Working with Multimedia Data in CMC Corpora

International Conference on
CMC and Social Media Corpora
for the Humanities

Dr. Thomas Schmidt

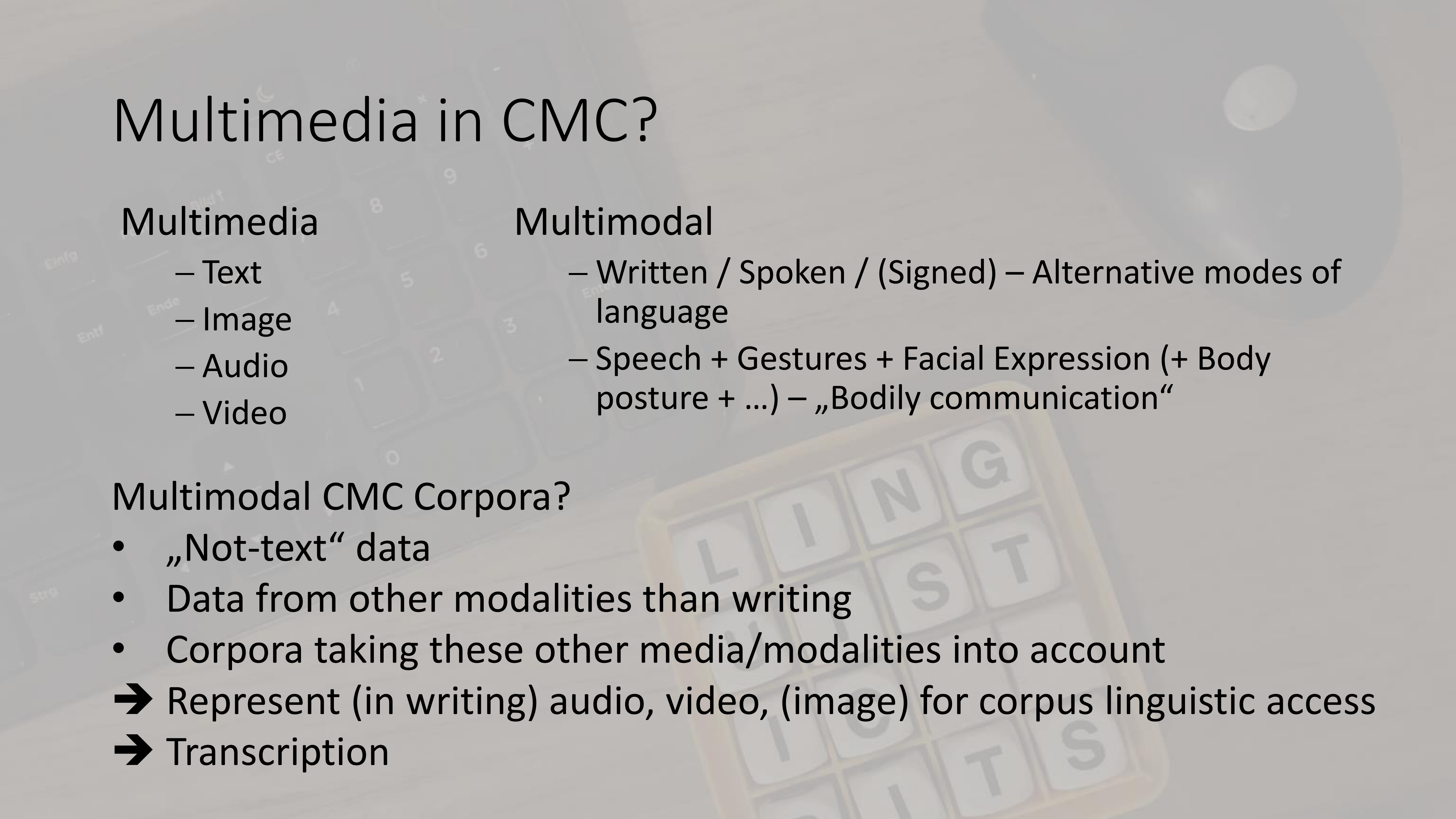
linguisticbits.de | Musical Bits GmbH

thomas@linguisticbits.de

Outline

1. Multimedia data in CMC?
2. Tools
   - Manual transcription and annotation
   - Automatic speech recognition
3. Formats, Standards, Interoperability
   - ISO/TEI standard
4. Q & A

Multimedia data in CMC?



Multimedia in CMC?

Multimedia
- Text
- Image
- Audio
- Video

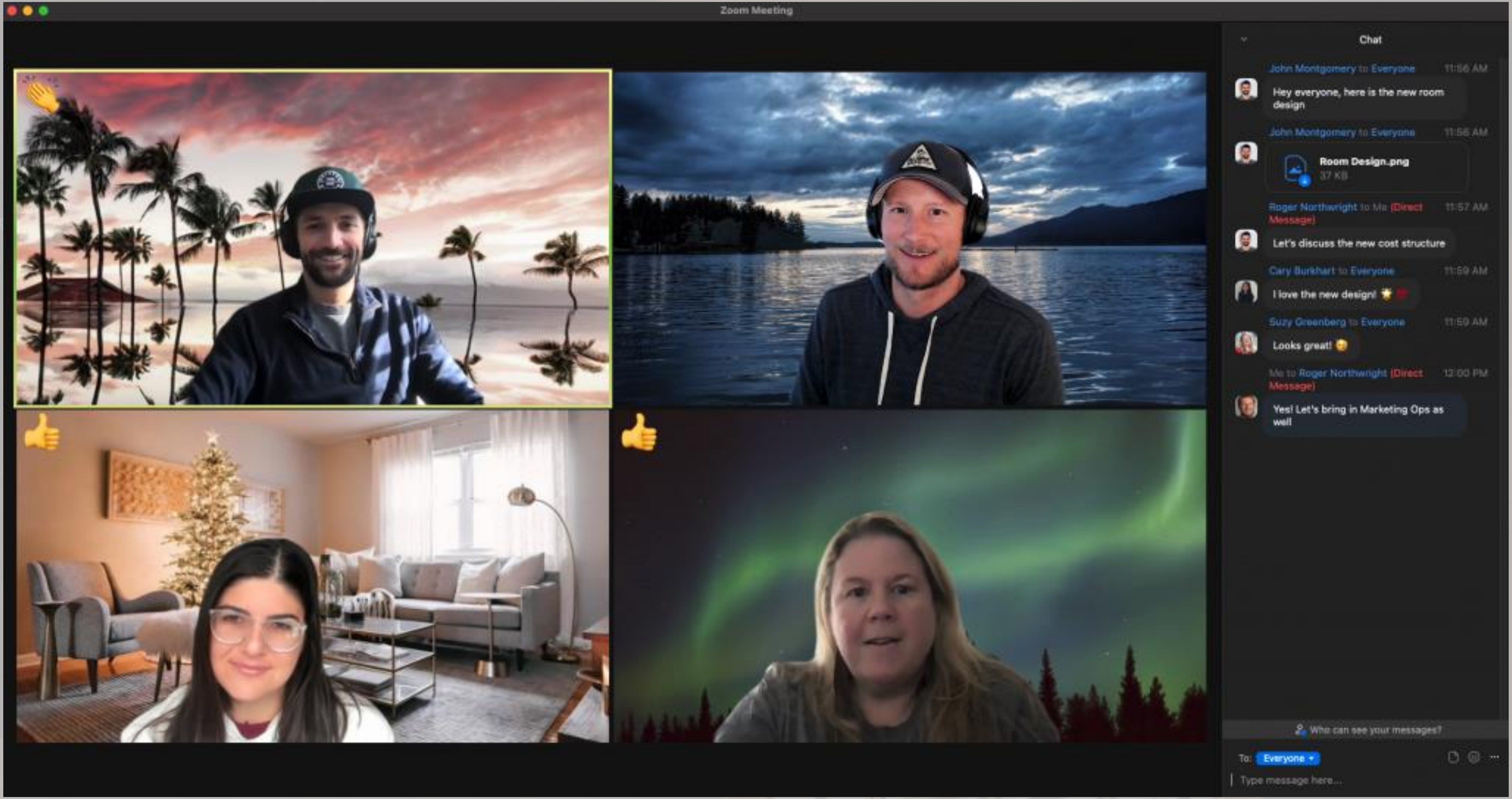
Multimodal
- Written / Spoken / (Signed) – Alternative modes of language
- Speech + Gestures + Facial Expression (+ Body posture + …) – „Bodily communication“

Multimodal CMC Corpora?

- „Not-text“ data
- Data from other modalities than writing
- Corpora taking these other media/modalities into account

➔ Represent (in writing) audio, video, (image) for corpus linguistic access

➔ Transcription



John Montgomery to Everyone 11:56 AM
Hey everyone, here is the new room design

John Montgomery to Everyone 11:56 AM
Room Design.png
37 KB

Roger Northwright to Me (Direct Message) 11:57 AM
Let's discuss the new cost structure

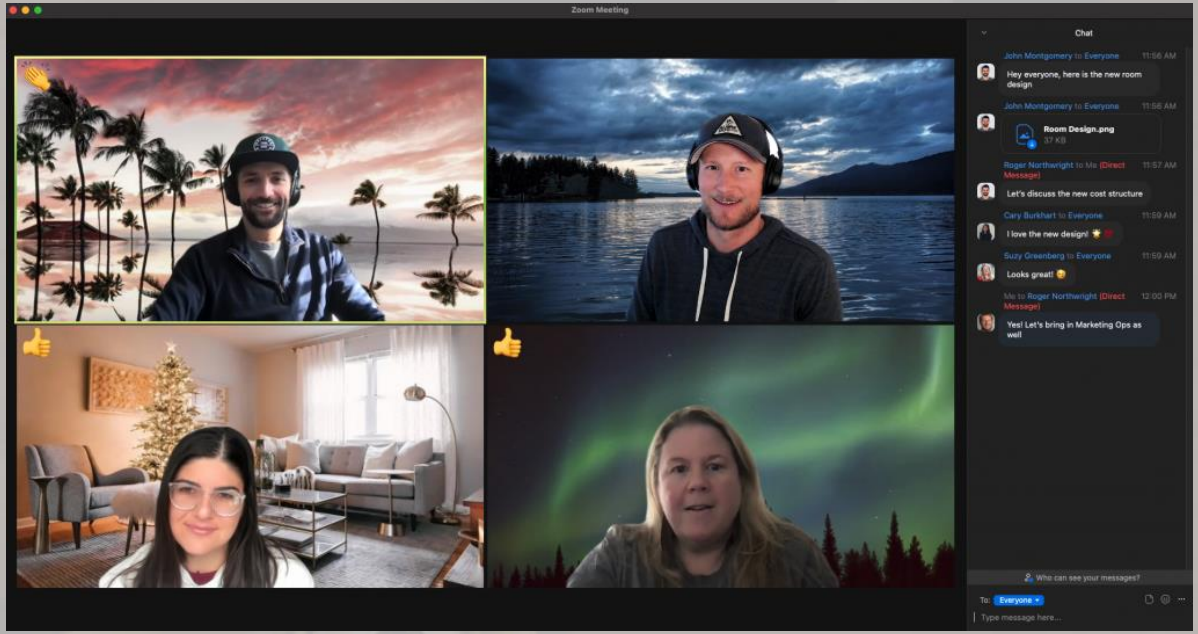
Cary Burkhart to Everyone 11:59 AM
I love the new design! 🌟 ❤

Suzy Greenberg to Everyone 11:59 AM
Looks great! 😊

Me to Roger Northwright (Direct Message) 12:00 PM
Yes! Let's bring in Marketing Ops as well

Video conferencing

- Text (chat)
- Video incl. audio
- Speech in video
- Gesture + facial expression in video
- Emoticons in video
- Simultaneity of text and video



John Montgomery to Everyone 11:56 AM
Hey everyone, here is the new room design

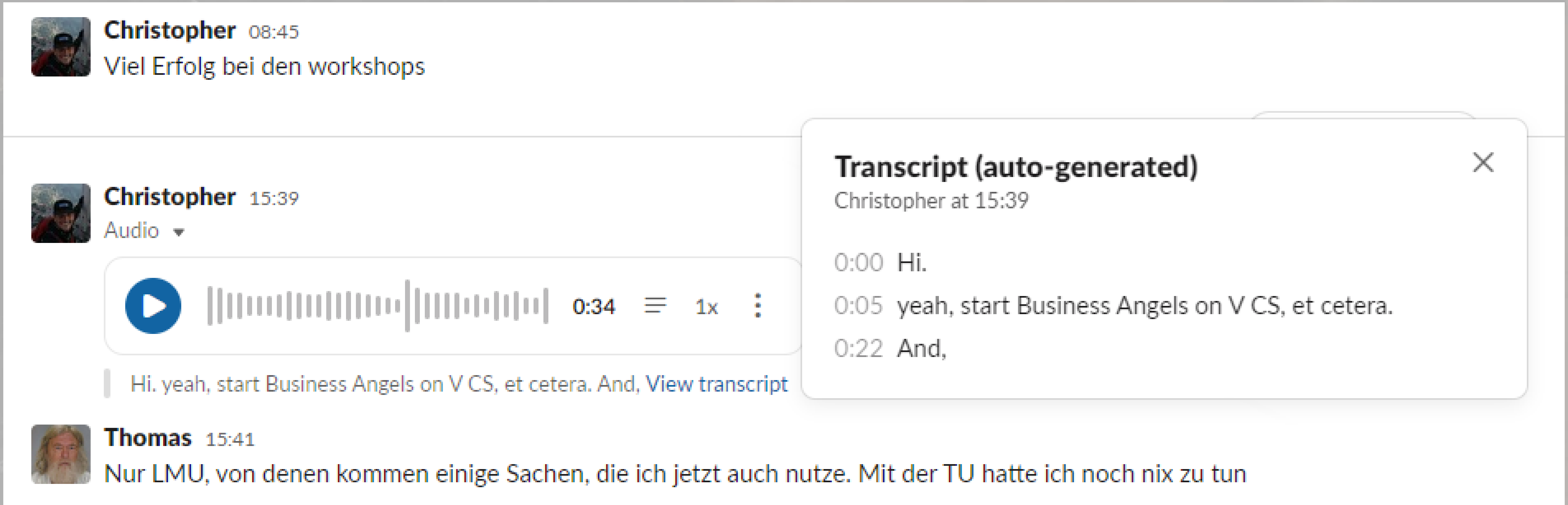
John Montgomery to Everyone 11:56 AM
Room Design.png
37 KB

Roger Northwright to Me (Direct Message) 11:57 AM
Let's discuss the new cost structure

Cary Burkhart to Everyone 11:59 AM
I love the new design! 🌟 ❤️

Suzy Greenberg to Everyone 11:59 AM
Looks great! 😊

Me to Roger Northwright (Direct Message) 12:00 PM
Yes! Let's bring in Marketing Ops as well



Christopher 08:45
Viel Erfolg bei den workshops

Christopher 15:39
Audio

Hi. yeah, start Business Angels on V CS, et cetera. And, View transcript

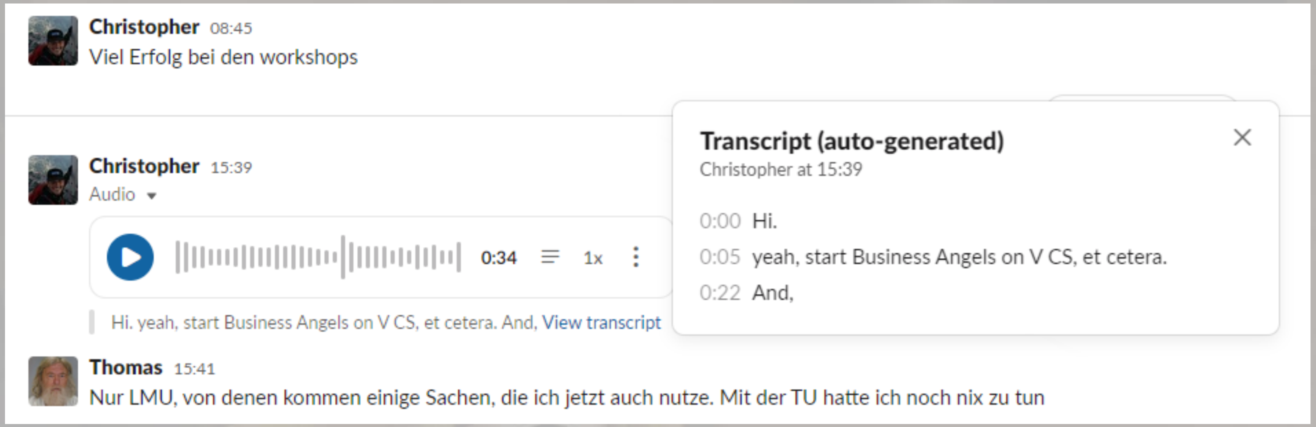
Transcript (auto-generated)
Christopher at 15:39

0:00 Hi.
0:05 yeah, start Business Angels on V CS, et cetera.
0:22 And,

Thomas 15:41
Nur LMU, von denen kommen einige Sachen, die ich jetzt auch nutze. Mit der TU hatte ich noch nix zu tun

Slack channel

- Text
- Audio (voice messages)
- Derived text (auto-generated transcript)
- Alternation between text and audio



Christopher 08:45
Viel Erfolg bei den workshops

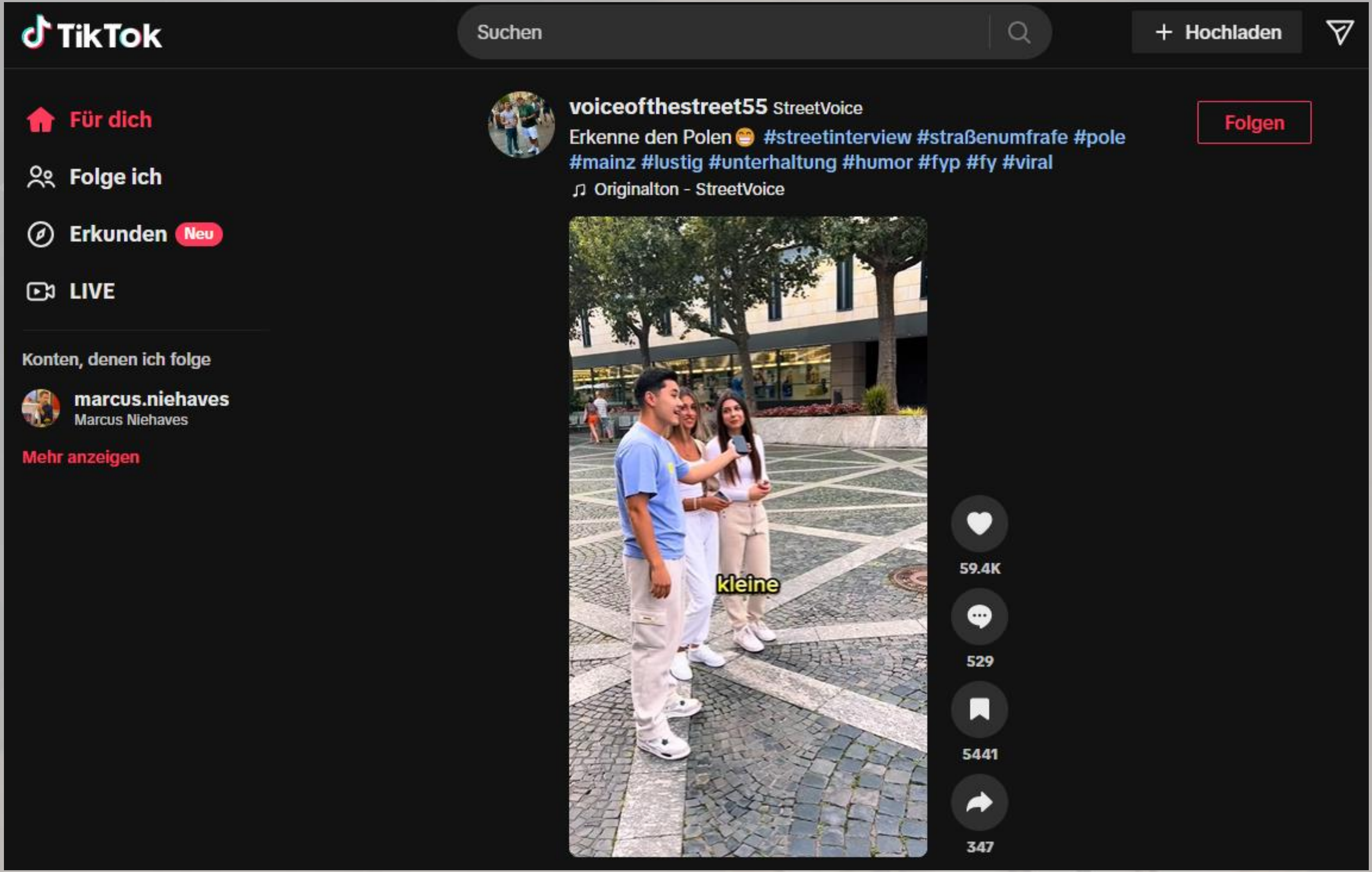
Christopher 15:39
Audio ▾

Hi. yeah, start Business Angels on V CS, et cetera. And, View transcript

Transcript (auto-generated)
Christopher at 15:39

0:00 Hi.
0:05 yeah, start Business Angels on V CS, et cetera.
0:22 And,

Thomas 15:41
Nur LMU, von denen kommen einige Sachen, die ich jetzt auch nutze. Mit der TU hatte ich noch nix zu tun



