SIGNLAB
AMSTERDAM

LYKE ESSELINK

UNIVERSITY OF AMSTERDAM

WWW.SIGNLAB-AMSTERDAM.NL

TEXT-TO-SIGN TRANSLATION: MAKING INFORMATION ACCESSIBLE

TEXT TO SIGN RESEARCHERS AT SIGNLAB AMSTERDAM

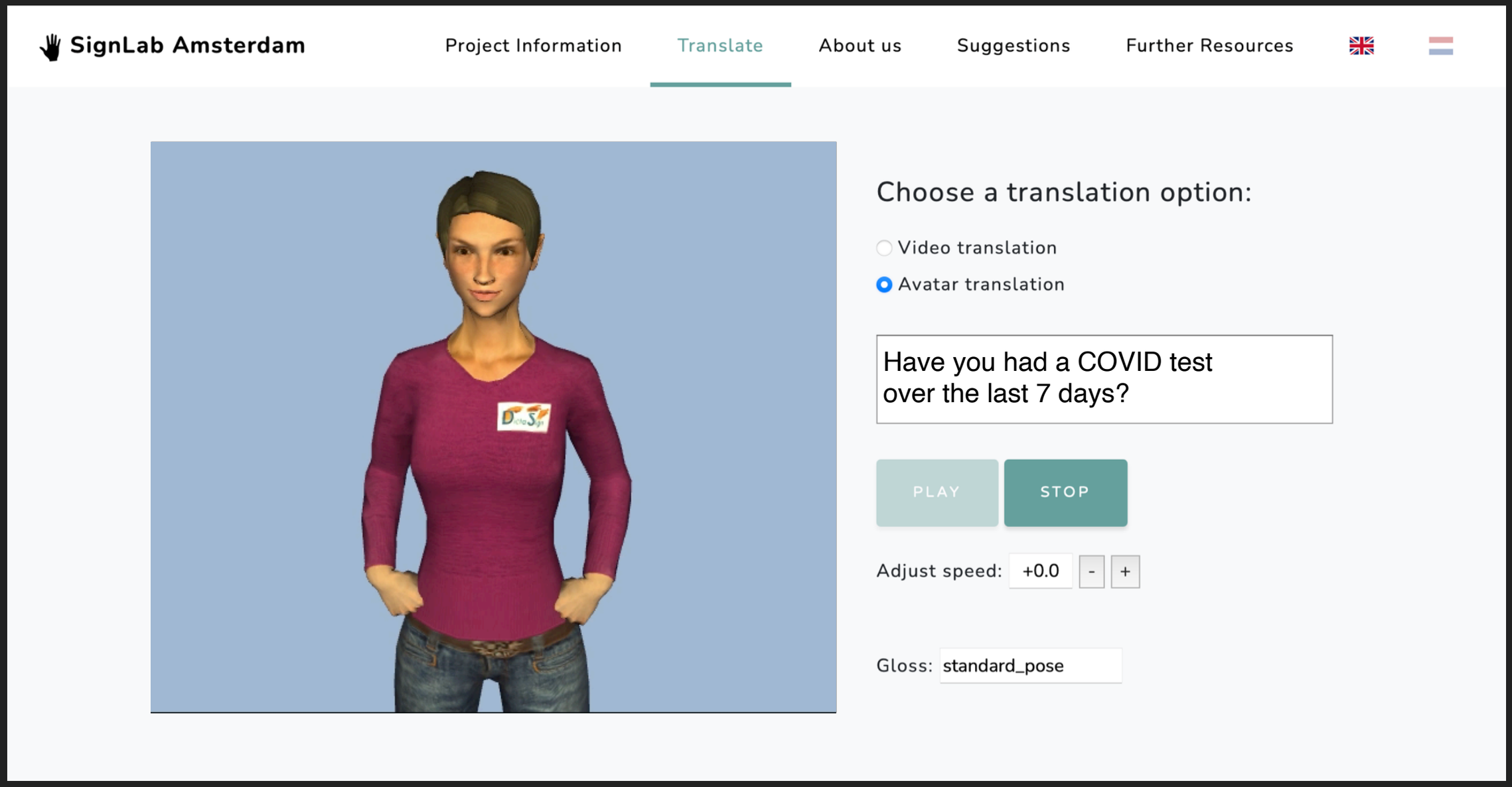
Britt van Gemert

Floris Roelofsen

Lyke Esselink

Shani Mende-Gillings

WHAT DOES IT LOOK LIKE?



