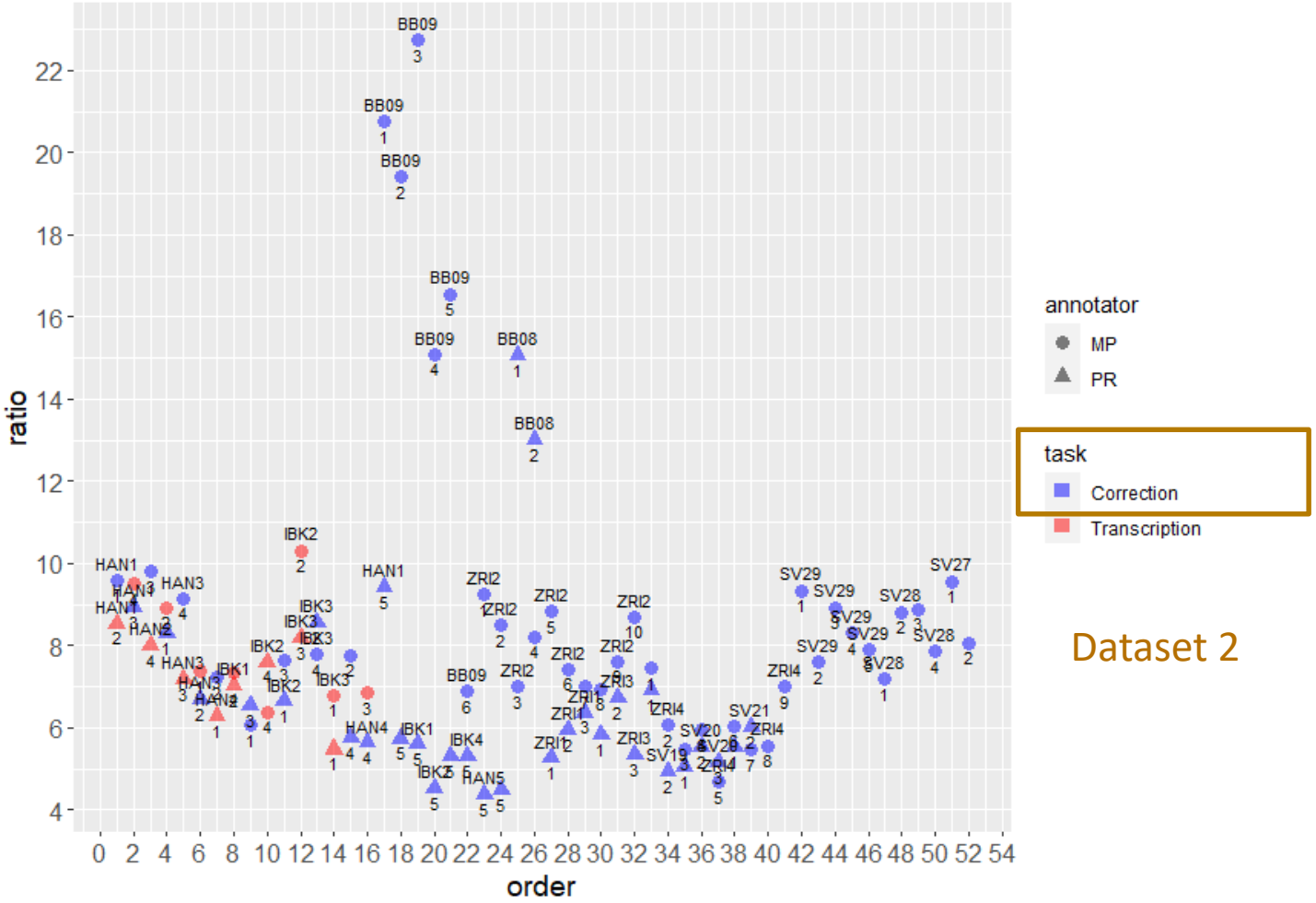
Adjusted R-squared = 0.33

CORRECTING ONLY

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Linear regression model

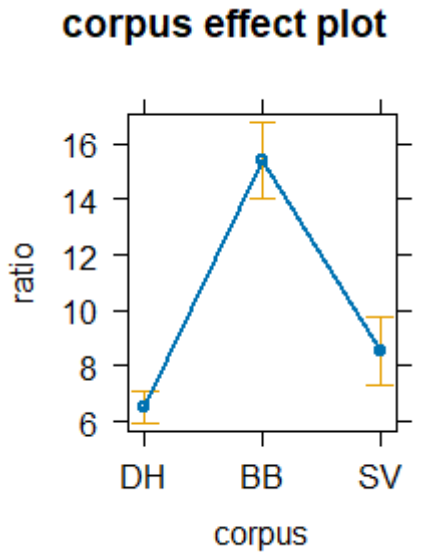
- Criterion variable:
  - ratio
- Predictors:
  - Corpus
  - Recording place
  - Annotator
  - Order

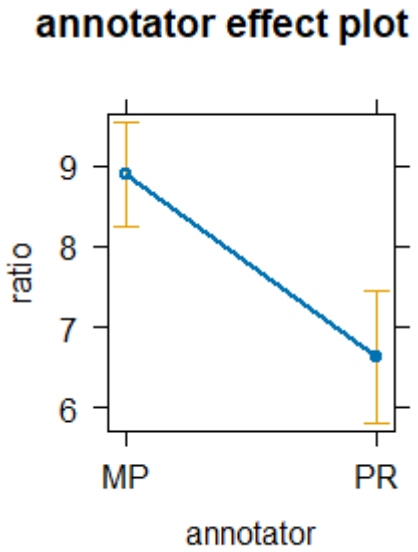


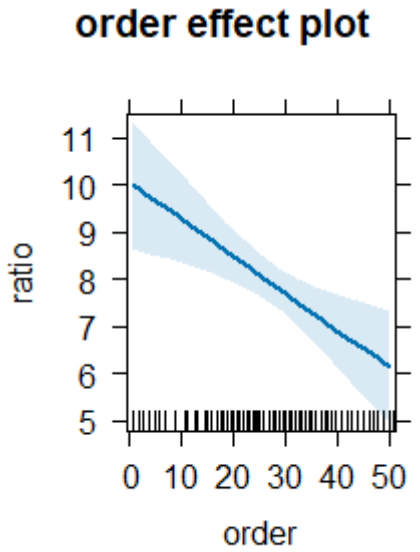
Dataset 2

CORRECTING

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE







Factor	Estimate	Std. Error	z	p	
(Intercept)	9.451	0.691	13.679	< 2e-16	***
corpusBB	8.888	0.7327	12.131	< 2e-16	***
corpusSV	2.0189	0.7552	2.637	0.0094	**
annotatorPR	−2.2656	0.5945	−3.811	0.0003	***
order	−0.0787	0.0244	−3.232	0.0019	**
audio_length_in_s	−0.0004	0.000 98	0.374	0.7094	