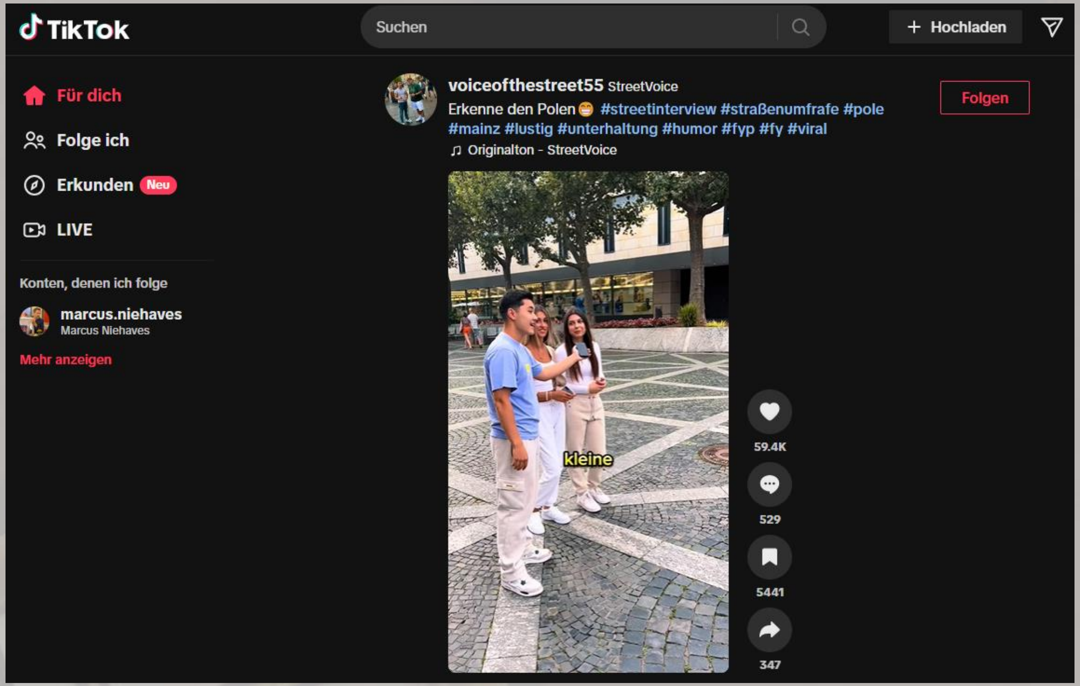
voiceofthestreet55 StreetVoice

Erkenne den Polen😁 #streetinterview #straßenumfrafe #pole #mainz #lustig #unterhaltung #humor #fyp #fy #viral

♫ Originalton - StreetVoice

TikTok post

- Text
- Video
- Text refers to the video



voiceofthestreet55 StreetVoice

Erkenne den Polen😁 #streetinterview #straßenumfrafe #pole #mainz #lustig #unterhaltung #humor #fyp #fy #viral

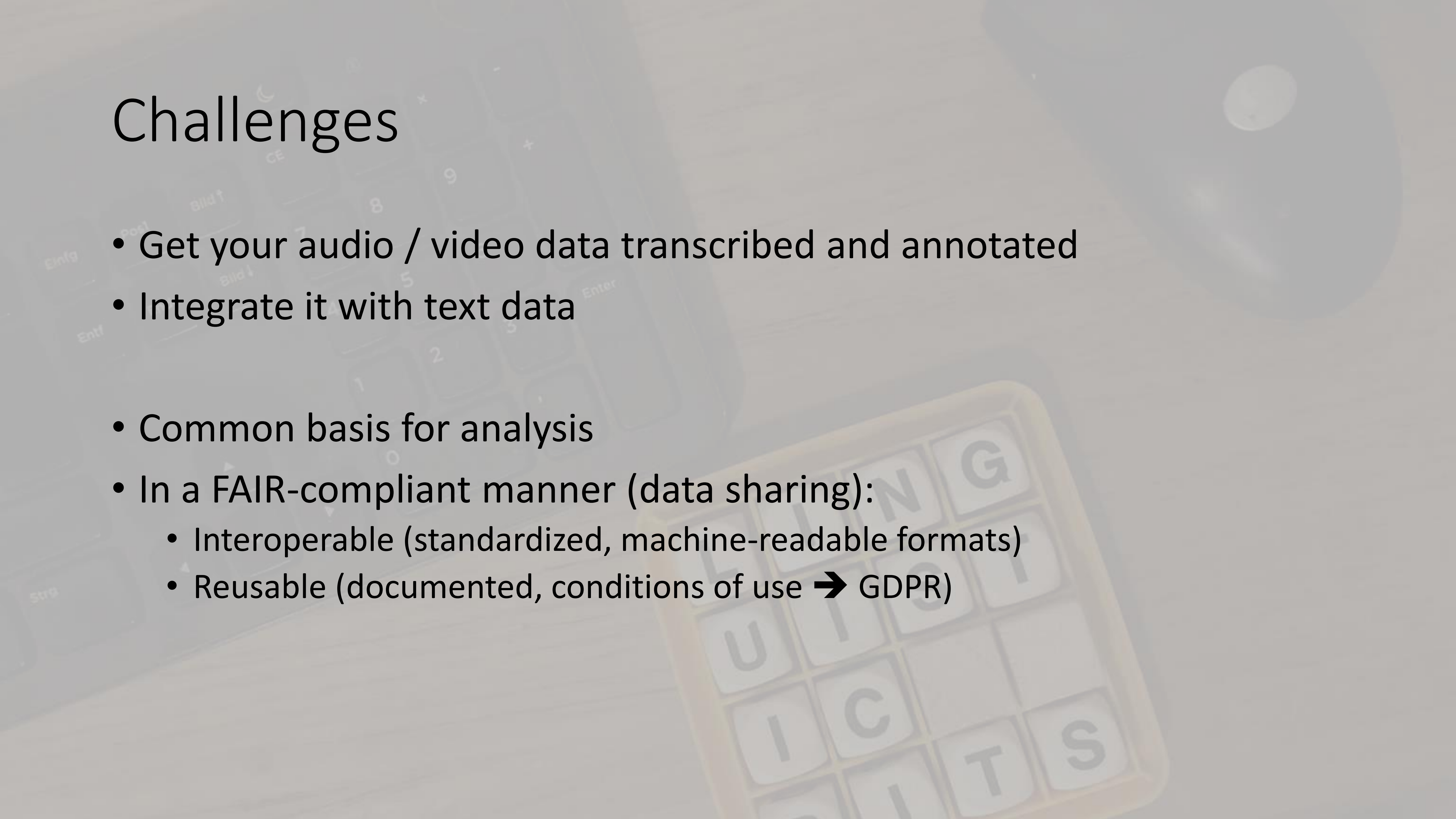
♫ Originalton - StreetVoice

This conference

- Podcasts [Babayode et al.]
- Audio/Video of Zoom and face-to-face-meetings (as comparison), interviews [Steinsiek]
- Video-conferencing in Zoom [Pabst et al.]
- Comments on Bilibili videos [Zheng]
- Podcasts (vs. blog posts) [Seemann et al.]
- Spoken corpora, gaze or walk annotations, kinect or motion capture data [Ferger et al.]
- Online video film reviews [Piroh]
- Short video data on TikTok [Helenius]
- Videos of Authentic Social Interaction [Krause et. al]
- Audio and video data from video sharing sites, streaming services and social media platforms [Coats]
- Multimodal WhatsApp discussion [Mäkinen]

- Some mono-modal (podcasts)
- Audio/video of very different durations (seconds to hours)
- Different status of audio/video
- Comparison CMC

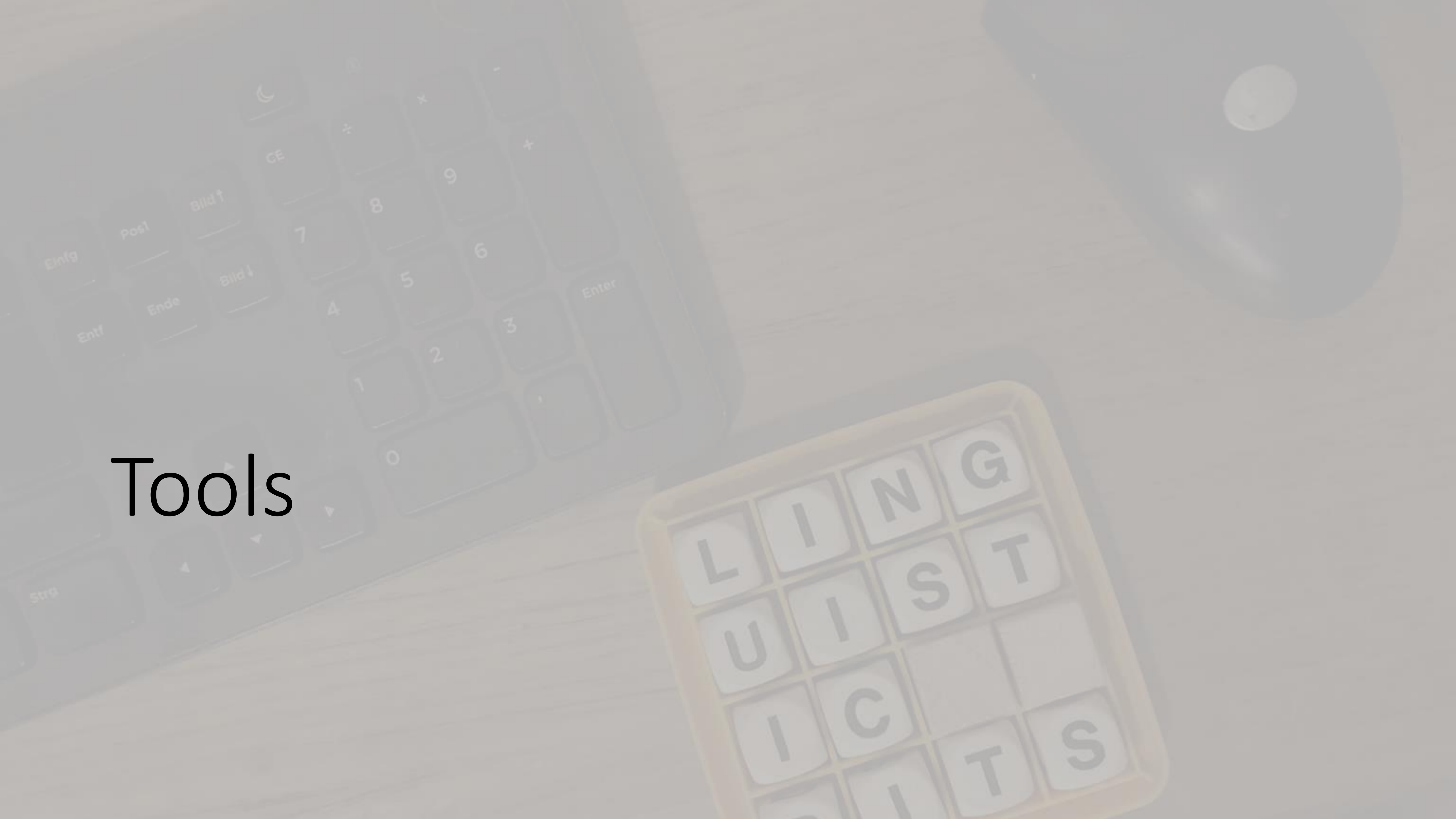
<> Face-to-Face interaction



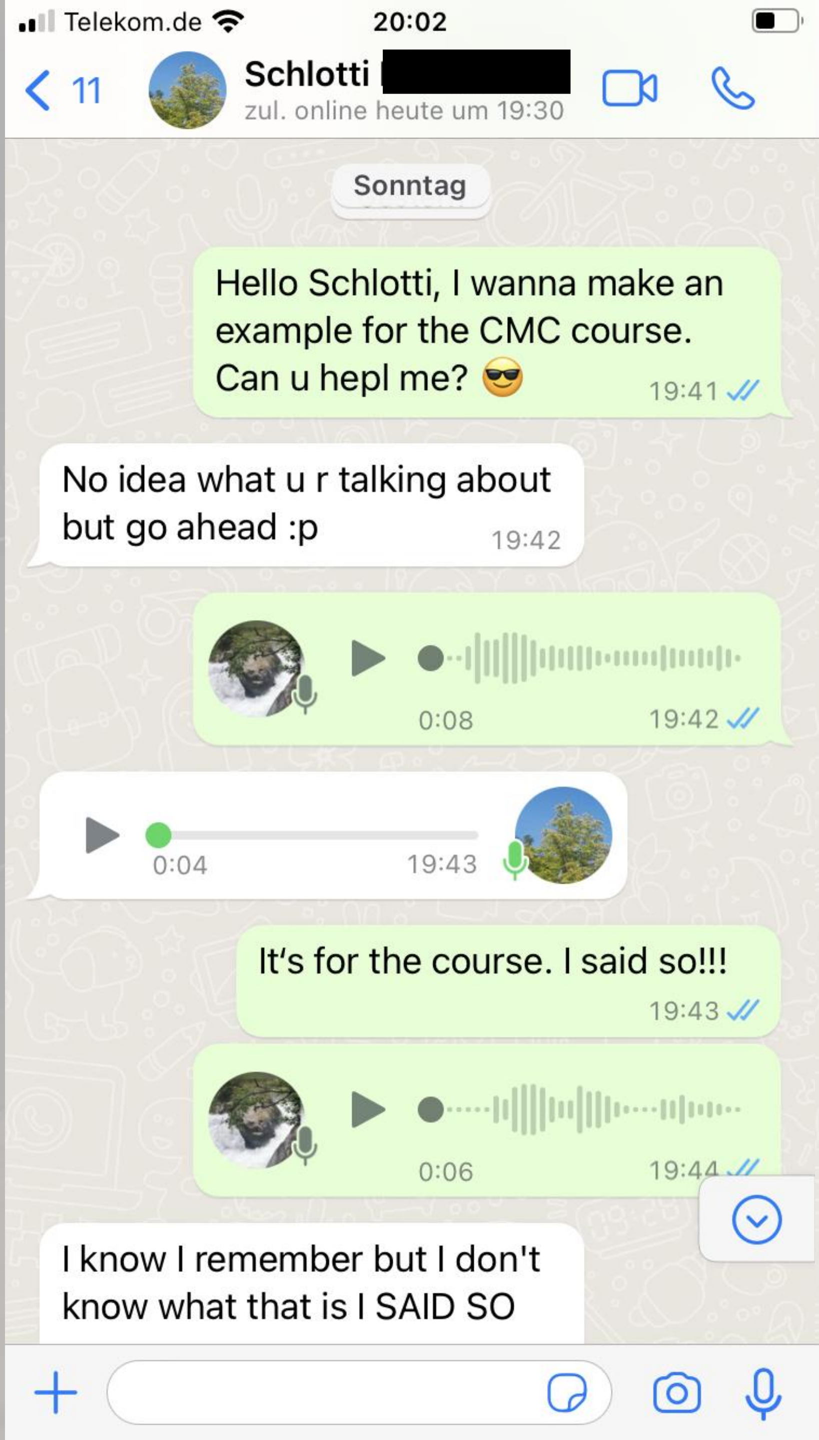
Challenges

- Get your audio / video data transcribed and annotated
- Integrate it with text data

- Common basis for analysis
- In a FAIR-compliant manner (data sharing):
  - Interoperable (standardized, machine-readable formats)
  - Reusable (documented, conditions of use ➔ GDPR)



Tools



Telekom.de 20:02

< 11 Schlotti [redacted]

zul. online heute um 19:30

Sonntag

Hello Schlotti, I wanna make an example for the CMC course. Can u hepl me? 😎 19:41

No idea what u r talking about but go ahead :p 19:42

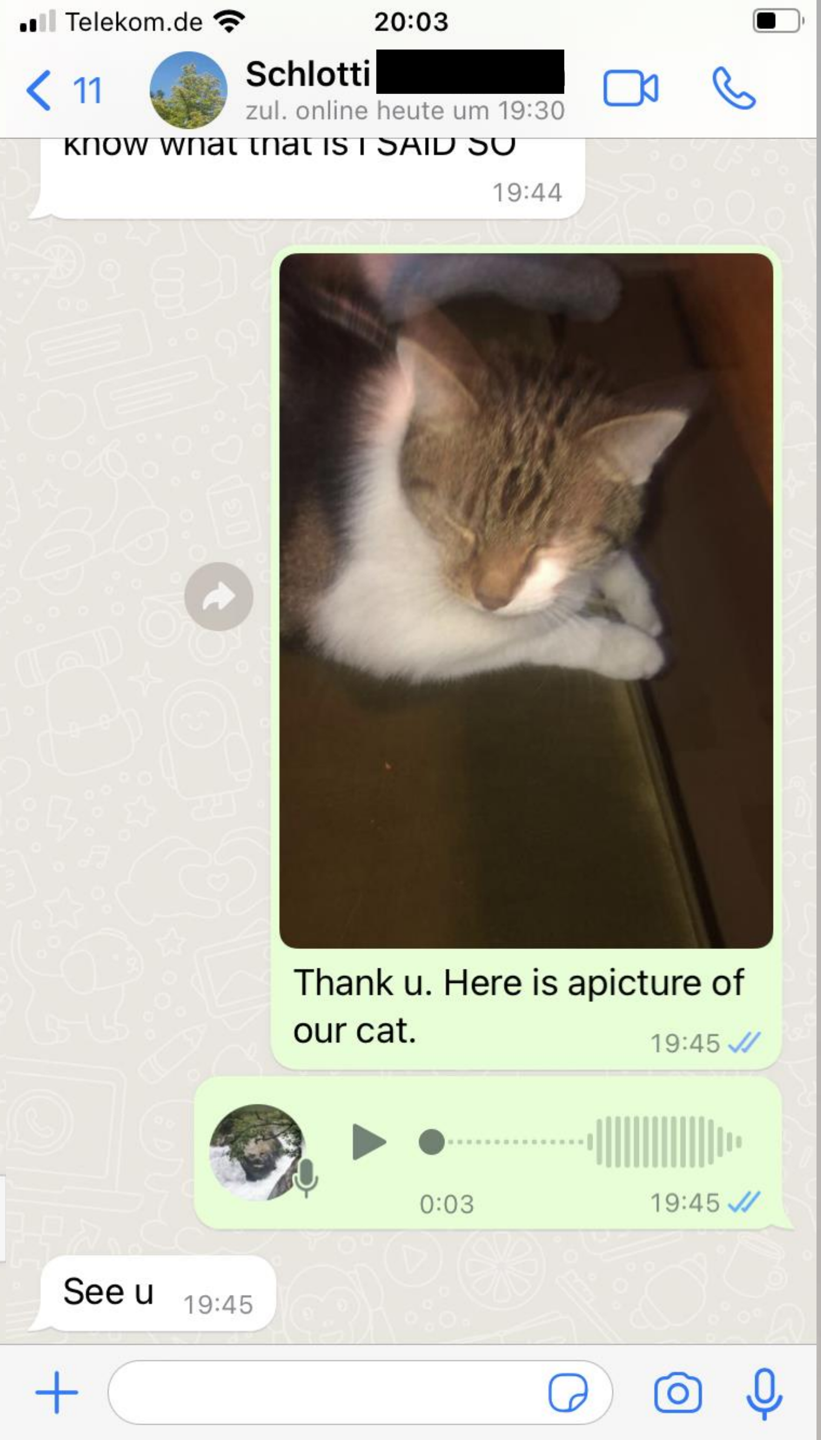
0:08 19:42

0:04 19:43

It‘s for the course. I said so!!! 19:43

0:06 19:44

I know I remember but I don't know what that is I SAID SO



Telekom.de 20:03

< 11 Schlotti [redacted]