OUTLINE

- Why?
- COVID-19 project
- Evaluation
- Results
- Discussion, conclusion & future work

WHY?

DEAF POPULATION AND SIGN LANGUAGES

- WFD: around 70 million deaf people around the world
- Evolved naturally in deaf communities around the world
  - No single, universal sign language
  - The sign language of a country is not directly linked to its spoken language
    - Example: American Sign Language and British Sign Language are very different
    - Both in lexicon and grammar
    - Even within a given country, there are often many variants/dialects
- Sign languages have no writing system (some have been proposed, but none is widely used)
- Sign languages are generally poorly documented, if at all

ACCESSING TEXT/SPEECH WITHOUT TRANSLATION TO SIGN LANGUAGE

- A devil's advocate might ask:
  - Is text/speech to sign translation really necessary?
  - Can't deaf people just read texts, speech transcripts, and subtitles?
  - Or can't they just manage with hearing aids, cochlear implants, and lipreading?

READING

- Deaf people often have great difficulty reading
- Median reading level of 18-year-old deaf adolescents in NL is comparable to 8-year-old hearing children (Wauters et al 2006)
- 68% of deaf adults in NL is low literate (Wauters et al 2017)
- For example: subtitles are too fast, newspapers too complex
- This makes sense: Imagine having to learn to read Thai without ever being told how the characters are pronounced.

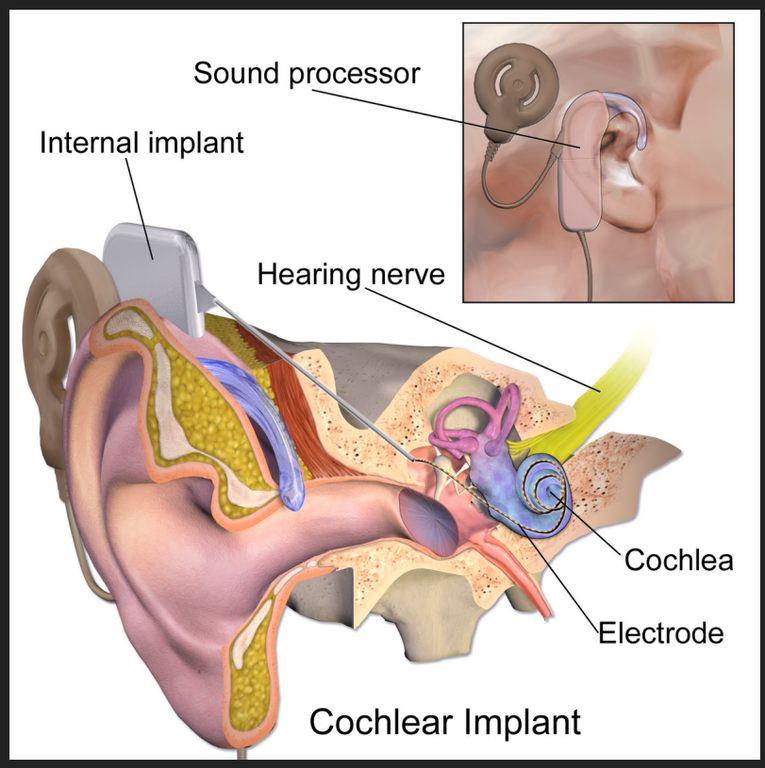
Good morning	สวัสดีตอนเช้า
How are you?	คุณเป็นอย่างไร?