Adjusted R-squared = 0.75

QUALITATIVE RESULTS
COMMENTS FROM ANNOTATORS

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Stretch	Comment
HAN2-1	Speakers are extremely often on the wrong tier. Much overlap.
HAN3-2	Again often the wrong tier. The interviewee speaks rather clear.
HAN4-1	Little overlap. Missing uhm and yes, as always.
HAN4-4	Extremely often wrong tier.
IBK1-2	Extremely often overlap.
IBK1-3	No opportunity is missed to overlap. On the contrary, the dialect is surprisingly well transcribed.
IBK2-4	As above; and the interviewee speaks extremely unclear and quiet.
IBK2-1	Again missing uhm and yes in overlap.
IBK3-3	The interviewee speaks a lot.
IBK3-2	Very often wrong tier; repetitions are almost never detected.
IBK4-1	Little talk and little overlap.
IBK1-5	The alignment of the last minute was completely off; the first time this kind of error.
HAN3-5	More often the wrong tier than in all previous transcripts.
HAN4-5	Both speak very clear and in longer sequences.
BB08	Correction might have taken longer as more listening was necessary.
SV19-1	Despite strong dialect, easy to transcribe because of little overlap.
SV22-1	One speaker speaks very unclearly and the interviewer is extremely quiet.
SV26-1	Possibly shorter correction time, as much was simply incomprehensible.

ANECDOTES I

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Example from Südwestdeutschland and Vorarlberg

Ref:	*wir*	*sind*	*zuerst*	*natürlich*	*selbstständiger*	*Konsum*	*gewesen*
Ref:	we	have	first	of course	autonomous	cooperative	been

'Ref: at the beginning we have been an autonomous cooperative of course'

Hyp:	*Wir*	*waren*	*zuerst*	*natürlich*	*selbstständiger*	*Konsum*
Hyp:	we	were	first	of course	autonomous	cooperative

'Hyp: first we were an autonomous cooperative of course'

Cut Excerpt:
SV--_E_00019 from 08:25.34 to 08:28.70
And add 1 s silence at the beginning

ANECDOTES II
ON AMBERSCRIPT

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Despite doing speaker recognition, Amberscript does not seem to be able to deal with:

Genus-accord.

Ségolène Royale (feminin) said:

- “je ne suis pas sortie”

Amberscript transcribes (masculin):

- “je ne suis pas sorti”

CONCLUSIONS & FUTURE WORK

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Problem: Transcription bottleneck

Solution: ASR

Improve and optimize the solution:

- Exploit prompting mechanism
- Avoid some normalizations
- Different post-processing
- Adapt system to data (fine-tuning)

Adjust or redefine the problem:

- Analyze the imperfect output
- Use additional ASR information, e.g. probability of word-detection

1. Generic query on ASR-transcribed corpus
2. Refine manually the results

Record audio as close to the (individual) speaker as possible

fin

EVOLVING MODELS… (LARGE) V1, V2, V3 AND AMOUNTS OF TRAINING DATA

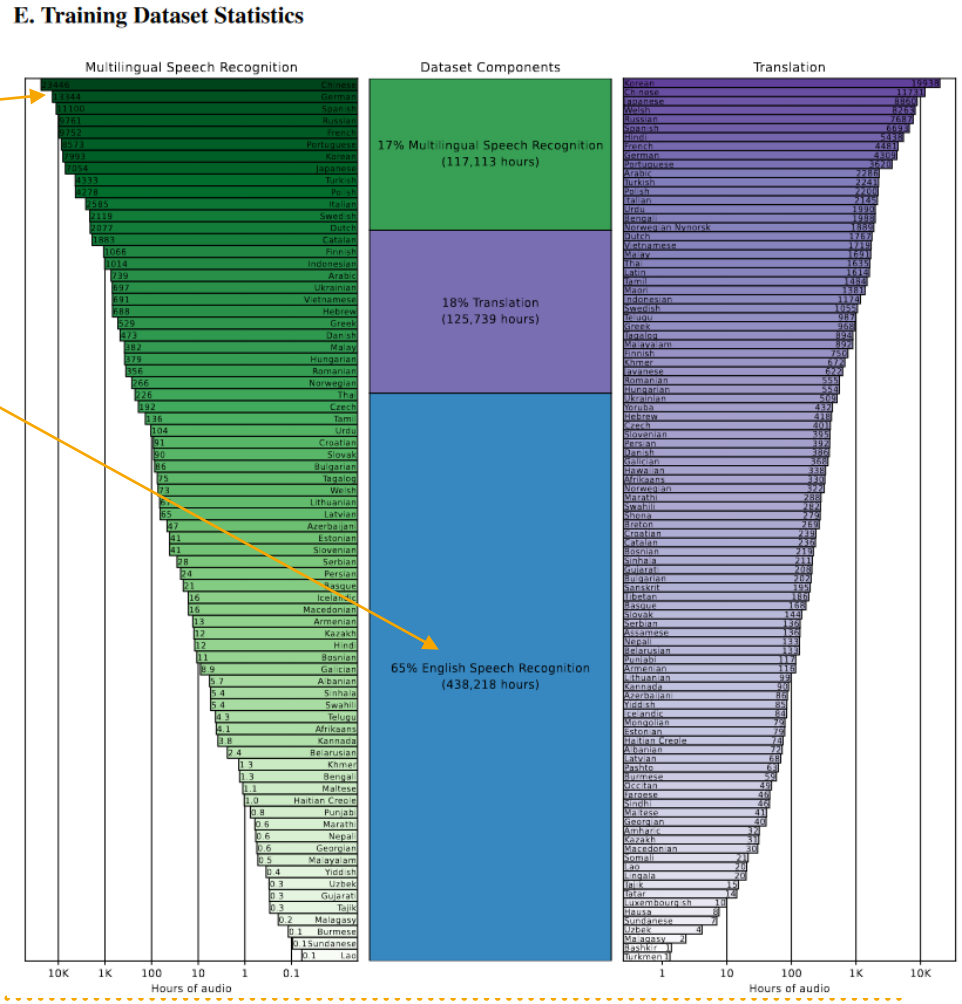
IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