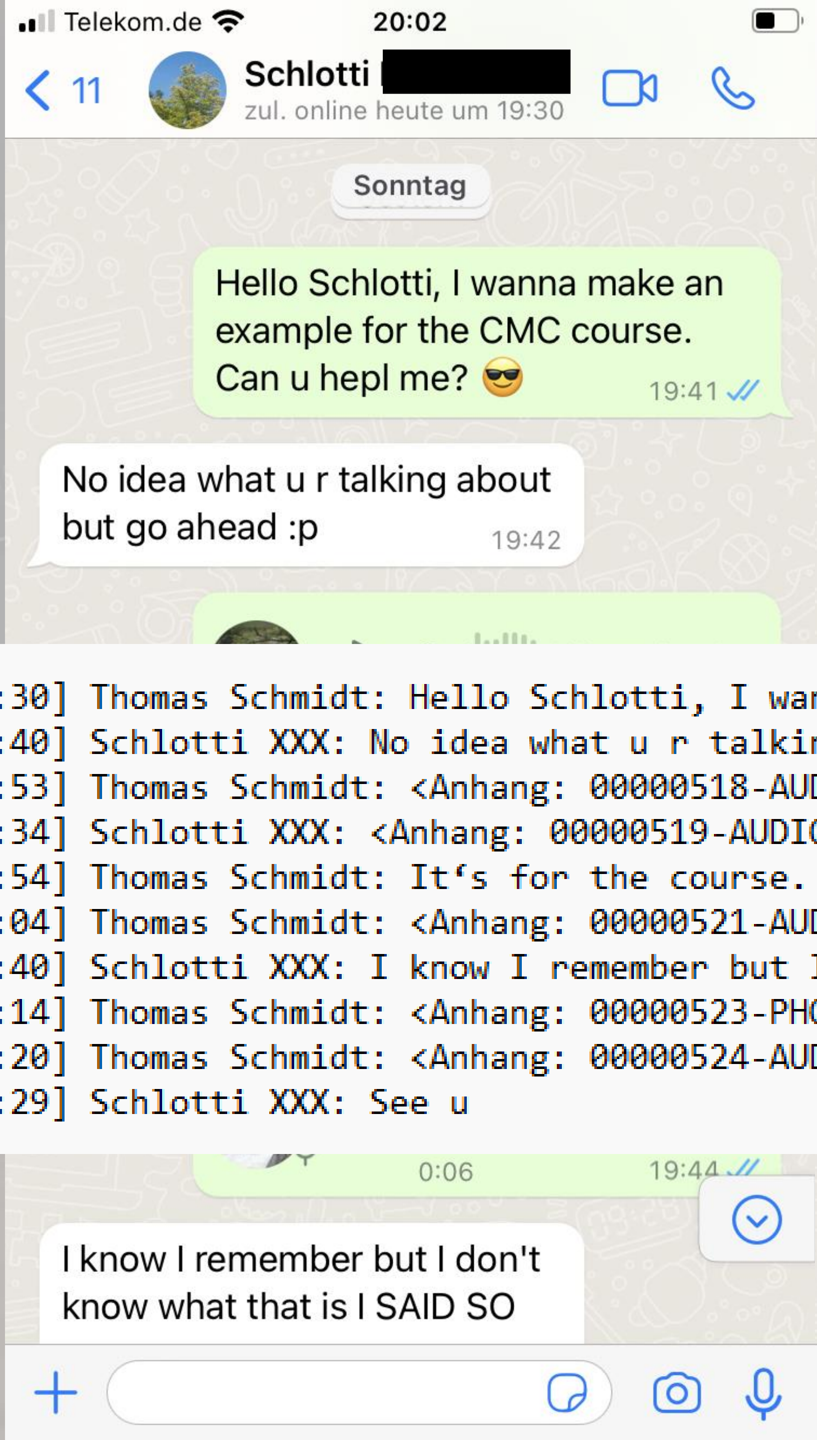
zul. online heute um 19:30

know what that is I SAID SO 19:44

Thank u. Here is apicture of our cat. 19:45

0:03 19:45

See u 19:45

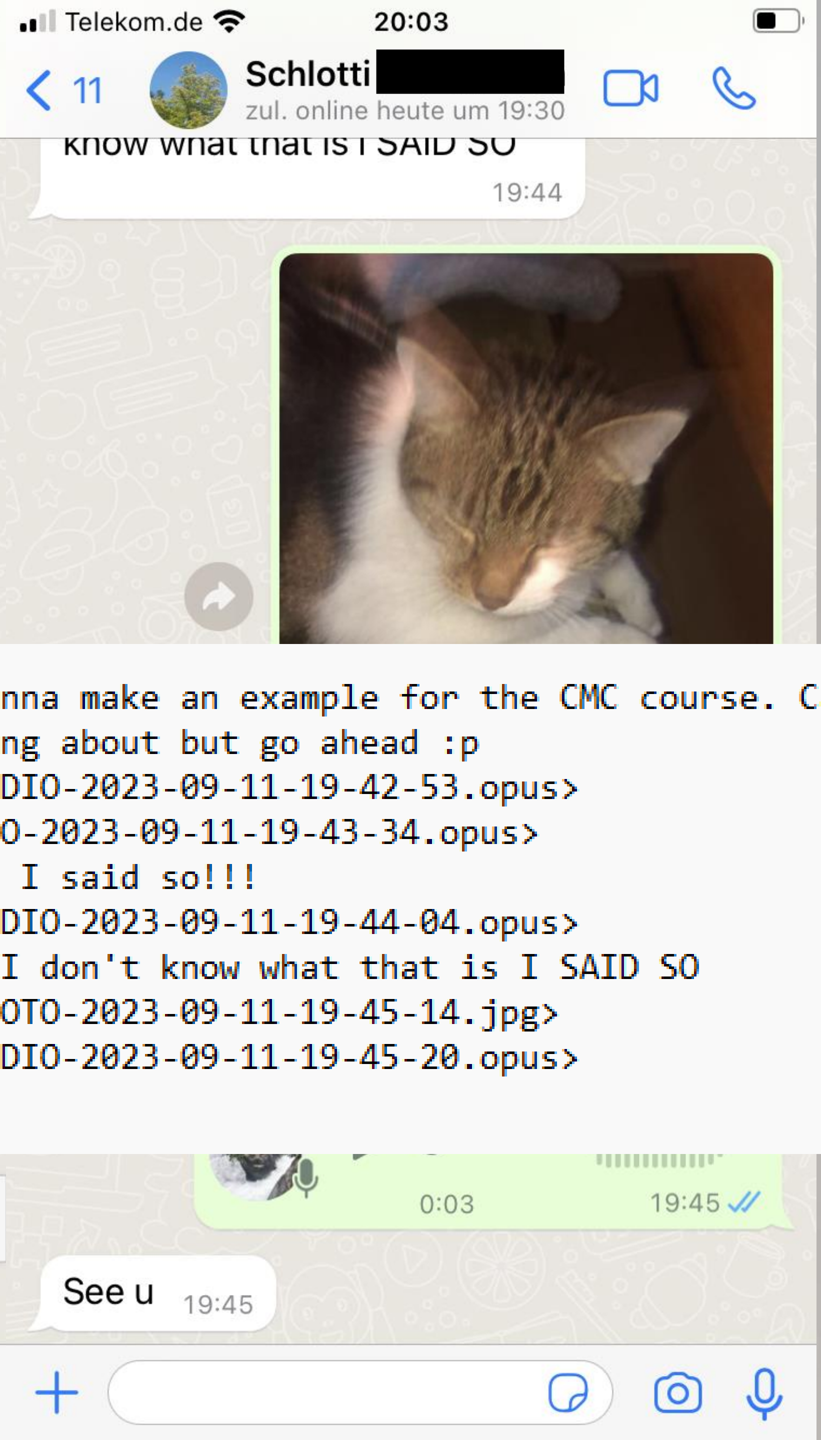


Hello Schlotti, I wanna make an example for the CMC course. Can u hepl me? 😎 19:41

No idea what u r talking about but go ahead :p 19:42

0:06 19:44

I know I remember but I don't know what that is I SAID SO

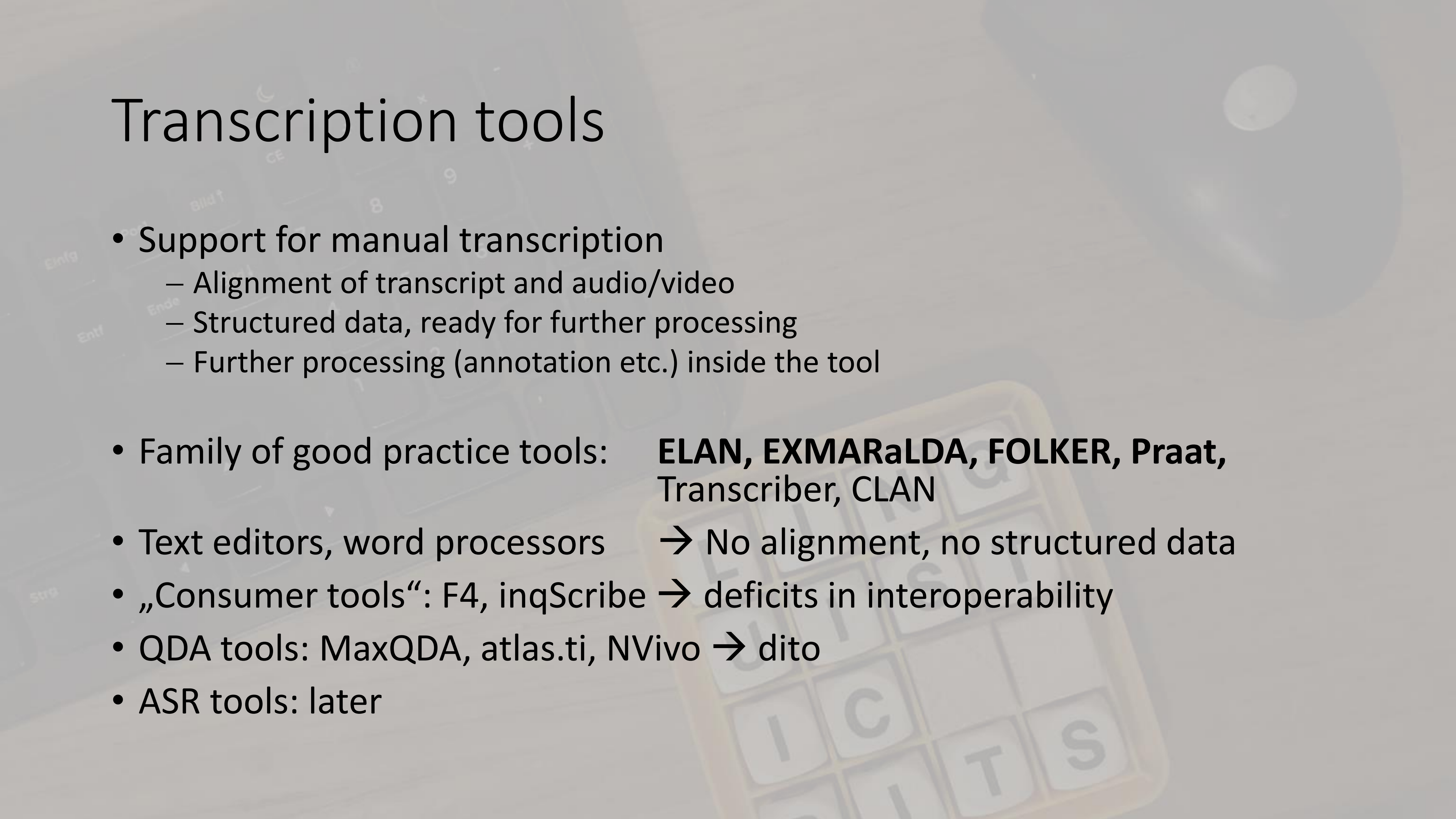


know what that is I SAID SO 19:44

0:03 19:45

See u 19:45

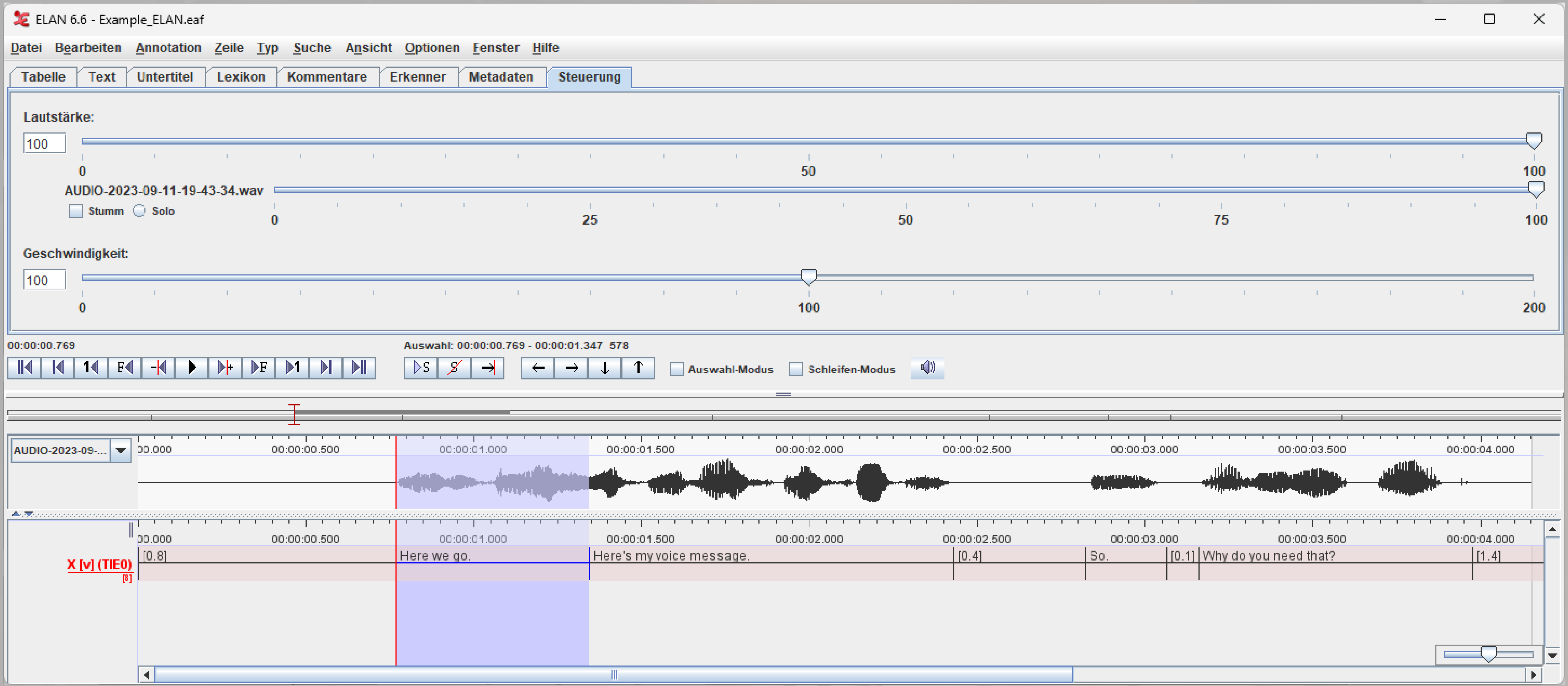
```
[11.09.23, 19:41:30] Thomas Schmidt: Hello Schlotti, I wanna make an example for the CMC course. Can u hepl me? 😎
[11.09.23, 19:42:40] Schlotti XXX: No idea what u r talking about but go ahead :p
[11.09.23, 19:42:53] Thomas Schmidt: <Anhang: 00000518-AUDIO-2023-09-11-19-42-53.opus>
[11.09.23, 19:43:34] Schlotti XXX: <Anhang: 00000519-AUDIO-2023-09-11-19-43-34.opus>
[11.09.23, 19:43:54] Thomas Schmidt: It‘s for the course. I said so!!!
[11.09.23, 19:44:04] Thomas Schmidt: <Anhang: 00000521-AUDIO-2023-09-11-19-44-04.opus>
[11.09.23, 19:44:40] Schlotti XXX: I know I remember but I don't know what that is I SAID SO
[11.09.23, 19:45:14] Thomas Schmidt: <Anhang: 00000523-PHOTO-2023-09-11-19-45-14.jpg>
[11.09.23, 19:45:20] Thomas Schmidt: <Anhang: 00000524-AUDIO-2023-09-11-19-45-20.opus>
[11.09.23, 19:45:29] Schlotti XXX: See u
```



Transcription tools

- Support for manual transcription
  - Alignment of transcript and audio/video
  - Structured data, ready for further processing
  - Further processing (annotation etc.) inside the tool

- Family of good practice tools: **ELAN, EXMARaLDA, FOLKER, Praat,** Transcriber, CLAN
- Text editors, word processors → No alignment, no structured data
- „Consumer tools“: F4, inqScribe → deficits in interoperability
- QDA tools: MaxQDA, atlas.ti, NVivo → dito
- ASR tools: later



Lautstärke: 100

AUDIO-2023-09-11-19-43-34.wav

Geschwindigkeit: 100

00:00:00.769

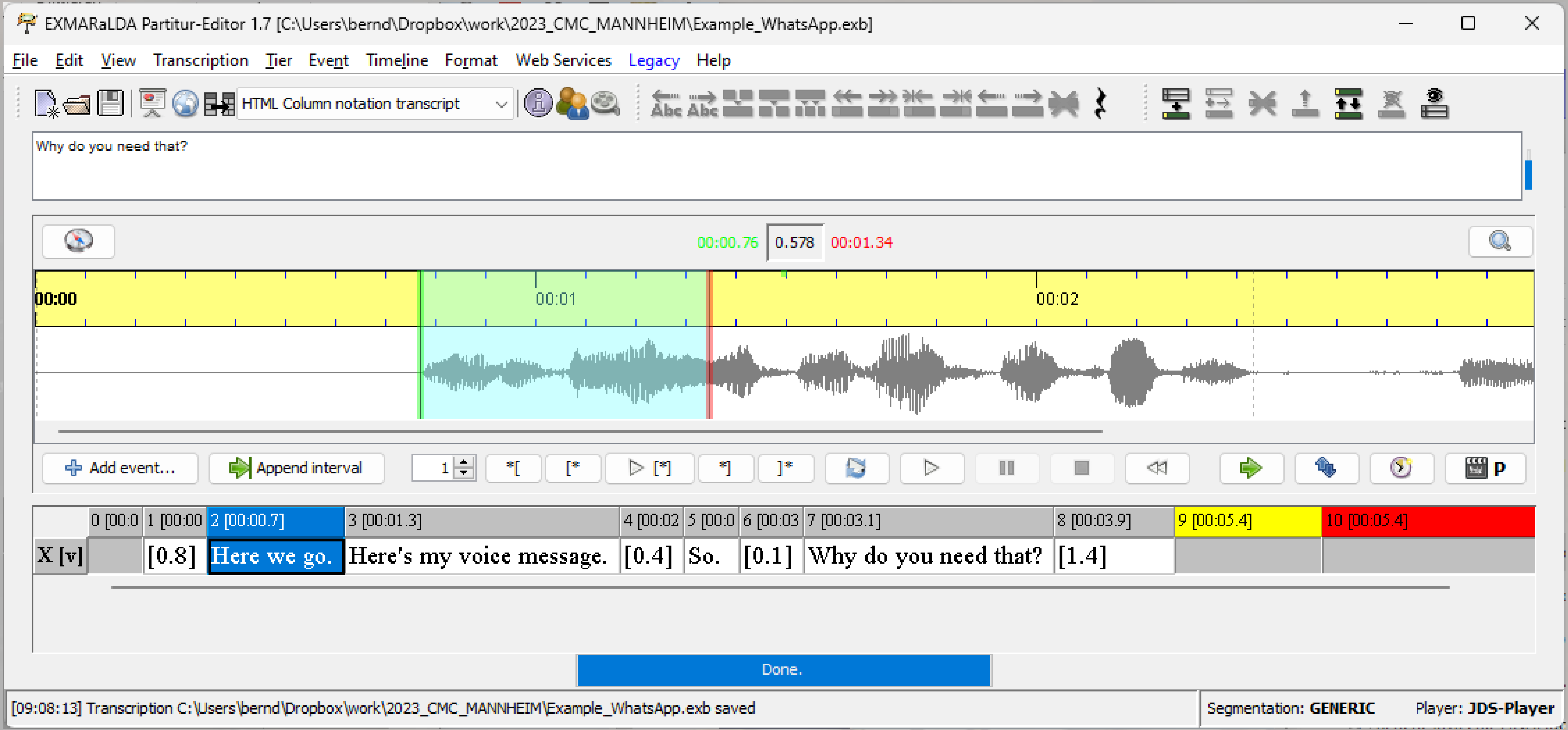
Auswahl: 00:00:00.769 - 00:00:01.347 578

□ Auswahl-Modus □ Schleifen-Modus

[0.8]	Here we go.	Here's my voice message.	[0.4]	So.	[0.1]	Why do you need that?	[1.4]

ELAN (Eudico Linguistic Annotator)

https://archive.mpi.nl/tla/elan



Why do you need that?