- `large-v1`: Training Data: 13344 hours for German

65% English Speech Recognition
(438,218 hours)

- `large-v2`: Compared to the Whisper large model, the large-v2 model is trained for 2.5x more epochs with added regularization

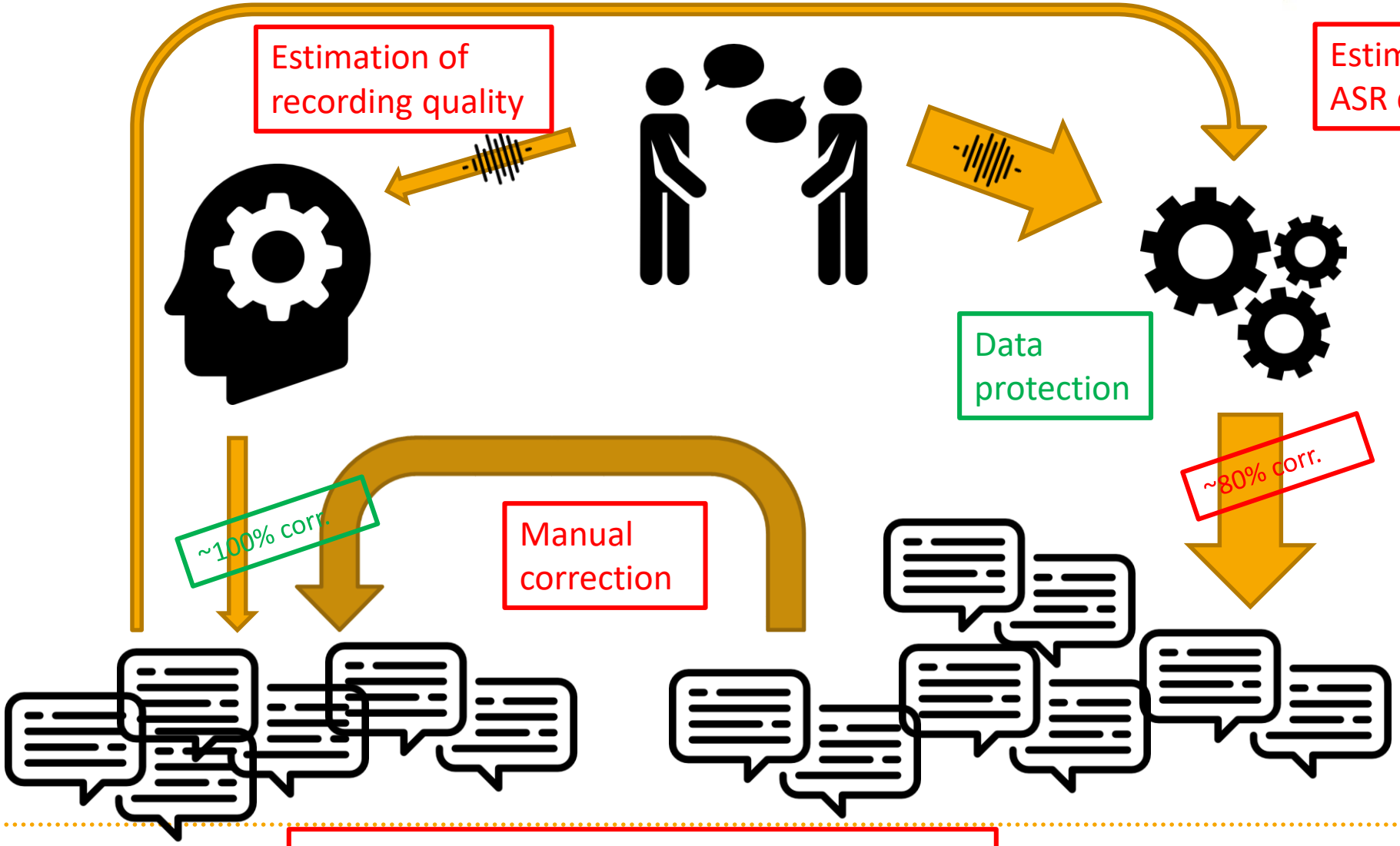
- `large-v3`:
  - The input uses 128 Mel frequency bins instead of 80
  - 1 million hours of weakly labeled audio and 4 million hours of pseudolabeled audio collected using large-v2.



Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, pages 28492–28518. PMLR.

CHALLENGES / IDEAS

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE



Evaluation: WER-estimation and exploitation

QUESTIONS TO THE AUDIENCE

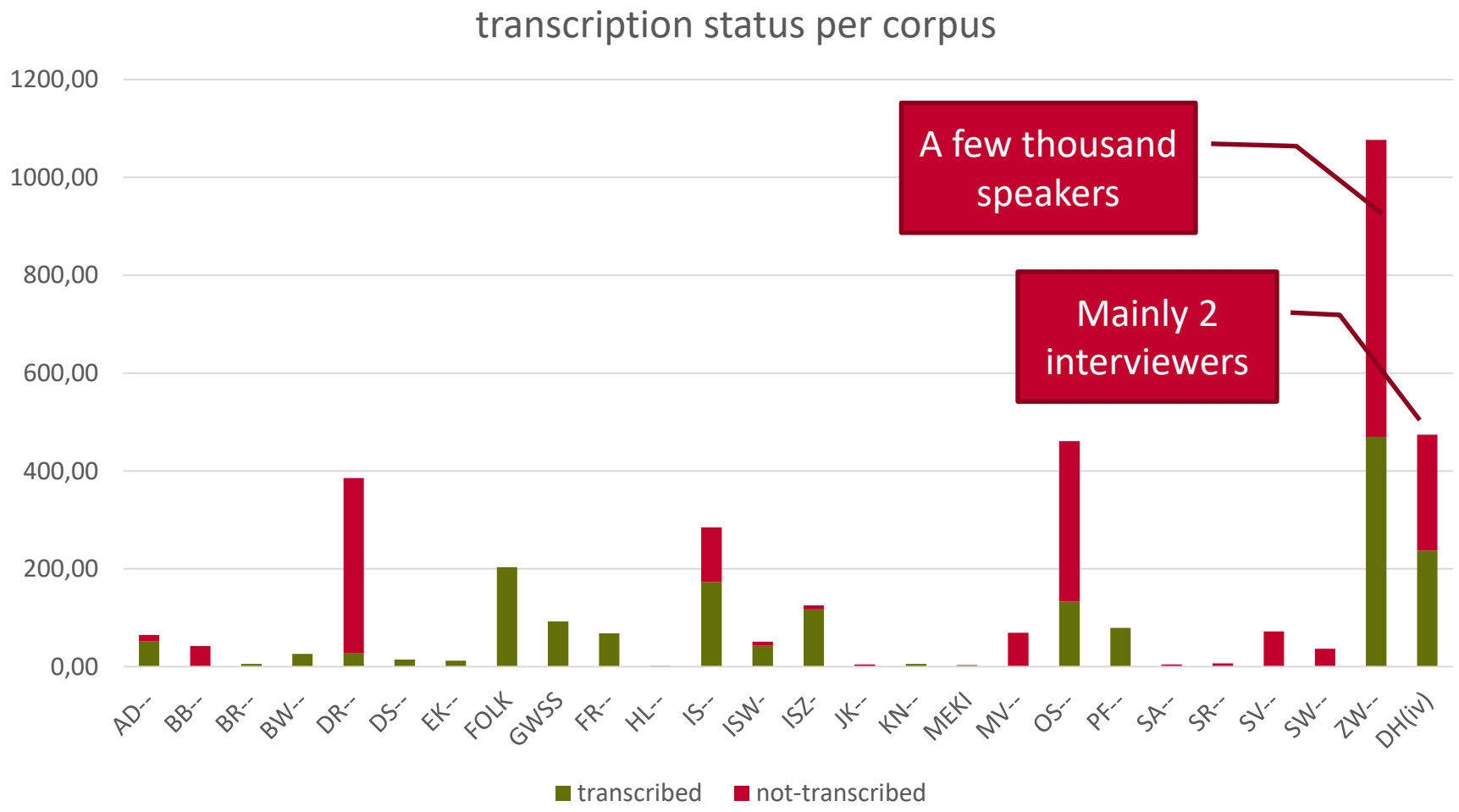
IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

- Who knows of literature on “average time spent for (orthographic) transcription”
  - According to types of transcription conventions (cGAT, normalized)
- Is there any literature on correcting ASR-output? Any tools for that specific purpose?
- Are there open (speech) recognizers around that can also recognize/output …
  - Backchannels
  - Hesitations
  - Repetitions
  - Annotations of non-verbal sounds / sounds of background noise
  - Dialectal speech from old speakers recorded in the 1960s in rural areas?
- Which tools can be used for evaluating/benchmarking ASR-output (WER, Insertions, Deletions, Substitutions)
  - Kaldi, HResult in HTK-3, …, ??diff in Gitlab??

REFERENCE(S)

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Brinckmann, Caren (2009): Transcription Bottleneck of Speech Corpus Exploitation. In: Lyding, Verena (Ed.): *LULCL II 2008 - Proceedings of the Second Colloquium on Lesser Used Languages and Computer Linguistics*. Bozen-Bolzano, 13th-14th November 2008. Bozen-Bolzano: EURAC. 165-179.


transcription status per corpus
1200,00
1000,00
800,00
600,00
400,00
200,00
0,00
A few thousand speakers
Mainly 2 interviewers
AD--
BB--
BR--
BW--
DR--
DS--
EK--
FOLK
GWSS
FR--
HL--
IS--
ISW-
ISZ-
JK--
KN--
MEKI
MV--
OS--
PF--
SA--
SR--
SV--
SW--
ZW--
DH(iv)
transcribed
not-transcribed

METHOD FOR THE

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

- 5-minute chunks
- Two student helpers (Maja Peer & Paul Rölle)
- Two tasks (T & K)
  - T(ranscribing) from scratch
  - C(orrecting) ASR-output
    - ✓ speaker association
    - ✓ content
    - ✓ alignment
- Training (30 Min. for C; 20 Min. for T)
- Measure: working time (stopwatch)
- Metric: ratio of $\frac{audio\ length}{working\ time}$

Hannover

task	1st chunk	2nd chunk	3rd chunk	4th chunk	5th chunk	interviewer	interviewee/ rec. name
C	Maja		Paul			NL	HAN1
T		Paul		Maja			
C	Paul		Maja			NL	HAN2
T		Maja		Paul			
C		Paul		Maja		NL	HAN3
T	Maja		Paul				
C		Maja		Paul		NL	HAN4
T	Paul		Maja				

Innsbruck

task	1st chunk	2nd chunk	3rd chunk	4th chunk	5th chunk	interviewer	interviewee/ rec. name
C	Maja		Paul			MF	IBK1
T		Paul		Maja			
C	Paul		Maja			MF	IBK2
T		Maja		Paul			
C		Paul		Maja		MF	IBK3
T	Maja		Paul				
C		Maja		Paul		MF	IBK4
T	Paul		Maja				