	0 [00:0	1 [00:00	2 [00:00.7]	3 [00:01.3]	4 [00:02	5 [00:0	6 [00:03	7 [00:03.1]	8 [00:03.9]	9 [00:05.4]	10 [00:05.4]
X [v]		[0.8]	Here we go.	Here's my voice message.	[0.4]	So.	[0.1]	Why do you need that?	[1.4]		

Done.

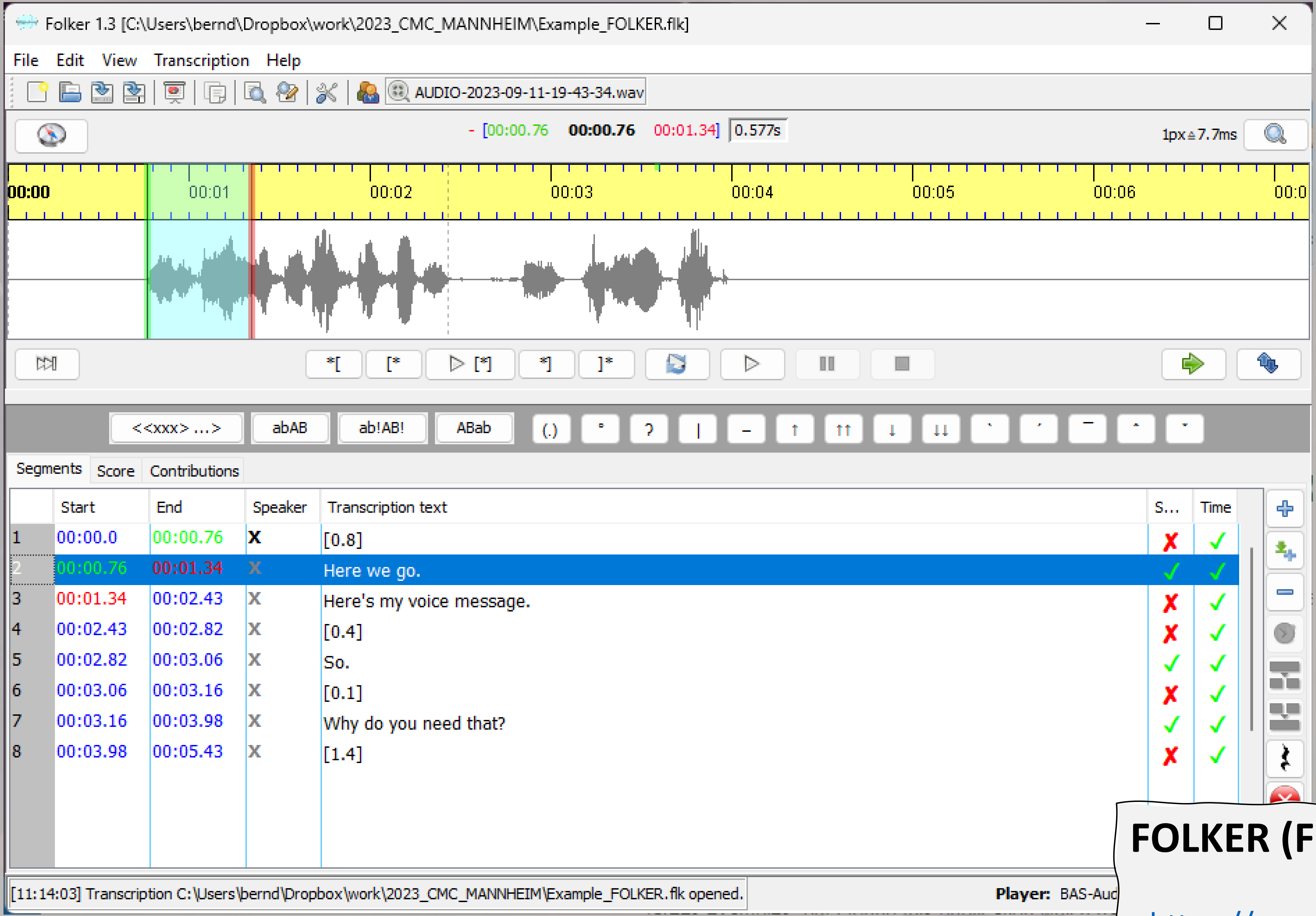
[09:08:13] Transcription C:\Users\bernd\Dropbox\work\2023_CMC_MANNHEIM\Example_WhatsApp.exb saved

Segmentation: **GENERIC**

Player: **JDS-Player**

EXMARaLDA Partitur-Editor

https://www.exmaralda.org



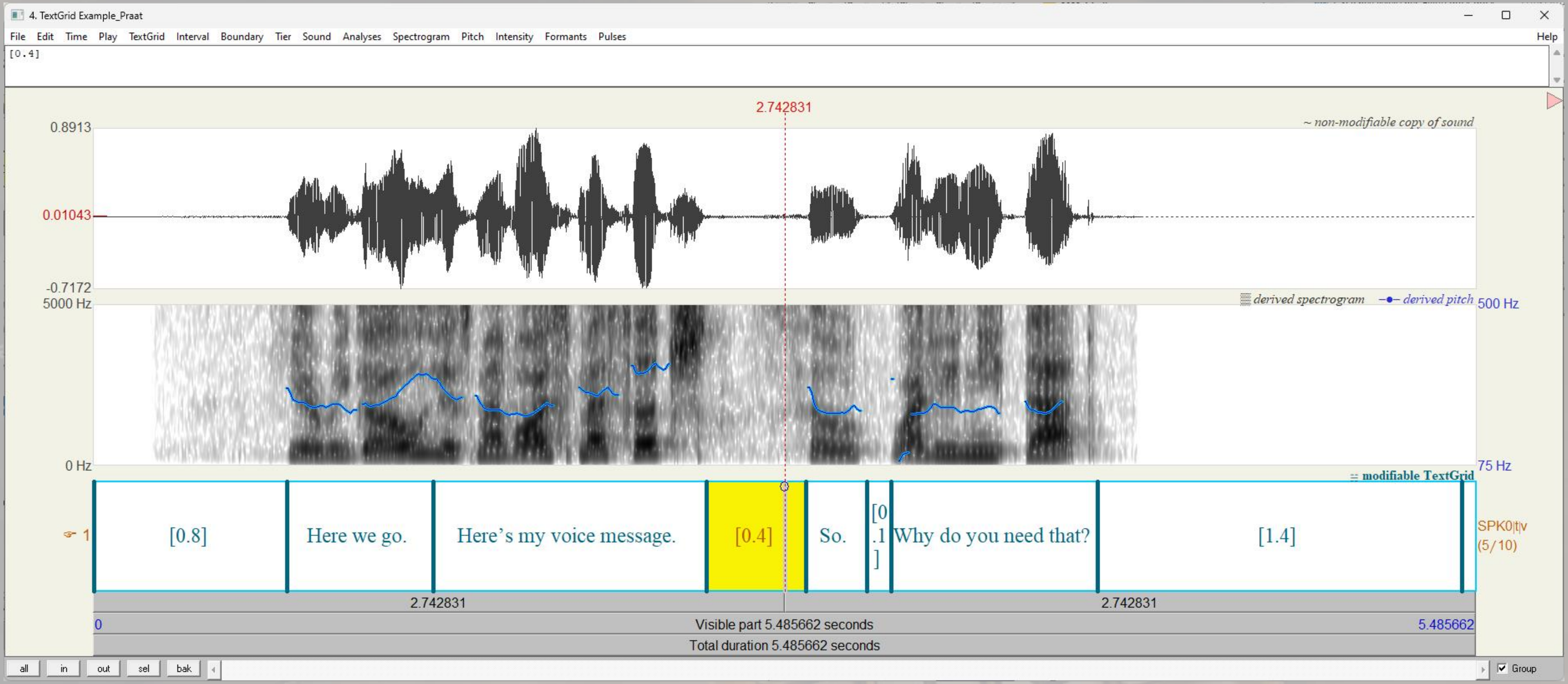
	Start	End	Speaker	Transcription text	S...	Time
1	00:00.0	00:00.76	X	[0.8]	✗	✓
2	00:00.76	00:01.34	X	Here we go.	✓	✓
3	00:01.34	00:02.43	X	Here's my voice message.	✗	✓
4	00:02.43	00:02.82	X	[0.4]	✗	✓
5	00:02.82	00:03.06	X	So.	✓	✓
6	00:03.06	00:03.16	X	[0.1]	✗	✓
7	00:03.16	00:03.98	X	Why do you need that?	✓	✓
8	00:03.98	00:05.43	X	[1.4]	✗	✓

[11:14:03] Transcription C:\Users\bernd\Dropbox\work\2023_CMC_MANNHEIM\Example_FOLKER.flk opened.

Player: BAS-Aud

FOLKER (FOLK-Editor)

https://www.exmaralda.org

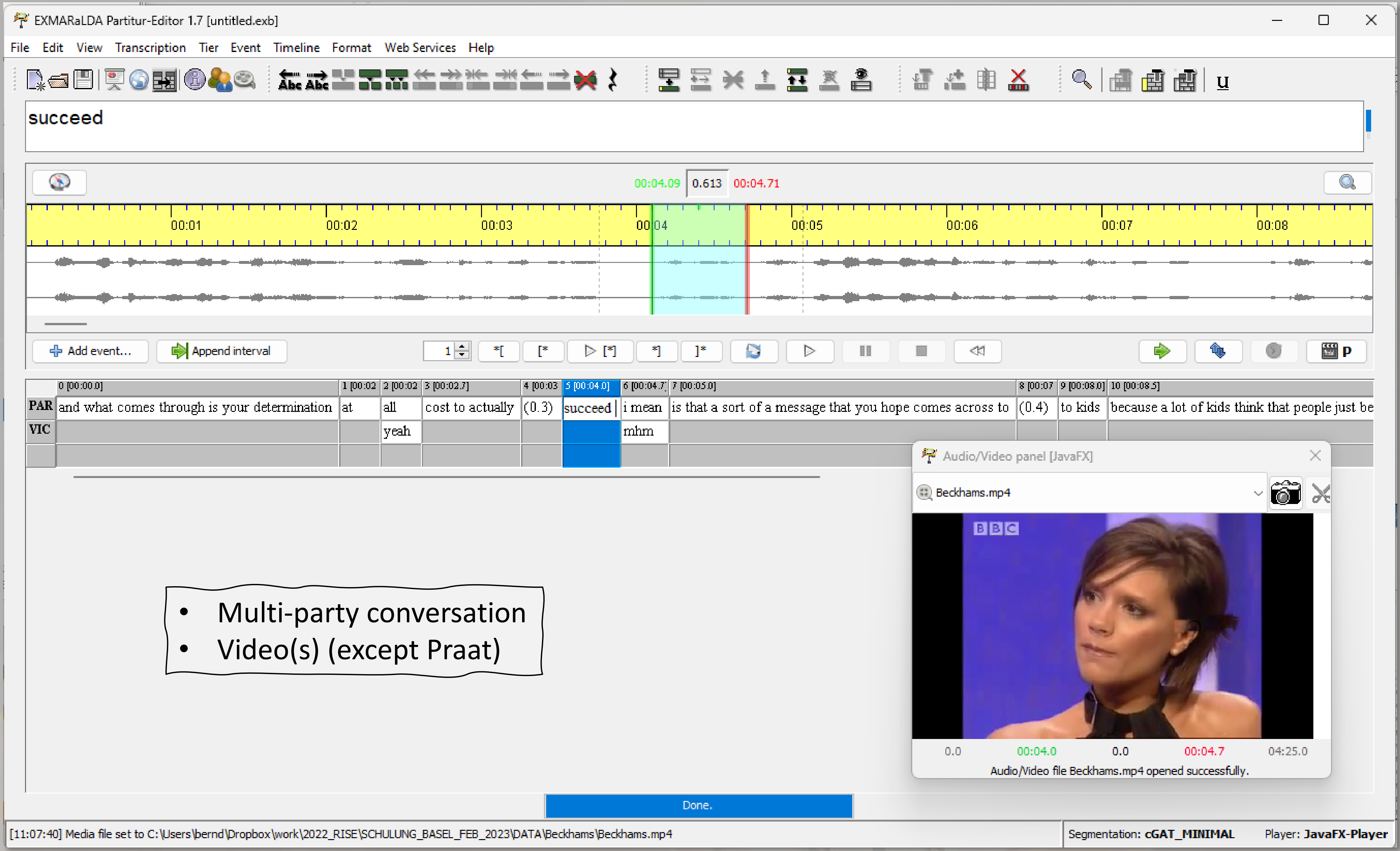

4. TextGrid Example_Praat
File Edit Time Play TextGrid Interval Boundary Tier Sound Analyses Spectrogram Pitch Intensity Formants Pulses
Help
2.742831
0.8913
0.01043
-0.7172
~ non-modifiable copy of sound
5000 Hz
0 Hz
derived spectrogram
derived pitch
500 Hz
75 Hz
modifiable TextGrid
1
SPK0|t|v
(5/10)
2.742831
2.742831
0
5.485662
Visible part 5.485662 seconds
Total duration 5.485662 seconds
all
in
out
sel
bak
Group

[0.4]

[0.8]	Here we go.	Here’s my voice message.	[0.4]	So.	[0.1]	Why do you need that?	[1.4]

Praat

https://www.fon.hum.uva.nl/praat/



succeed

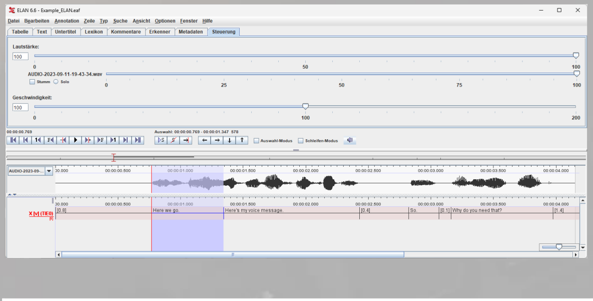
	0 [00:00.0]	1 [00:02	2 [00:02	3 [00:02.7]	4 [00:03	5 [00:04.0]	6 [00:04.7]	7 [00:05.0]	8 [00:07	9 [00:08.0]	10 [00:08.5]
PAR	and what comes through is your determination	at	all	cost to actually	(0.3)	succeed	i mean	is that a sort of a message that you hope comes across to	(0.4)	to kids	because a lot of kids think that people just be
VIC			yeah				mhm				

Audio/Video file Beckhams.mp4 opened successfully.

[11:07:40] Media file set to C:\Users\bernd\Dropbox\work\2022_RISE\SCHULUNG_BASEL_FEB_2023\DATA\Beckhams\Beckhams.mp4

Segmentation: cGAT_MINIMAL Player: JavaFX-Player

- Multi-party conversation
- Video(s) (except Praat)



ELAN 6.6 - Example_ELAN.eaf

Lautstärke: 100

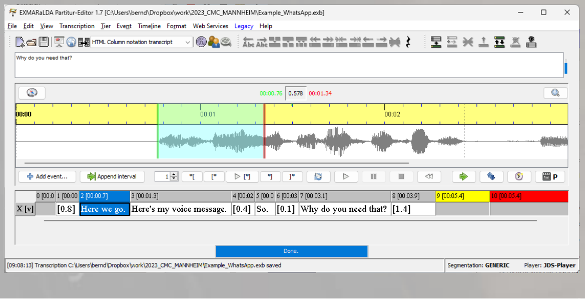
AUDIO-2023-09-11-19-43-34.wav

Geschwindigkeit: 100

00:00:00.769

Auswahl: 00:00:00.769 - 00:00:01.347 578

X [v] (TIE0)	[0.8]	Here we go.	Here's my voice message.	[0.4]	So.	[0.1]	Why do you need that?	[1.4]



EXMARaLDA Partitur-Editor 1.7 [C:\Users\bernd\Dropbox\work\2023_CMC_MANNHEIM\Example_WhatsApp.exb]

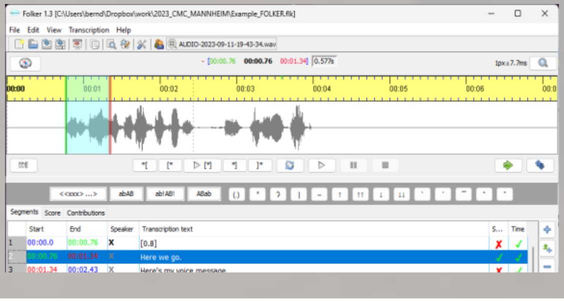
Why do you need that?