COMMENTS FROM ANNOTATOR

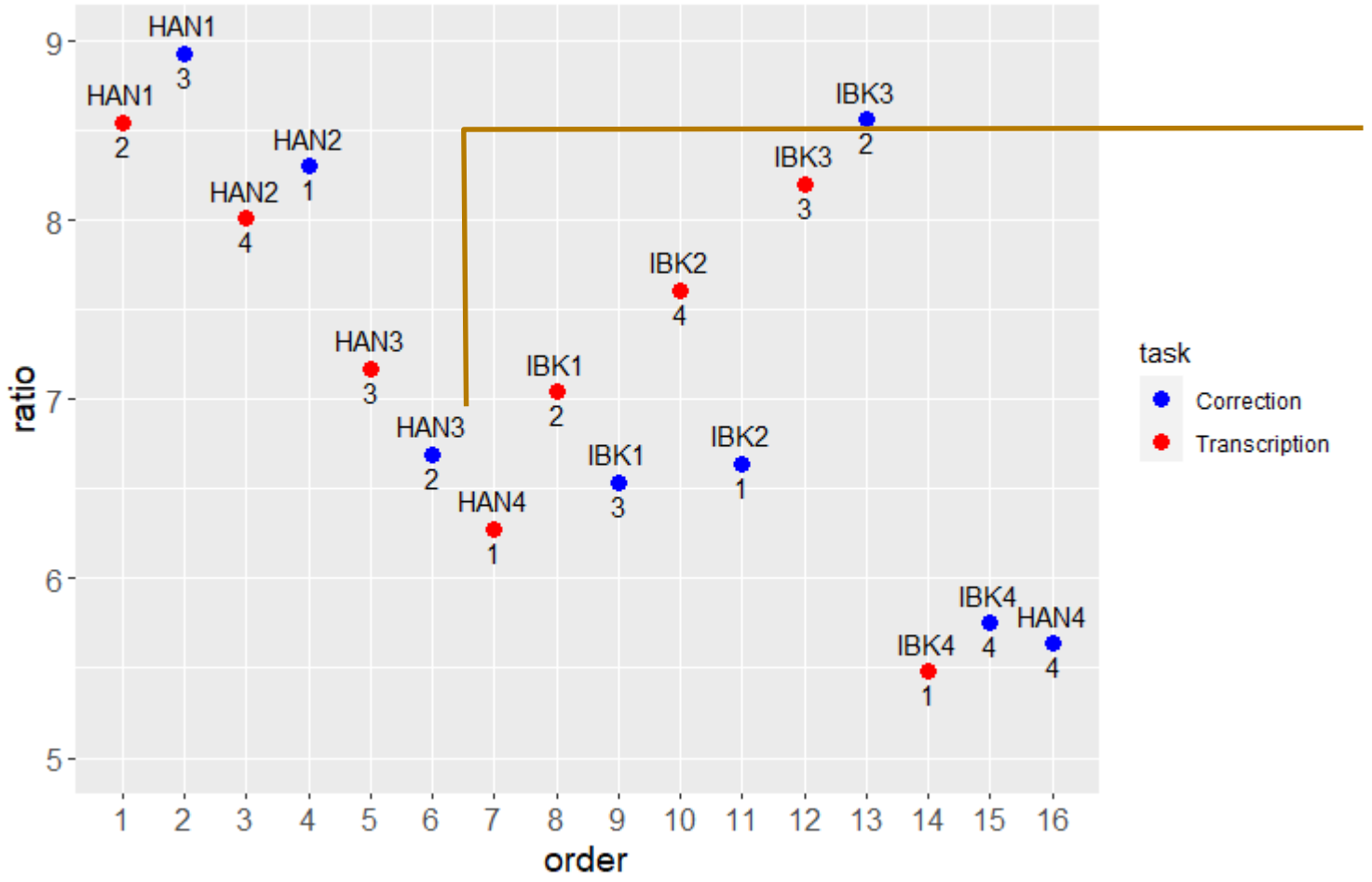
IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Paul (after training):

- “Transcribing is much more pleasant.”
- “it is easier to get into a rhythm/flow.”
- “At the correcting task one has to concentrate on several things at once.”
- “The alignment is often slightly off”
- “In most cases, backchannels and hesitation markers are missing”