00:00.76 0.578 00:01.34

	0 [00:0	1 [00:00	2 [00:00.7]	3 [00:01.3]	4 [00:02	5 [00:0	6 [00:03	7 [00:03.1]	8 [00:03.9]	9 [00:05.4]	10 [00:05.4]
X [v]		[0.8]	Here we go.	Here's my voice message.	[0.4]	So.	[0.1]	Why do you need that?	[1.4]		

[09:08:13] Transcription C:\Users\bernd\Dropbox\work\2023_CMC_MANNHEIM\Example_WhatsApp.exb saved

Segmentation: GENERIC Player: JDS-Player

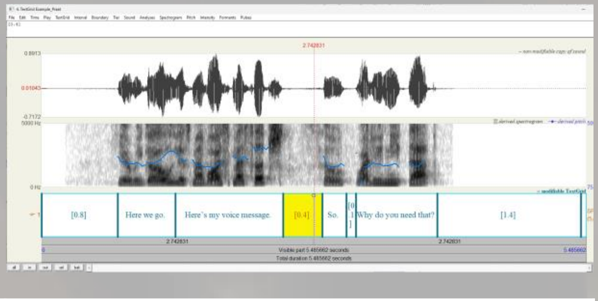


Folker 1.3 [C:\Users\bernd\Dropbox\work\2023_CMC_MANNHEIM\Example_FOLKER.flk]

AUDIO-2023-09-11-19-43-34.wav

[00:00.76 00:00.76 00:01.34] 0.577s 1px≙7.7ms

	Start	End	Speaker	Transcription text	S...	Time
1	00:00.0	00:00.76	X	[0.8]	✗	✓
2	00:00.76	00:01.34	X	Here we go.	✓	✓
3	00:01.34	00:02.43	X	Here's my voice message.	✗	✓



[0.8]	Here we go.	Here's my voice message.	[0.4]	So.	[0.1]	Why do you need that?	[1.4]

Visible part 5.485662 seconds

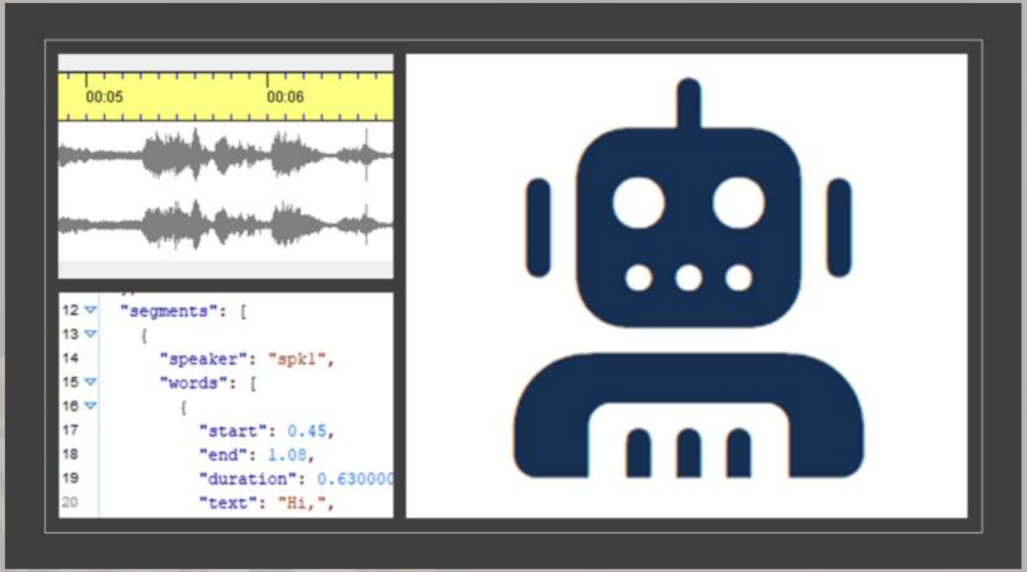
Total duration 5.485662 seconds

ELAN	EXMARaLDA Partitur-Editor	FOLKER	Praat
Advanced video functionality	Part of a larger system: COMA (Corpus Manager), OrthoNormal (token annotation), EXAKT (Query)		Advanced functionality for phonetic analysis
Complex annotation hierarchies	Direct support for ISO/TEI Direct support for transcription systems (cGAT, HIAT)		Scriptable
	Built-in support for tokenisation Built-in support for masking / pseudonymization		
Complex	**Not too simple ;-)**	**One tier per speaker**	**No video support**

- Tools are interoperable (more later)
- Things to consider:
  - ✓ complexity of your data
  - ✓ workflow integration (other tasks)
  - ✓ expertise in your team / network
  - ✓ collaboration
  - ✓ recommendations by data centres

Automatic Speech Recognition

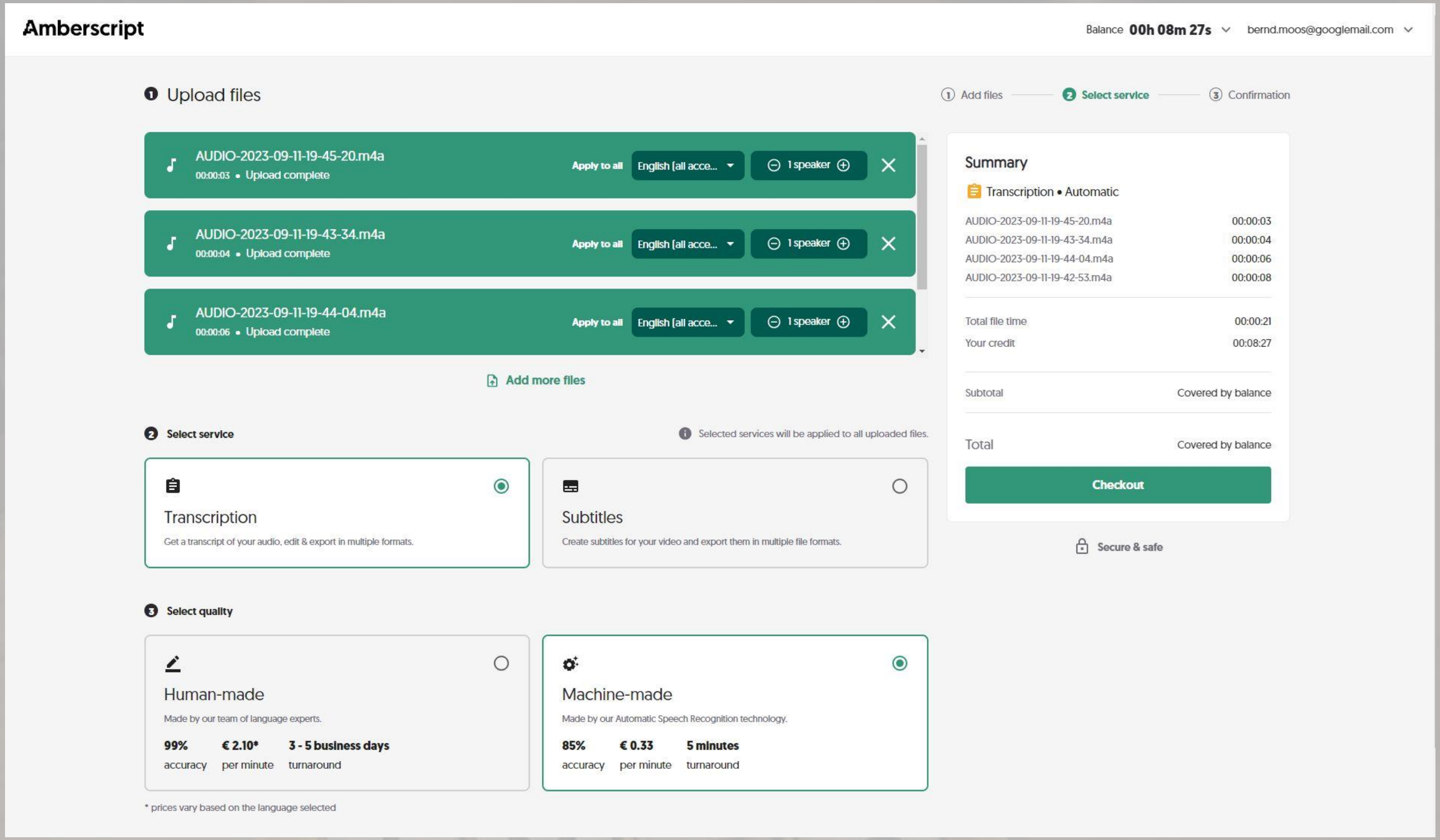
- „Speech to text“ – machine transcription
- Dramatic improvements in the last few years
- Potential to save lots (and lots) of effort (manual transcription 1:10 up to 1:100)
- Commercial companies: Amberscript, Trint, Google, SpeechMatics, …
- Big questions:
    - Quality and precision?
    - Data protection?



```
  "segments": [
    {
      "speaker": "spkl",
      "words": [
        {
          "start": 0.45,
          "end": 1.08,
          "duration": 0.630000
          "text": "Hi,",
```

ASR Example: Amberscript

- Upload to platform
- Pay (EUR 10 to 20 per hour)
- Submit for ASR
- Edit result online
- Download result



Amberscript

Balance **00h 08m 27s** bernd.moos@googlemail.com

❶ Upload files

① Add files — ❷ **Select service** — ③ Confirmation

AUDIO-2023-09-11-19-45-20.m4a
00:00:03 • Upload complete
Apply to all English (all acce... ⊖ 1 speaker ⊕ ✕

AUDIO-2023-09-11-19-43-34.m4a
00:00:04 • Upload complete
Apply to all English (all acce... ⊖ 1 speaker ⊕ ✕

AUDIO-2023-09-11-19-44-04.m4a
00:00:06 • Upload complete
Apply to all English (all acce... ⊖ 1 speaker ⊕ ✕

Add more files

❷ **Select service**

Selected services will be applied to all uploaded files.

● Transcription
Get a transcript of your audio, edit & export in multiple formats.

○ Subtitles
Create subtitles for your video and export them in multiple file formats.

❸ **Select quality**

○ Human-made
Made by our team of language experts.

99%	**€ 2.10***	**3 - 5 business days**
accuracy	per minute	turnaround

● Machine-made
Made by our Automatic Speech Recognition technology.

85%	**€ 0.33**	**5 minutes**
accuracy	per minute	turnaround

* prices vary based on the language selected

Summary

Transcription • Automatic

AUDIO-2023-09-11-19-45-20.m4a	00:00:03
AUDIO-2023-09-11-19-43-34.m4a	00:00:04
AUDIO-2023-09-11-19-44-04.m4a	00:00:06
AUDIO-2023-09-11-19-42-53.m4a	00:00:08
Total file time	00:00:21
Your credit	00:08:27
Subtotal	Covered by balance
Total	Covered by balance

Checkout

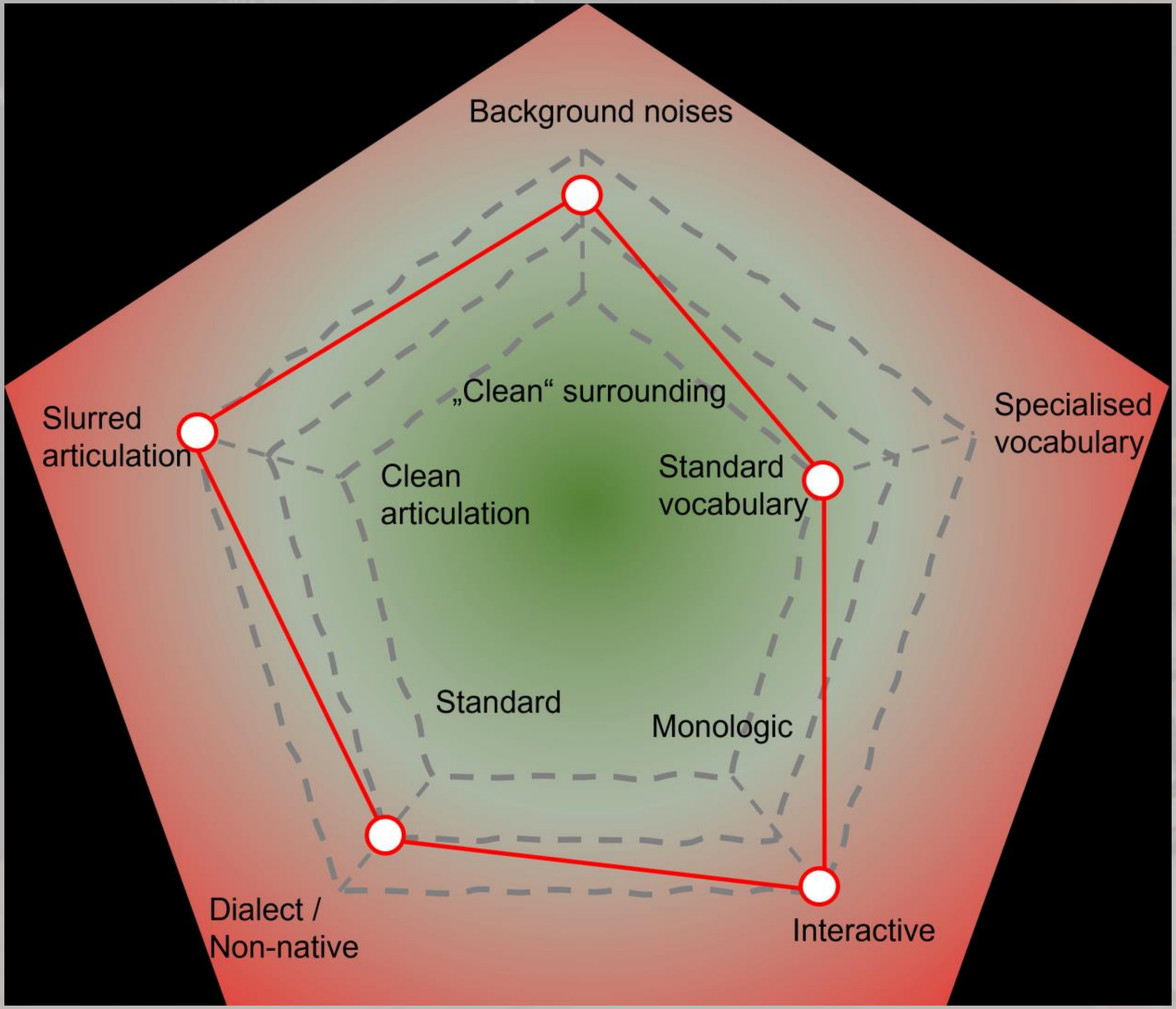
Secure & safe

ASR: Quality and precision

Manual transcription	Amberscript
[0.8] Here we go. Here's my voice message. **[0.4]** So. **[0.1]** Why do you need that? **[1.4]**	Here we go. Here's my voice message. So why do you need that?
The thing is **((clears throat))**, I would need an **ehm ehm** what's it called voice message from you. Can you do that?	The thing is, I would need an what's it called voice message from you. Can you do that?
Oh, **m/ ((clears throat))** maybe I should have put that in a voice message. It's for the **course**. I said so.	Oh, maybe I should have put that in a voice message. It's for the **cause**. I said so.

- Word Error Rate (WER): <5% very good // <10% good // <20% acceptable // >20% ???
- Smoothing of „performance phenomena“ – disfluencies, non-verbal
- Automatic Recognition → Manual correction : difficult with WER > 20%, difficult when high precision is required
- What WER to expect?

ASR: Quality



- Ideal: Professional podcast – expect WER < 5%
- Voice message:

  + monologic

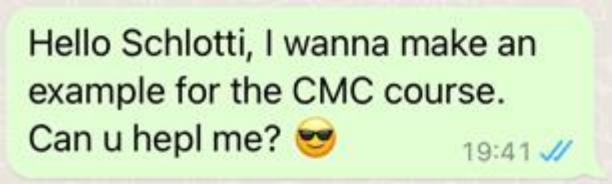
  +/- clean, standard, vocabulary
- Zoom conference

  + for clean surrounding

  +/- for all others

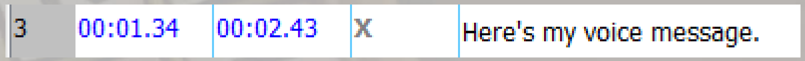
Tasks

- Transcription
- Normalisation
- Masking (de-identification, ~~anonymization~~) in audio
- Masking (pseudonyms) in transcript text
- Lemmatisation
- POS tagging
- …



Hello Schlotti, I wanna make an example for the CMC course. Can u hepl me? 😎 19:41

Hello **Linda**
I **want to** make an example …
Can **you help** me?

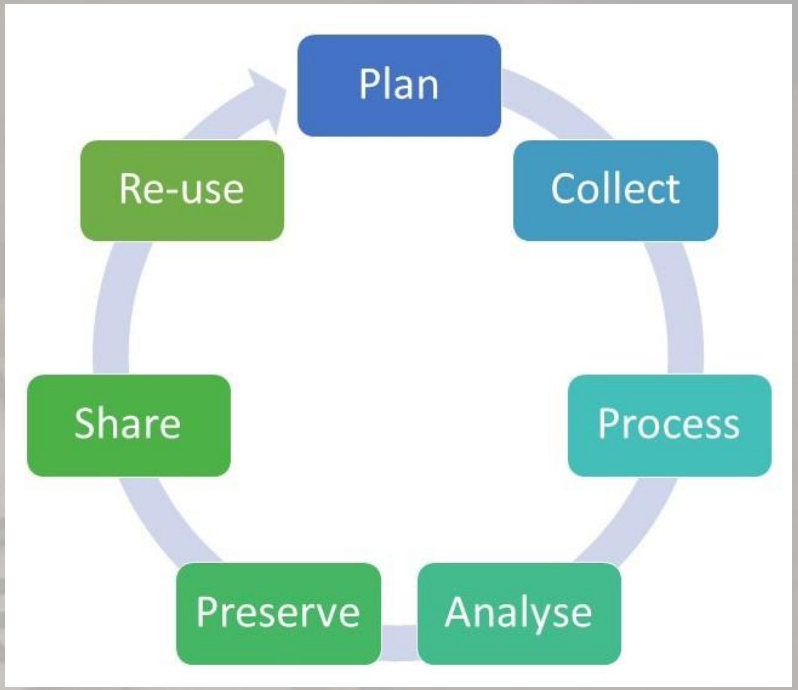


3 00:01.34 00:02.43 X Here's my voice message.

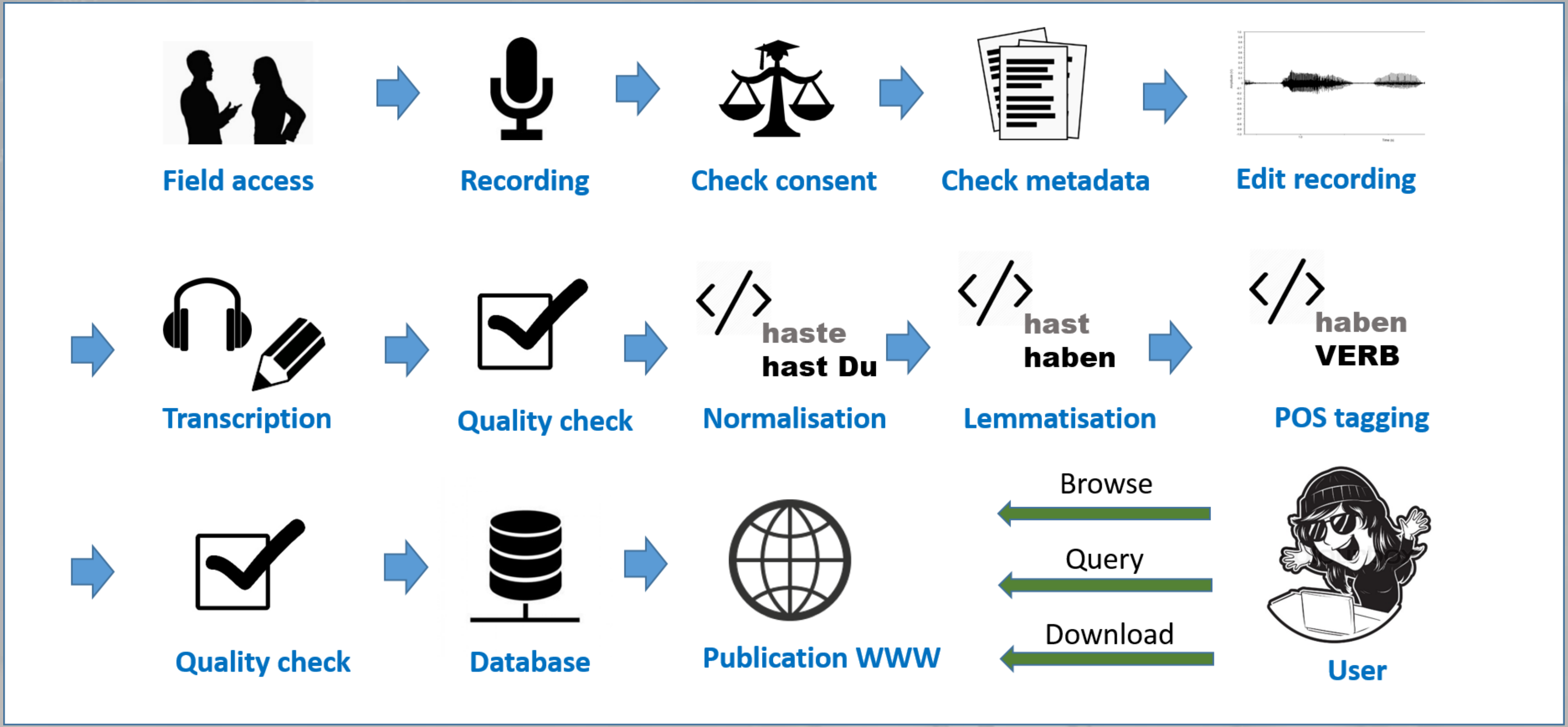
Here is my **speech** message.