RESULTS FROM ANNOTATOR PAUL

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

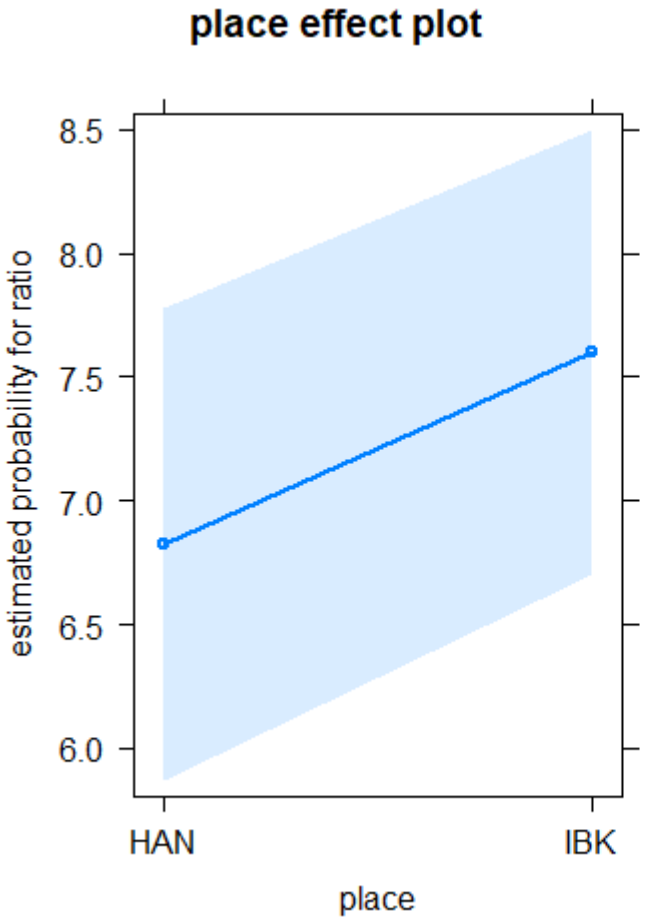


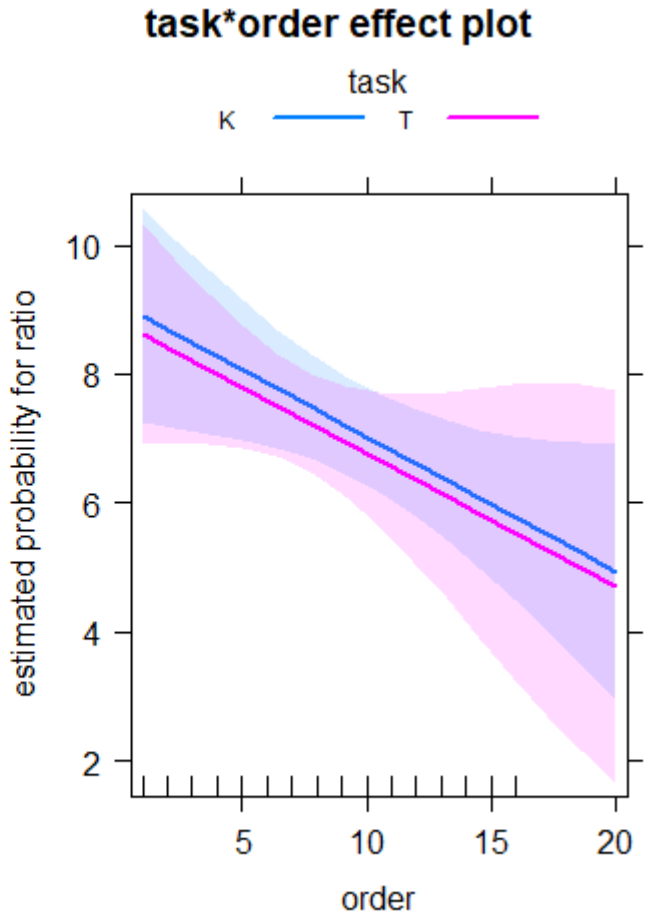
Paul (after 6 files): “in the meantime, I also got into a flow with correcting. I now know what to look for and thus have a better rhythm.”

RESULTS

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

```
mod <- lm(ratio ~ task * order + place, data = df)
```



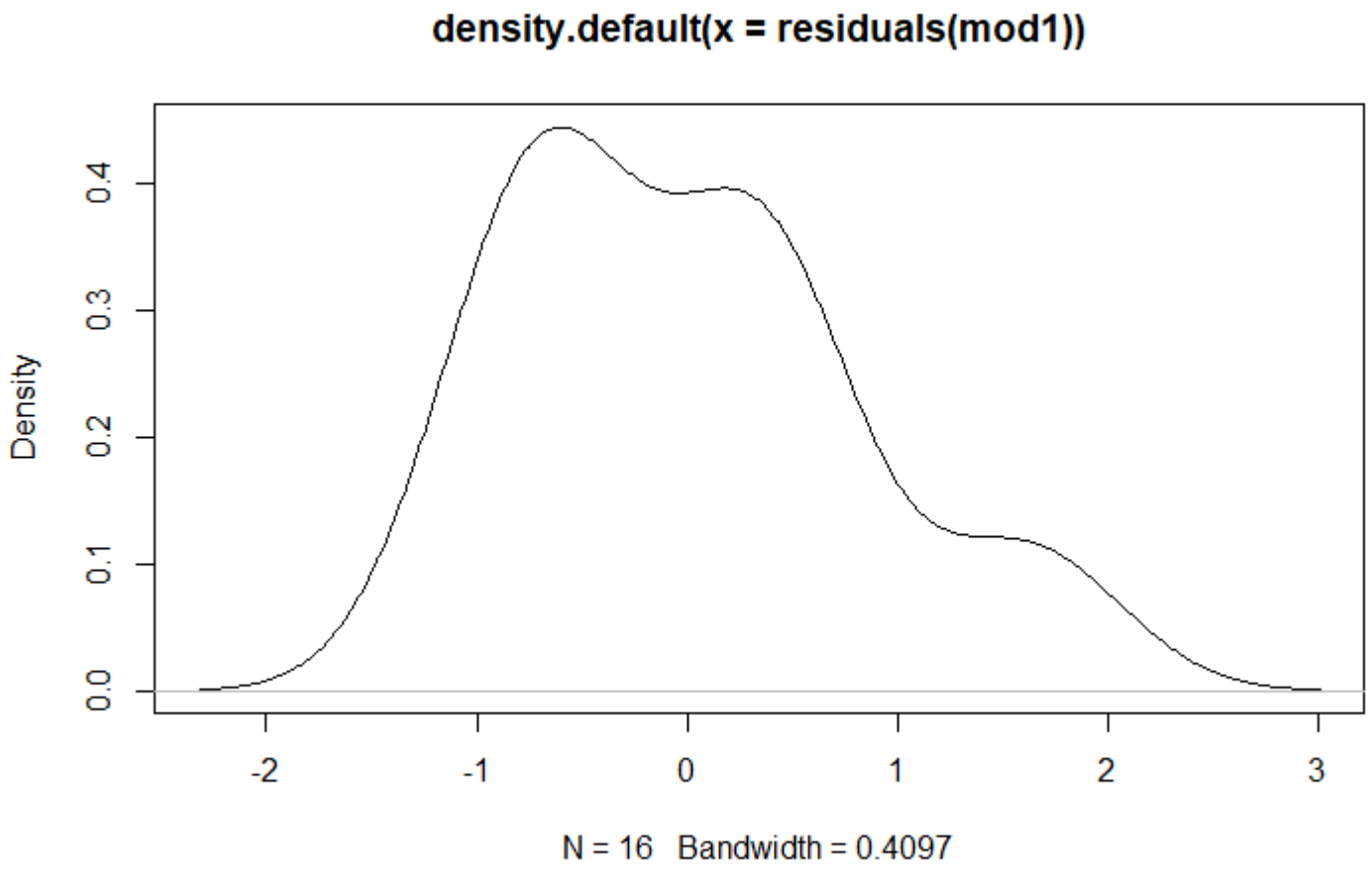


- Both transcribing and correcting seem to get faster over time
- Independent from the recording place (Hannover or Innsbruck)

DISTRIBUTION OF RESIDUALS

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

```
plot(density(residuals(mod1)))
```