Tasks and tools in workflows

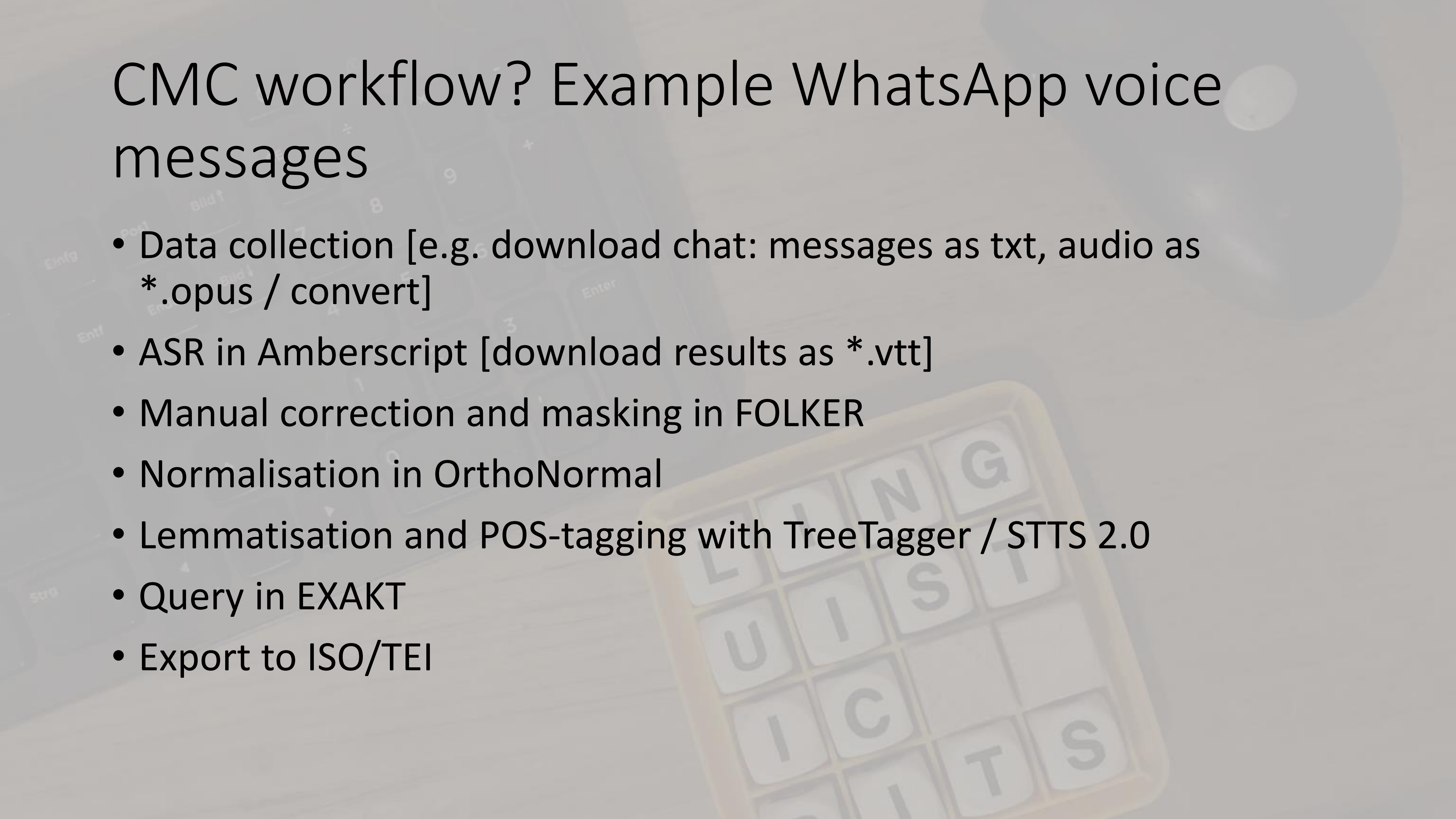
- Methodological coherence: Data models, Annotation schemes
- Interoperability of tools: Getting data from A to B
- With a view to your own analytical demands
- With a view to re-usability
- From data acquistion to data preservation (and back: research data lifecycle)



FOLK workflow



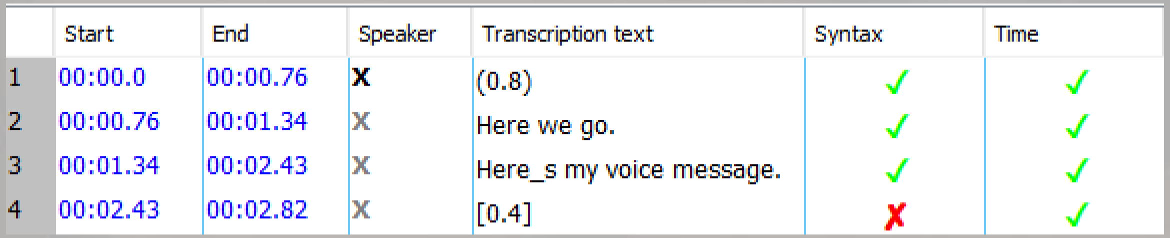
CMC workflow? Example WhatsApp voice messages



- Data collection [e.g. download chat: messages as txt, audio as *.opus / convert]
- ASR in Amberscript [download results as *.vtt]
- Manual correction and masking in FOLKER
- Normalisation in OrthoNormal
- Lemmatisation and POS-tagging with TreeTagger / STTS 2.0
- Query in EXAKT
- Export to ISO/TEI

CMC workflow? WhatsApp voice messages

- Data collection
- ASR in Amberscript
- Manual **correction** and masking in FOLKER
- Normalisation in OrthoNormal
- Lemmatisation and POS-tagging with TreeTagger / STTS 2.0
- Query in EXAKT
- Export to ISO/TEI

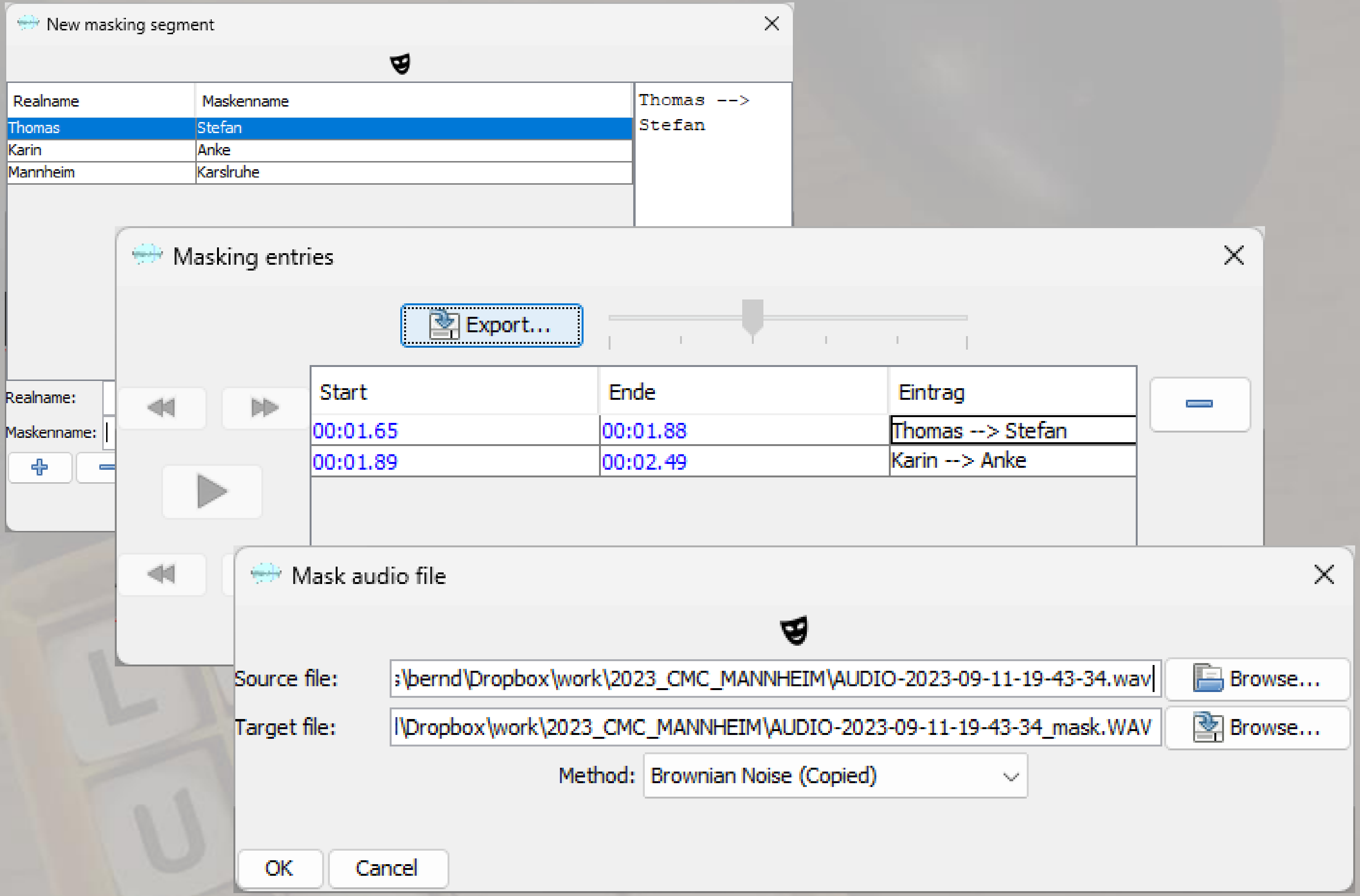


	Start	End	Speaker	Transcription text	Syntax	Time
1	00:00.0	00:00.76	X	(0.8)	✓	✓
2	00:00.76	00:01.34	X	Here we go.	✓	✓
3	00:01.34	00:02.43	X	Here_s my voice message.	✓	✓
4	00:02.43	00:02.82	X	[0.4]	✗	✓

- cGAT syntax control
- automatic tokenisation

CMC workflow? WhatsApp voice messages

- Data collection
- ASR in Amberscript
- Manual correction and **masking** in FOLKER
- Normalisation in OrthoNormal
- Lemmatisation and POS-tagging with TreeTagger / STTS 2.0
- Query in EXAKT
- Export to ISO/TEI



Realname	Maskenname
Thomas	Stefan
Karin	Anke
Mannheim	Karslruhe

Thomas --> Stefan

Start	Ende	Eintrag
00:01.65	00:01.88	Thomas --> Stefan
00:01.89	00:02.49	Karin --> Anke

Source file: :\bernd\Dropbox\work\2023_CMC_MANNHEIM\AUDIO-2023-09-11-19-43-34.wav

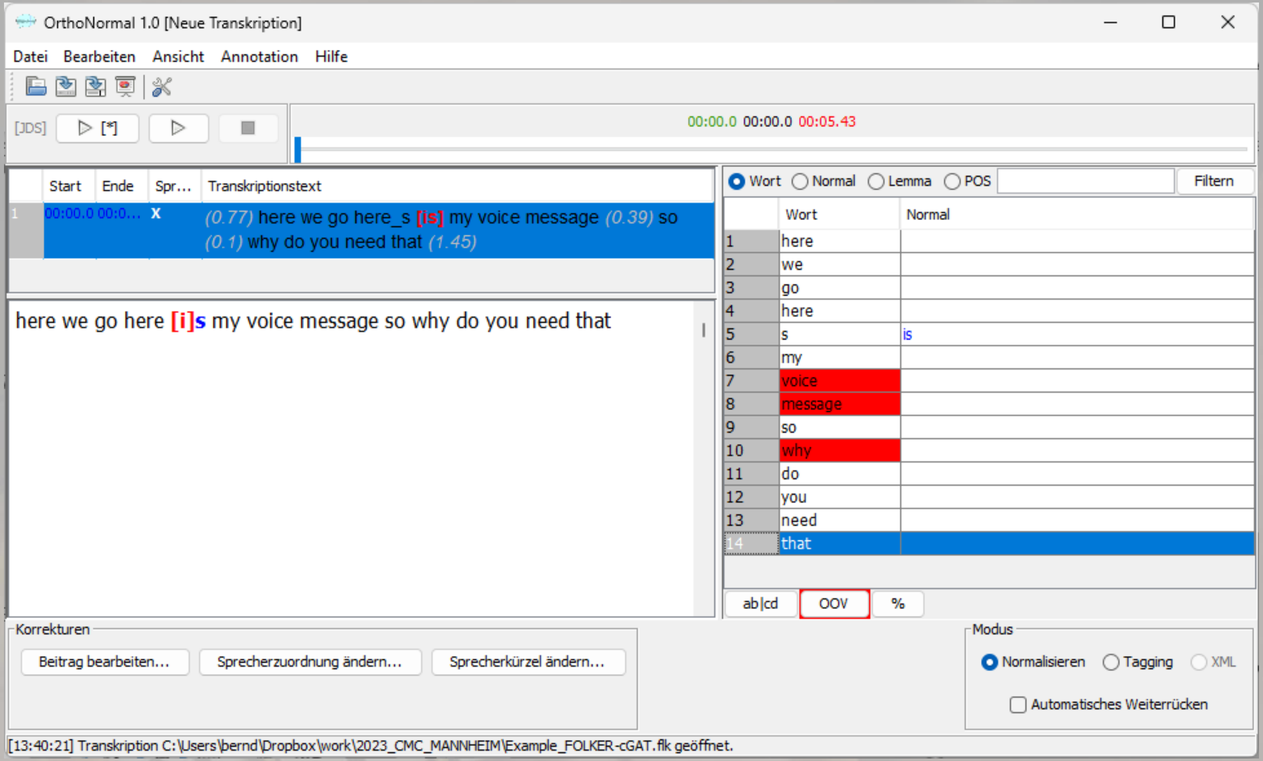
Target file: l\Dropbox\work\2023_CMC_MANNHEIM\AUDIO-2023-09-11-19-43-34_mask.WAV

Method: Brownian Noise (Copied)

- Management of pseudonyms
- Selection of audio stretches to be masked
- Automatic insertion of noises into the audio

CMC workflow? WhatsApp voice messages

- Data collection
- ASR in Amberscript
- Manual correction and masking in FOLKER
- **Normalisation** in OrthoNormal
- Lemmatisation and POS-tagging with TreeTagger / STTS 2.0
- Query in EXAKT
- Export to ISO/TEI



00:00.0 00:00.0 00:05.43

	Start	Ende	Spr...	Transkriptionstext
1	00:00.0	00:0...	X	(0.77) here we go here_s [is] my voice message (0.39) so (0.1) why do you need that (1.45)

here we go here [i]s my voice message so why do you need that

● Wort ○ Normal ○ Lemma ○ POS

	Wort	Normal
1	here	
2	we	
3	go	
4	here	
5	s	is
6	my	
7	voice	
8	message	
9	so	
10	why	
11	do	
12	you	
13	need	
14	that	

Modus: ● Normalisieren ○ Tagging ○ XML

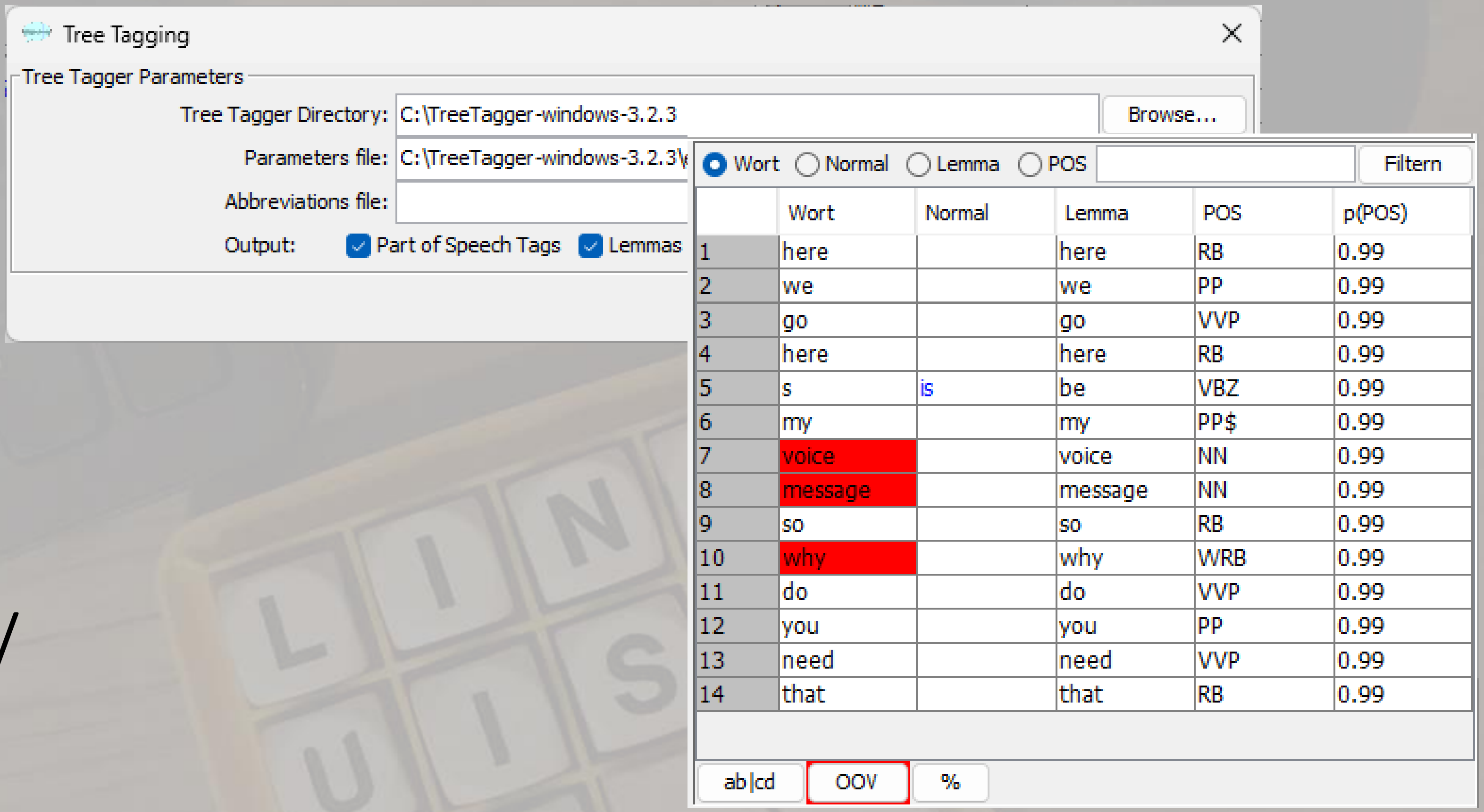
□ Automatisches Weiterrücken

[13:40:21] Transkription C:\Users\bernd\Dropbox\work\2023_CMC_MANNHEIM\Example_FOLKER-cGAT.flk geöffnet.

- Based on tokenisation
- Automatic for German (FOLK lexicon)

CMC workflow? WhatsApp voice messages

- Data collection
- ASR in Amberscript
- Manual correction and masking in FOLKER
- Normalisation in OrthoNormal
- **Lemmatisation and POS-tagging with TreeTagger** / STTS 2.0
- Query in EXAKT
- Export to ISO/TEI



Tree Tagger Directory: C:\TreeTagger-windows-3.2.3

Parameters file: C:\TreeTagger-windows-3.2.3\

Abbreviations file:

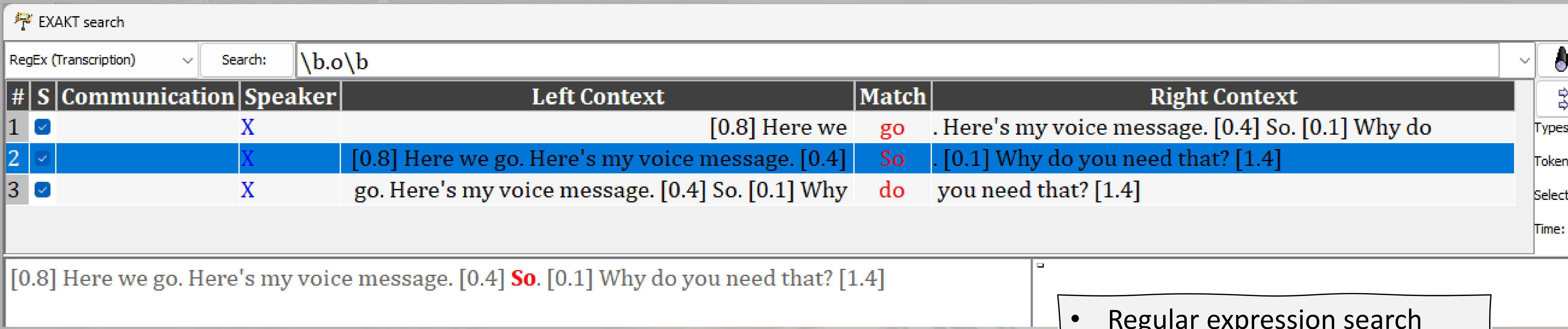
Output: ☑ Part of Speech Tags ☑ Lemmas

● Wort ○ Normal ○ Lemma ○ POS

	Wort	Normal	Lemma	POS	p(POS)
1	here		here	RB	0.99
2	we		we	PP	0.99
3	go		go	VVP	0.99
4	here		here	RB	0.99
5	s	is	be	VBZ	0.99
6	my		my	PP$	0.99
7	voice		voice	NN	0.99
8	message		message	NN	0.99
9	so		so	RB	0.99
10	why		why	WRB	0.99
11	do		do	VVP	0.99
12	you		you	PP	0.99
13	need		need	VVP	0.99
14	that		that	RB	0.99

- Based on normalisation
- Can also be done in OrthoNormal
- OrthoNormal for manual correction

CMC workflow? WhatsApp voice messages



EXAKT search

RegEx (Transcription)

Search: \b.o\b

#	S	Communication	Speaker	Left Context	Match	Right Context
1	☑		X	[0.8] Here we	go	. Here's my voice message. [0.4] So. [0.1] Why do
2	☑		X	[0.8] Here we go. Here's my voice message. [0.4]	So	. [0.1] Why do you need that? [1.4]
3	☑		X	go. Here's my voice message. [0.4] So. [0.1] Why	do	you need that? [1.4]

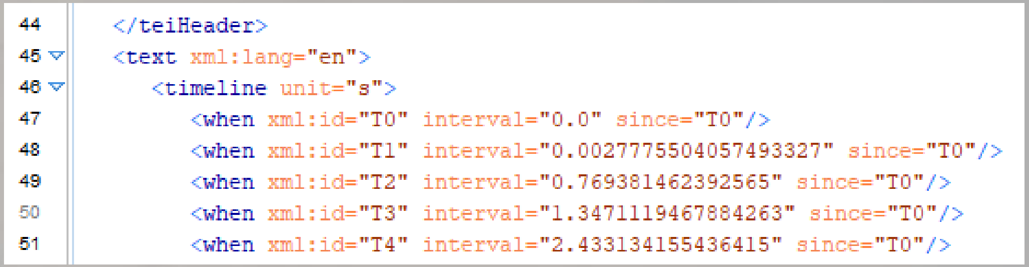
[0.8] Here we go. Here's my voice message. [0.4] **So**. [0.1] Why do you need that? [1.4]

- Regular expression search
- Also on annotations (POS etc.)

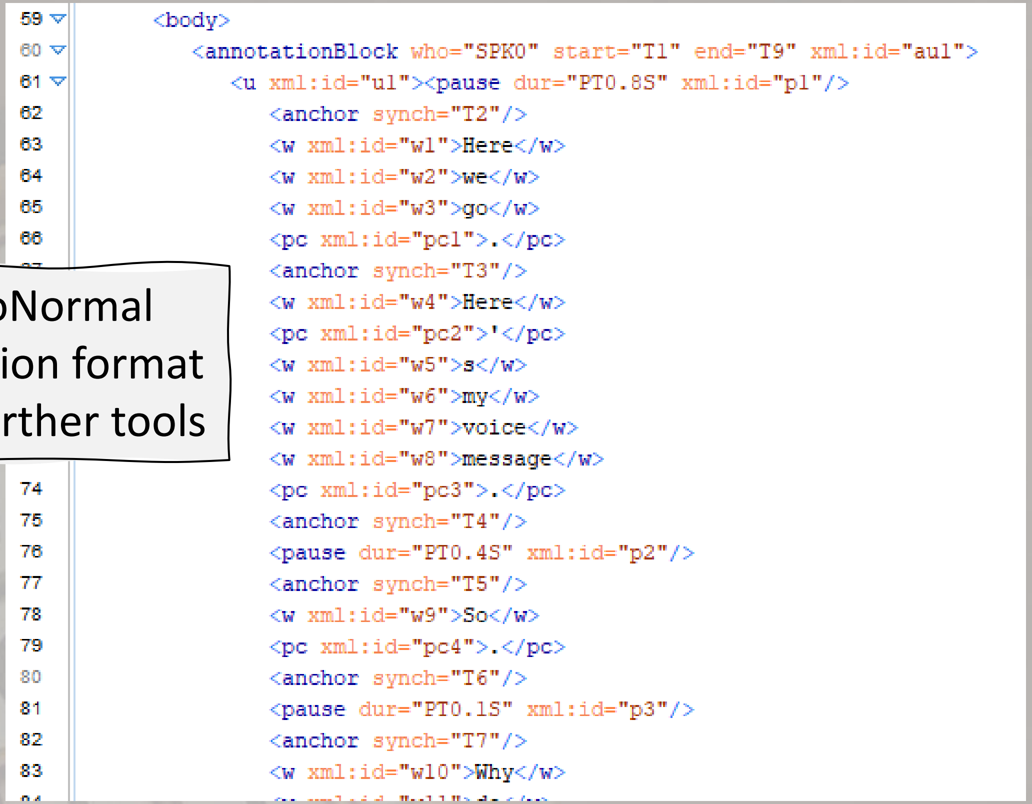
- Data collection
- ASR in Amberscript
- Manual correction and masking in FOLKER
- Normalisation in OrthoNormal
- Lemmatisation and POS-tagging with TreeTagger / STTS 2.0
- **Query** in EXAKT
- Export to ISO/TEI

CMC workflow? WhatsApp voice messages

- Data collection
- ASR in Amberscript
- Manual correction and masking in FOLKER
- Normalisation in OrthoNormal
- Lemmatisation and POS-tagging with TreeTagger / STTS 2.0
- Query in EXAKT
- Export to **ISO/TEI**

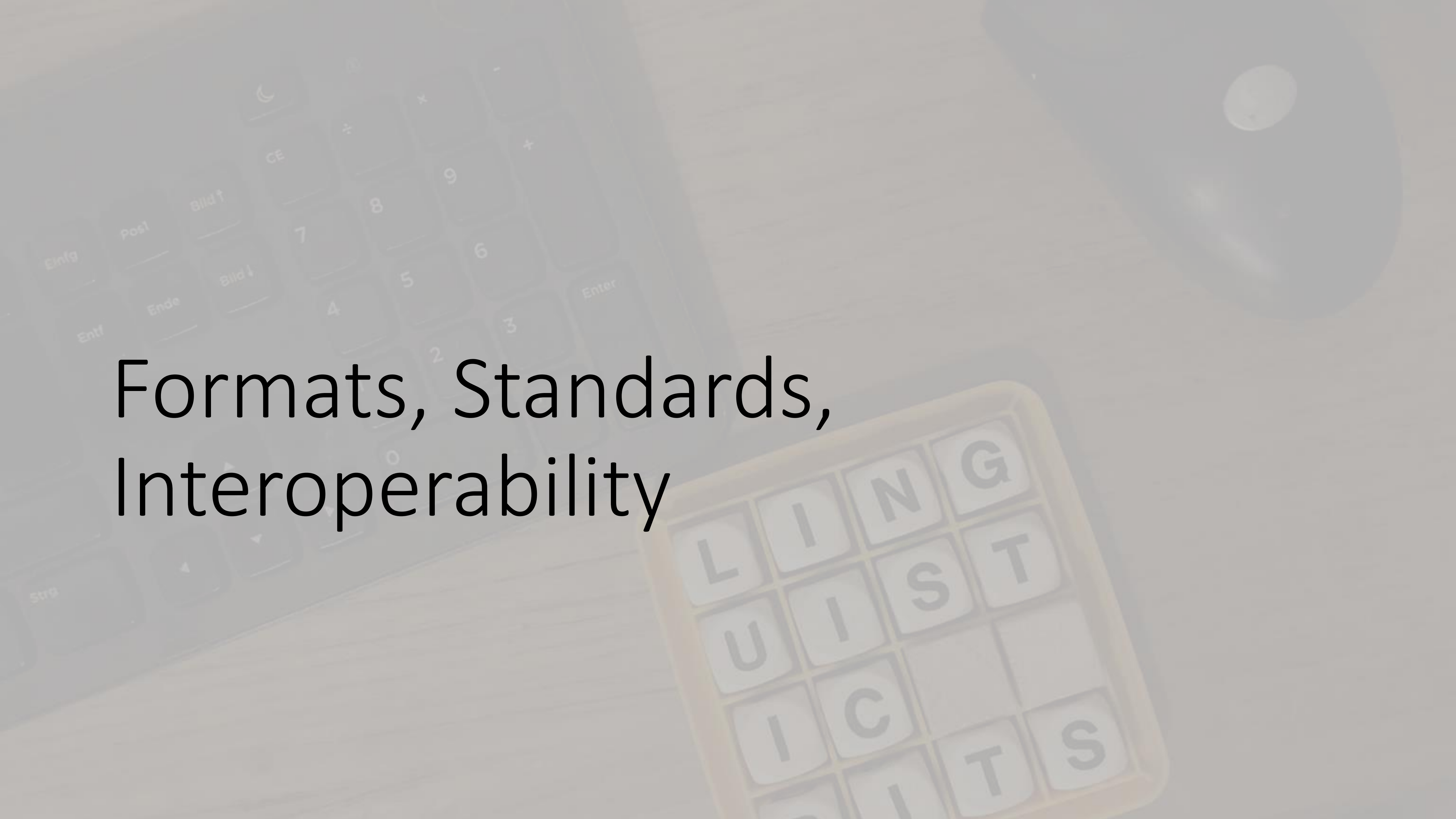


```
</teiHeader>
<text xml:lang="en">
  <timeline unit="s">
    <when xml:id="T0" interval="0.0" since="T0"/>
    <when xml:id="T1" interval="0.0027775504057493327" since="T0"/>
    <when xml:id="T2" interval="0.769381462392565" since="T0"/>
    <when xml:id="T3" interval="1.3471119467884263" since="T0"/>
    <when xml:id="T4" interval="2.433134155436415" since="T0"/>
```

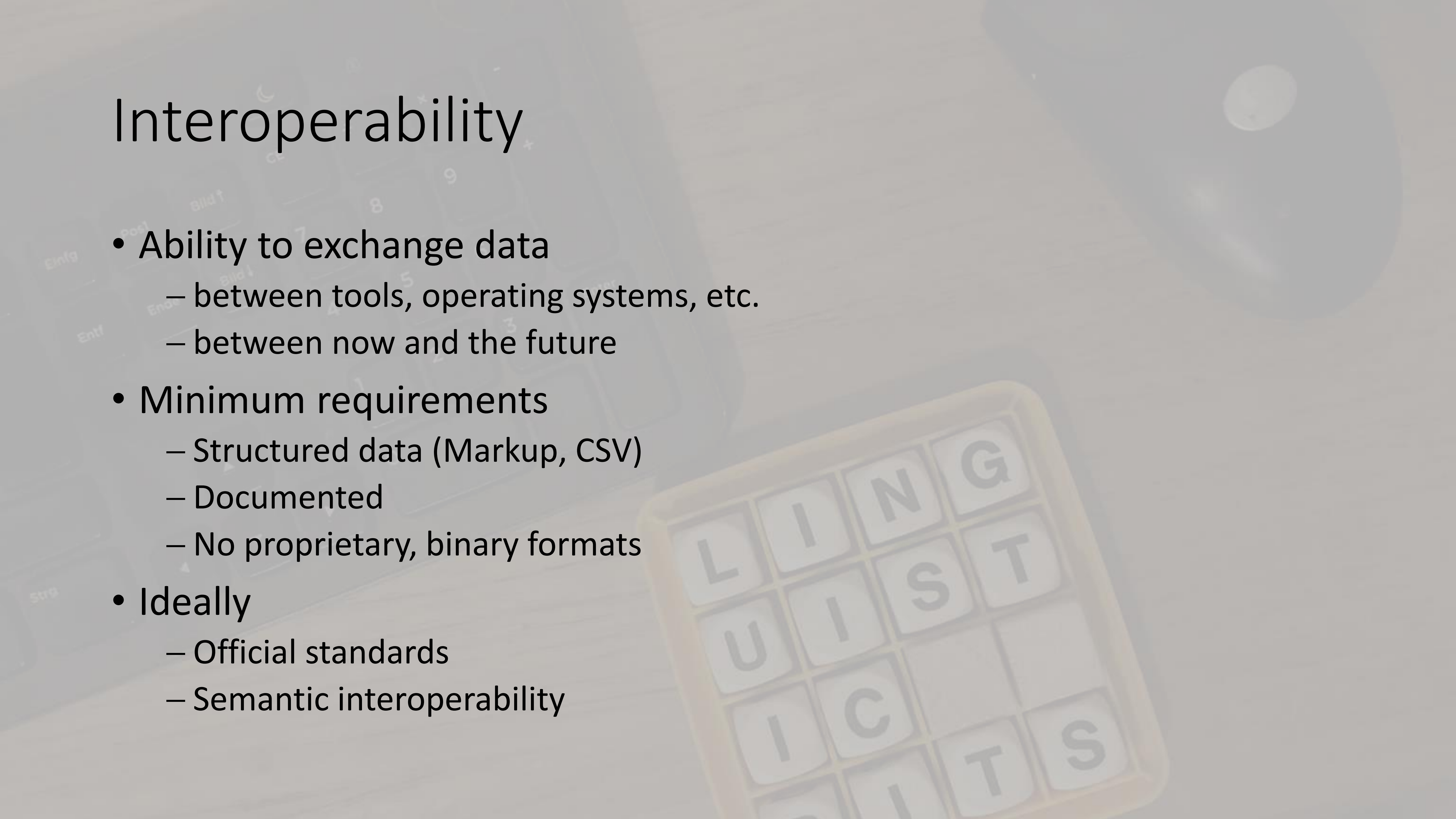


```
<body>
  <annotationBlock who="SPK0" start="T1" end="T9" xml:id="au1">
    <u xml:id="u1"><pause dur="PT0.8S" xml:id="p1"/>
      <anchor synch="T2"/>
      <w xml:id="w1">Here</w>
      <w xml:id="w2">we</w>
      <w xml:id="w3">go</w>
      <pc xml:id="pc1">.</pc>
      <anchor synch="T3"/>
      <w xml:id="w4">Here</w>
      <pc xml:id="pc2">'</pc>
      <w xml:id="w5">s</w>
      <w xml:id="w6">my</w>
      <w xml:id="w7">voice</w>
      <w xml:id="w8">message</w>
      <pc xml:id="pc3">.</pc>
      <anchor synch="T4"/>
      <pause dur="PT0.4S" xml:id="p2"/>
      <anchor synch="T5"/>
      <w xml:id="w9">So</w>
      <pc xml:id="pc4">.</pc>
      <anchor synch="T6"/>
      <pause dur="PT0.1S" xml:id="p3"/>
      <anchor synch="T7"/>
      <w xml:id="w10">Why</w>
```

- Directly from OrthoNormal
- Standard preservation format
- Compatible with further tools

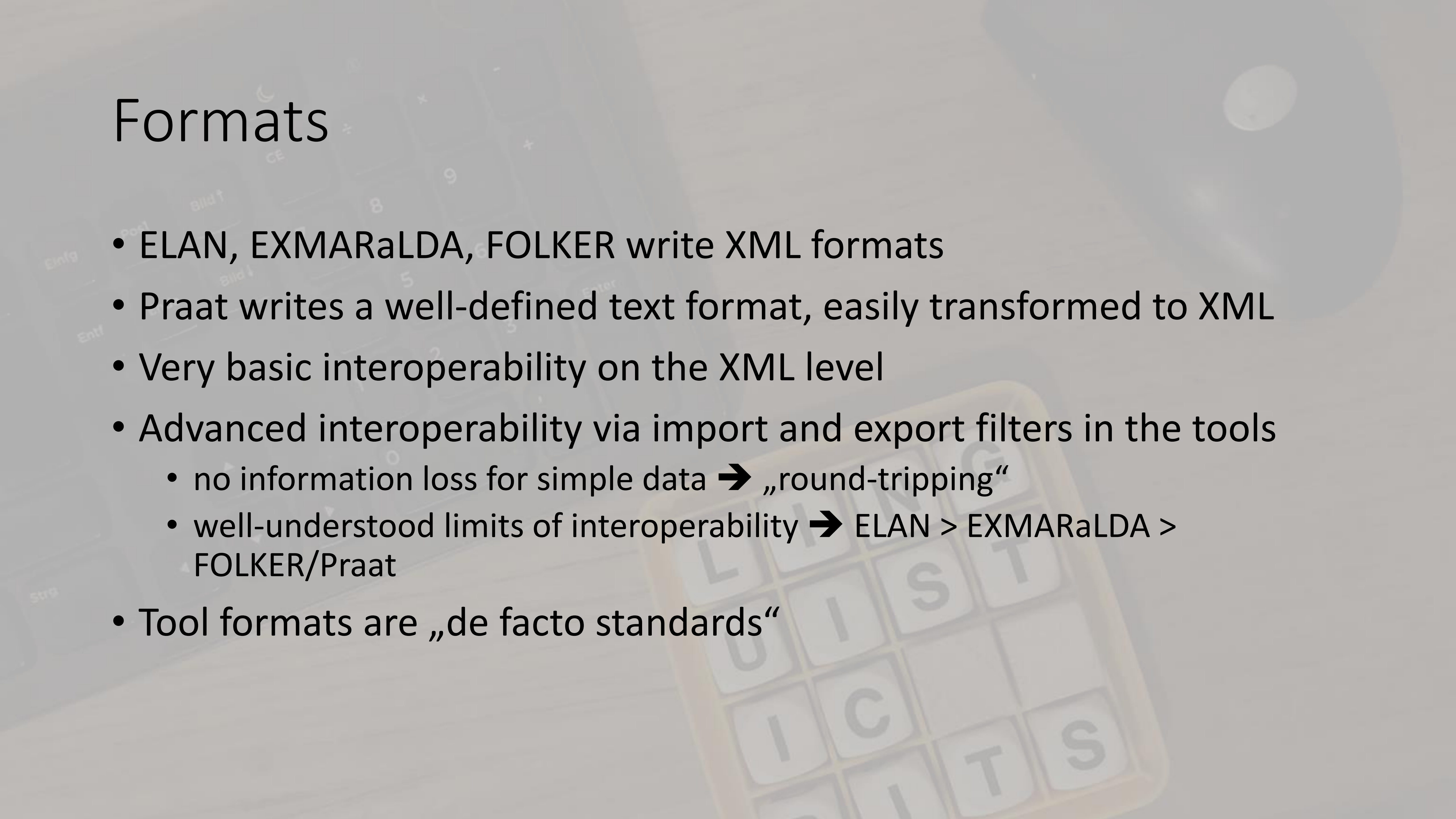


Formats, Standards, Interoperability



Interoperability

- Ability to exchange data
  - between tools, operating systems, etc.
  - between now and the future
- Minimum requirements
  - Structured data (Markup, CSV)
  - Documented
  - No proprietary, binary formats
- Ideally
  - Official standards
  - Semantic interoperability



Formats

- ELAN, EXMARaLDA, FOLKER write XML formats
- Praat writes a well-defined text format, easily transformed to XML
- Very basic interoperability on the XML level
- Advanced interoperability via import and export filters in the tools
  - no information loss for simple data ➔ „round-tripping“
  - well-understood limits of interoperability ➔ ELAN > EXMARaLDA > FOLKER/Praat
- Tool formats are „de facto standards“