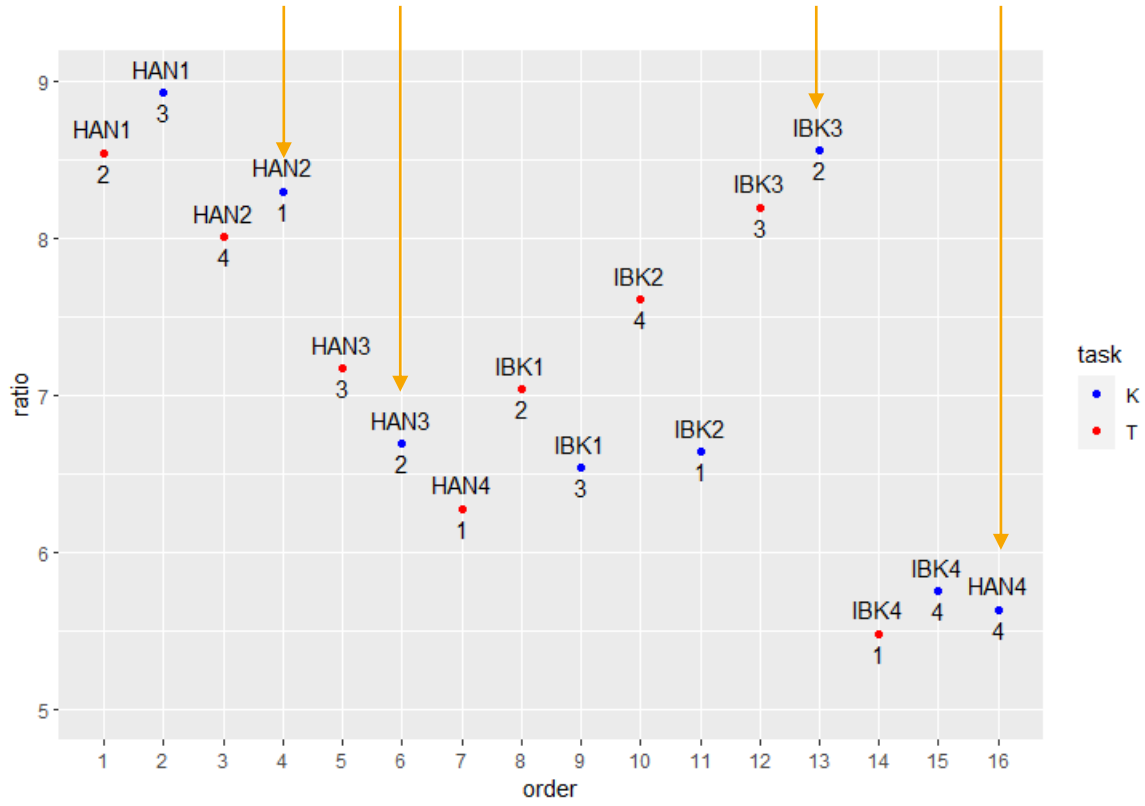


- Not really normally distributed
- !!! only 16 data points
- ??? we'll see with the next 16

COMMENTS FROM ANNOTATOR

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

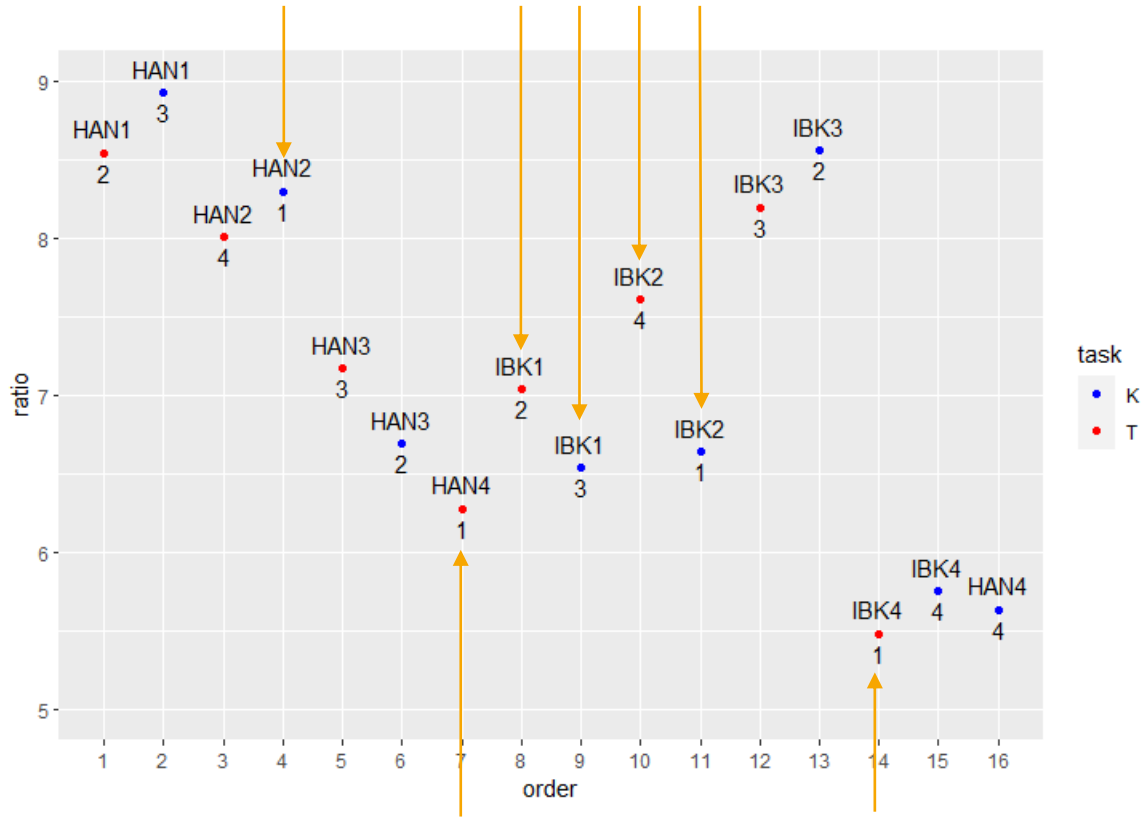
Speakers are very often on the wrong tier