Standards

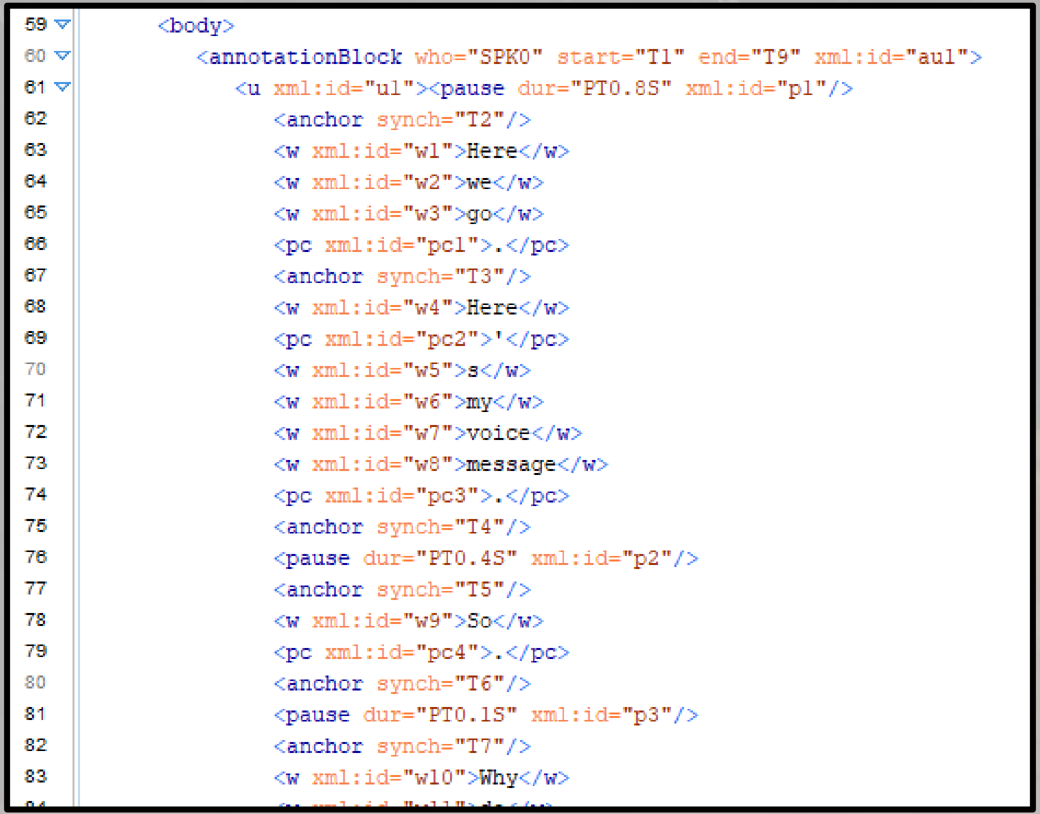
- ISO 24624:2016 “Language resource management — Transcription of spoken language”
  - published by ISO in 2016, reviewed and confirmed in 2022
  - “endorsed” by the Text Encoding Initiative (TEI)
    - based on the TEI guidelines
    - TEI guidelines adapted to concepts needed for the standard
  - cross-relations to other parts of the guidelines
    - written text corpora
    - **CMC corpora!**
  - compatible with and supported directly or indirectly (via interoperability) by more than one tool
  - recommended / required by some CLARIN data centres

Michael Beißwenger, Harald Lüngen (2020): **CMC-core: a schema for the representation of CMC corpora in TEI.** *Corpus.*

Hedeland, Hanna / Schmidt, Thomas (2022): **The TEI-based ISO Standard ‘Transcription of spoken language’ as an Exchange Format within CLARIN and beyond.** *Selected Papers from the CLARIN Annual Conference 2021.*

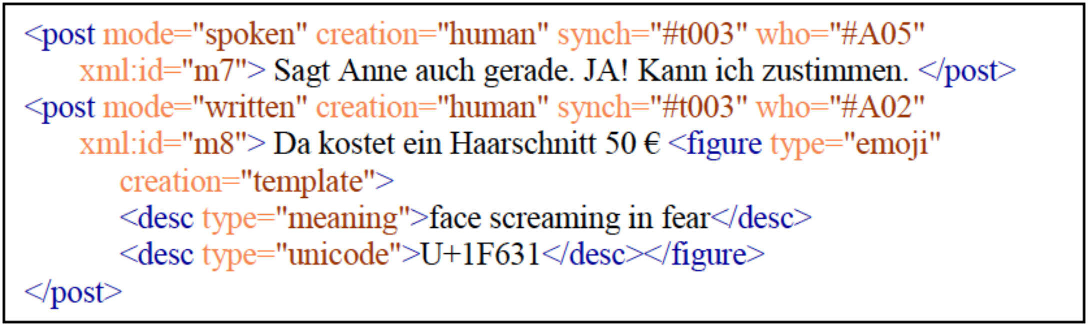
ISO/TEI Spoken vs. CMC TEI

Transcript in ISO/TEI Spoken



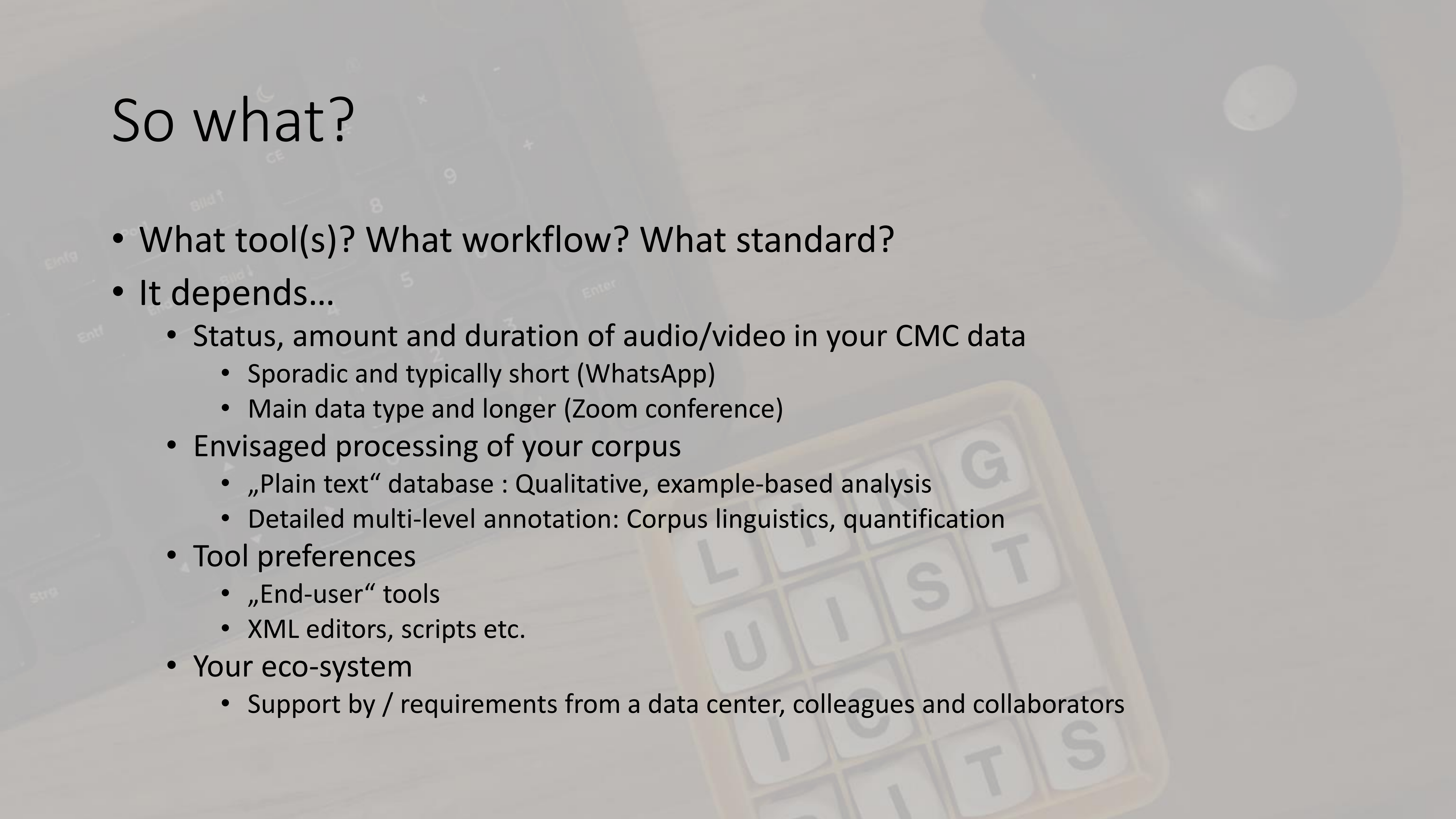
```
<body>
   <annotationBlock who="SPK0" start="T1" end="T9" xml:id="au1">
      <u xml:id="u1"><pause dur="PT0.8S" xml:id="p1"/>
         <anchor synch="T2"/>
         <w xml:id="w1">Here</w>
         <w xml:id="w2">we</w>
         <w xml:id="w3">go</w>
         <pc xml:id="pc1">.</pc>
         <anchor synch="T3"/>
         <w xml:id="w4">Here</w>
         <pc xml:id="pc2">'</pc>
         <w xml:id="w5">s</w>
         <w xml:id="w6">my</w>
         <w xml:id="w7">voice</w>
         <w xml:id="w8">message</w>
         <pc xml:id="pc3">.</pc>
         <anchor synch="T4"/>
         <pause dur="PT0.4S" xml:id="p2"/>
         <anchor synch="T5"/>
         <w xml:id="w9">So</w>
         <pc xml:id="pc4">.</pc>
         <anchor synch="T6"/>
         <pause dur="PT0.1S" xml:id="p3"/>
         <anchor synch="T7"/>
         <w xml:id="w10">Why</w>
```

<post mode=„spoken“> in CMC-TEI



```
<post mode="spoken" creation="human" synch="#t003" who="#A05"
    xml:id="m7"> Sagt Anne auch gerade. JA! Kann ich zustimmen. </post>
<post mode="written" creation="human" synch="#t003" who="#A02"
    xml:id="m8"> Da kostet ein Haarschnitt 50 € <figure type="emoji"
       creation="template">
       <desc type="meaning">face screaming in fear</desc>
       <desc type="unicode">U+1F631</desc></figure>
</post>
```

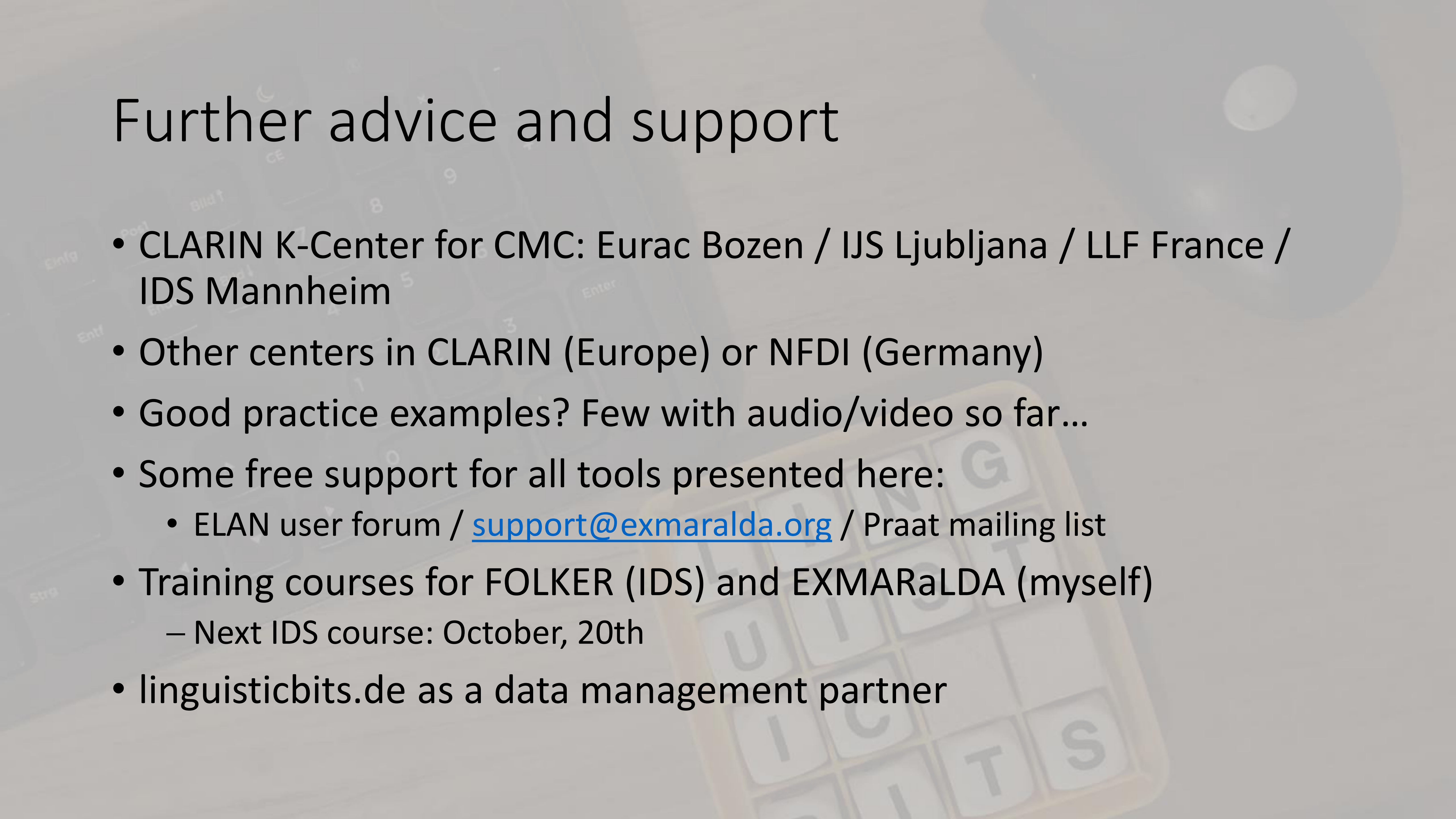
- Internal structure / Level of detail
  - „post“-internal time anchors
  - tokenisation
  - distinction words vs. non-words



So what?

- What tool(s)? What workflow? What standard?
- It depends…
  - Status, amount and duration of audio/video in your CMC data
    - Sporadic and typically short (WhatsApp)
    - Main data type and longer (Zoom conference)
  - Envisaged processing of your corpus
    - „Plain text“ database : Qualitative, example-based analysis
    - Detailed multi-level annotation: Corpus linguistics, quantification
  - Tool preferences
    - „End-user“ tools
    - XML editors, scripts etc.
  - Your eco-system
    - Support by / requirements from a data center, colleagues and collaborators

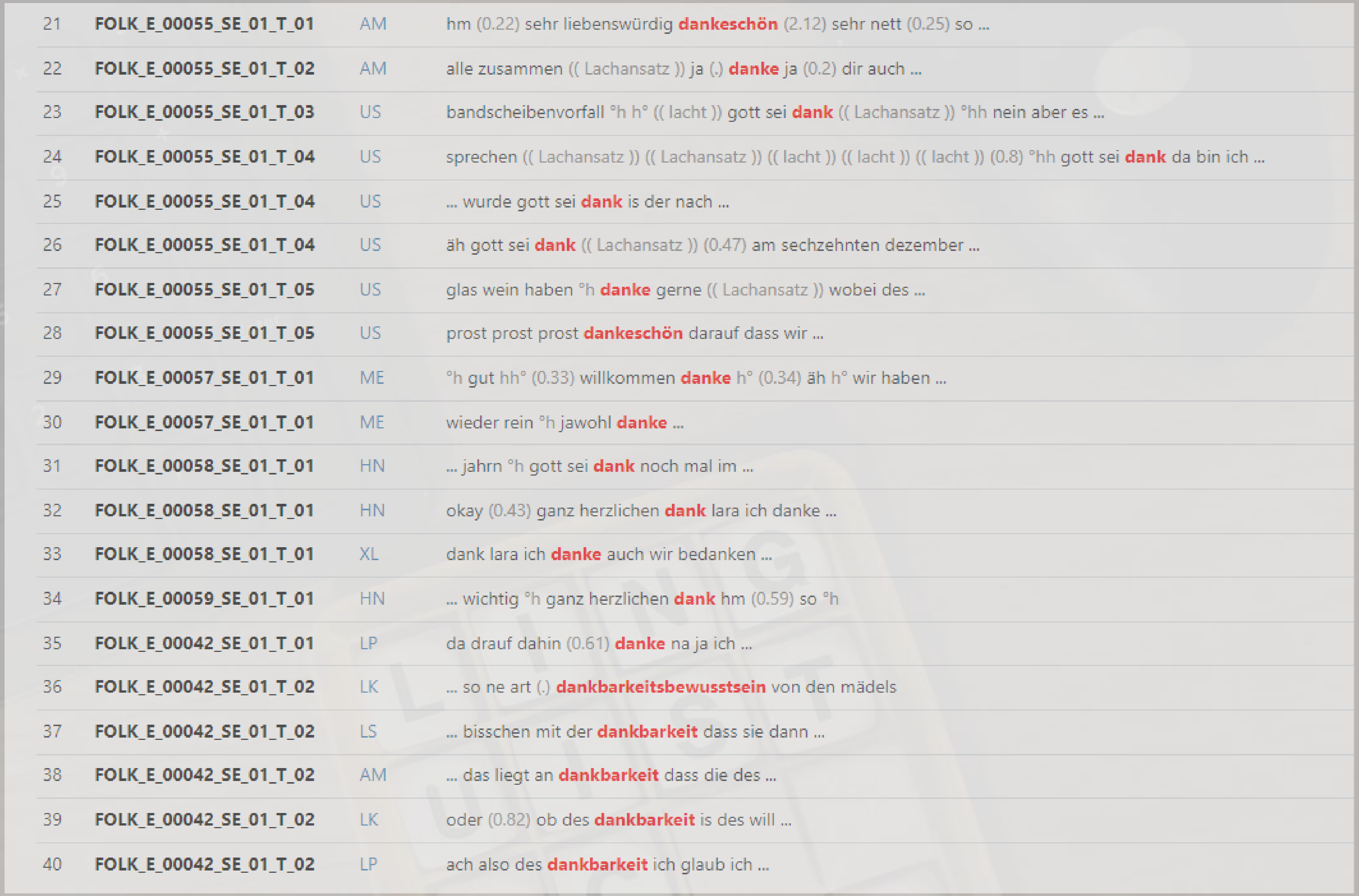
Questions &
attempts at Answers?



Further advice and support

- CLARIN K-Center for CMC: Eurac Bozen / IJS Ljubljana / LLF France / IDS Mannheim
- Other centers in CLARIN (Europe) or NFDI (Germany)
- Good practice examples? Few with audio/video so far…
- Some free support for all tools presented here:
  - ELAN user forum / support@exmaralda.org / Praat mailing list
- Training courses for FOLKER (IDS) and EXMARaLDA (myself)
  - – Next IDS course: October, 20th
- linguisticbits.de as a data management partner

Dr. Thomas Schmidt
https://linguisticbits.de
thomas@linguisticbits.de



21	**FOLK_E_00055_SE_01_T_01**	AM	hm (0.22) sehr liebenswürdig **dankeschön** (2.12) sehr nett (0.25) so ...
22	**FOLK_E_00055_SE_01_T_02**	AM	alle zusammen ((Lachansatz)) ja (.) **danke** ja (0.2) dir auch ...
23	**FOLK_E_00055_SE_01_T_03**	US	bandscheibenvorfall °h h° ((lacht)) gott sei **dank** ((Lachansatz)) °hh nein aber es ...
24	**FOLK_E_00055_SE_01_T_04**	US	sprechen ((Lachansatz)) ((Lachansatz)) ((lacht)) ((lacht)) ((lacht)) (0.8) °hh gott sei **dank** da bin ich ...
25	**FOLK_E_00055_SE_01_T_04**	US	... wurde gott sei **dank** is der nach ...
26	**FOLK_E_00055_SE_01_T_04**	US	äh gott sei **dank** ((Lachansatz)) (0.47) am sechzehnten dezember ...
27	**FOLK_E_00055_SE_01_T_05**	US	glas wein haben °h **danke** gerne ((Lachansatz)) wobei des ...
28	**FOLK_E_00055_SE_01_T_05**	US	prost prost prost **dankeschön** darauf dass wir ...
29	**FOLK_E_00057_SE_01_T_01**	ME	°h gut hh° (0.33) willkommen **danke** h° (0.34) äh h° wir haben ...
30	**FOLK_E_00057_SE_01_T_01**	ME	wieder rein °h jawohl **danke** ...
31	**FOLK_E_00058_SE_01_T_01**	HN	... jahrn °h gott sei **dank** noch mal im ...
32	**FOLK_E_00058_SE_01_T_01**	HN	okay (0.43) ganz herzlichen **dank** lara ich danke ...
33	**FOLK_E_00058_SE_01_T_01**	XL	dank lara ich **danke** auch wir bedanken ...
34	**FOLK_E_00059_SE_01_T_01**	HN	... wichtig °h ganz herzlichen **dank** hm (0.59) so °h
35	**FOLK_E_00042_SE_01_T_01**	LP	da drauf dahin (0.61) **danke** na ja ich ...
36	**FOLK_E_00042_SE_01_T_02**	LK	... so ne art (.) **dankbarkeitsbewusstsein** von den mädels
37	**FOLK_E_00042_SE_01_T_02**	LS	... bisschen mit der **dankbarkeit** dass sie dann ...
38	**FOLK_E_00042_SE_01_T_02**	AM	... das liegt an **dankbarkeit** dass die des ...
39	**FOLK_E_00042_SE_01_T_02**	LK	oder (0.82) ob des **dankbarkeit** is des will ...
40	**FOLK_E_00042_SE_01_T_02**	LP	ach also des **dankbarkeit** ich glaub ich ...