COMMENTS FROM ANNOTATOR

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Much overlap

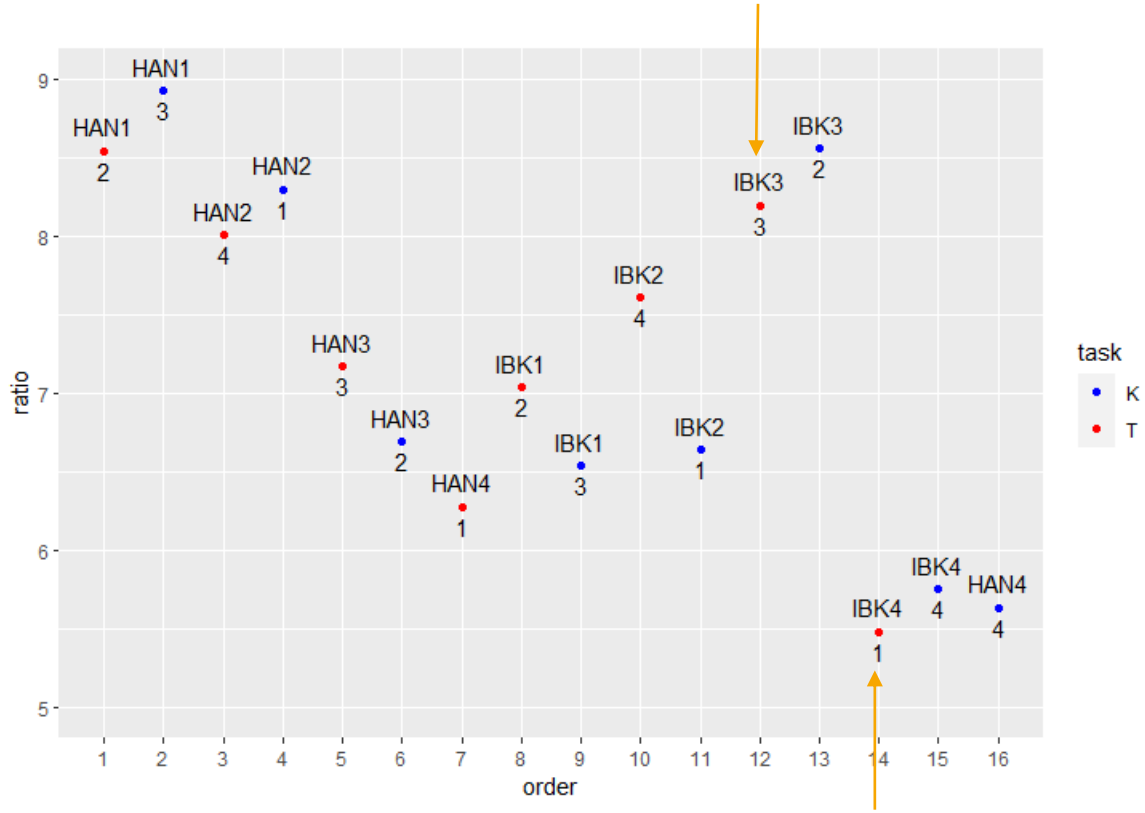


Little overlap

COMMENTS FROM ANNOTATOR

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

The interviewee speaks a lot

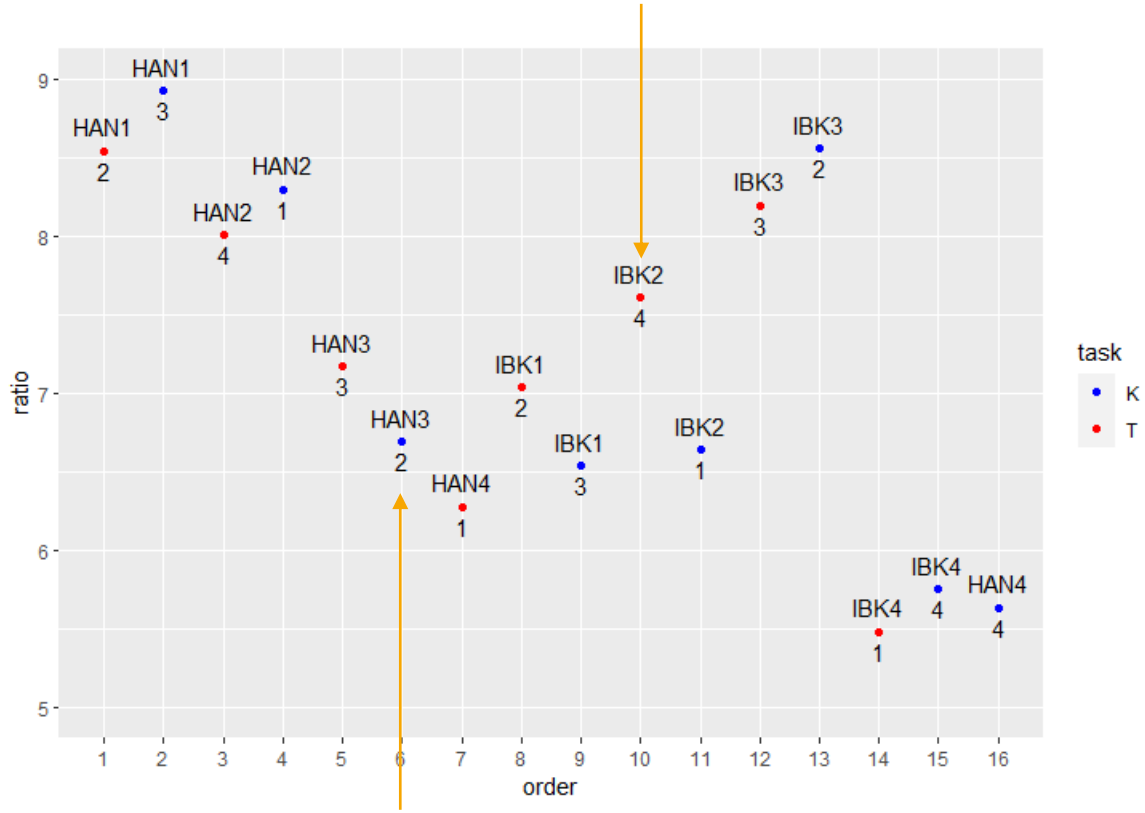


Speakers don’t speak much

COMMENTS FROM ANNOTATOR

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

The interviewee speaks extremely unclear and quiet



The interviewee speaks rather clear

COMMENTS FROM ANNOTATOR

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Word repetitions are almost never recognized by the ASR system