HEARING AIDS

- Hearing aids only work if someone has hearing loss but is not completely deaf
- Not as effective for children who are born with hearing loss as for people who incur hearing loss later in life (they can ‘fill in the gaps’)

COCHLEAR IMPLANTS

- Cochlear implants can give access to speech even if someone is completely deaf



COCHLEAR IMPLANTS

- In some (rich) parts of the world, most deaf children are given an implant around their first birthday.
  - In NL around 80% (de Raeve and van Hardeveld 2021)
  - In the US around 50% (Sorkin 2013)

- Controversial in deaf communities:

  "Rather than locating it in a history of medical progress, they have located it within a history of their own oppression." (Blume, 1999)

COCHLEAR IMPLANT CAVEATS

- Mixed results: many CI users still have poor access to speech
- Effectiveness decreases with background noise and in multi-party conversations
- Surgery is risky, long-term health repercussions are unknown
- CI users become dependent on technology
- CIs have affected language policies and the attitudes of parents, family, and teachers of deaf children to favour speech over sign language

REPERCUSSIONS

- Bottom line: deaf people have limited access to speech and written text
- Reading is often very difficult; hearing aids and cochlear implants can help but do not provide a complete solution and have significant downsides as well
- This language barrier leads to **inequality** and **social exclusion**, with repercussions such as **unemployment** and **depression**. For instance:
  - In 2017 only **53.3% of deaf people in the US were employed**, compared to **75.8%** of hearing people (Garberoglio et al., 2019)
  - In a Norwegian well-being survey in 2007, **21% of deaf people said they felt hopeless**, compared to **4%** of hearing people (Kvam et al., 2007)

COVID-19 PROJECT

PROBLEM STATEMENT

- Communication between healthcare professionals and deaf patients is very challenging
- Even more so during the ongoing COVID-19 pandemic
  - Sign language interpreters (SLIs) often cannot enter hospitals/clinics
  - Interpreting via video relay not always viable
  - Face masks hinder lipreading

DEAF PERCEPTION OF COMMUNICATION CHALLENGES DUE TO COVID-19

- Online questionnaire
  - January - February 2021
  - 179 participants from Dutch deaf community, aged 20-84
- Communication barrier perceived as major threat
  - 88% worry about communication barriers if hospitalised with COVID-19
  - For comparison, only 33% worry about not seeing family and friends

TRANSLATION SYSTEM AT A GLANCE

- Source languages: Dutch and English
- Target language: Sign Language of the Netherlands (NGT)
  - Problem is worldwide, system designed to accommodate other source/target languages
- Domain: Phrases frequently used in hospital setting, especially related to COVID-19
- Format:
  - Some translations offered through videos of human signer
  - Most translations offered by means of signing avatar
- Intended use: Only when a qualified human SLI cannot be employed

IMPORTANT FEATURES OF SIGN LANGUAGES

- Signs are generally not just articulated with the hands
  - Also involve facial expressions, movement of the head, mouth, shoulders, or upper body
    - Known as non-manual components
    - Text-to-sign translation should take both manual and non-manual components into account
- Signs have phonetic properties
  - E.g. initial location, shape and orientation of the hands, possibly movements of hands and other body parts, facial expressions

IMPORTANT FEATURES OF SIGN LANGUAGES

- Non-manual components are also used to convey grammatical features
  - Comparable to intonation in spoken languages
  - E.g. raised eyebrows indicate a question, head shake indicates negation
  - Typically supra-segmental: span across a sequence of signs in a sentence

IMPORTANT FEATURES OF SIGN LANGUAGES

- Sign language utterances are represented as glosses

brow raise

YOU HOLIDAY GO

- Lexical signs in small-caps
- Always involve manual, and often also non-manual components
- Upper tier shows non-manual grammatical markers, horizontal line indicates duration

RESULTING REQUIREMENTS FOR A TEXT-TO-SIGN TRANSLATION SYSTEM

- System should be able to integrate:
  - Manual and non-manual components of lexical signs
  - Non-manual elements that convey grammatical information
- Translating sentences word by word is not fully satisfactory
  - Even when reordering signs in accordance with word order rules of target sign language
  - Will miss grammatical information

TEXT TO SIGN TRANSLATION: TWO APPROACHES

DIRECT: END-TO-END LEARNING

- INPUT: SPEECH/TEXT
- OUTPUT: COORDINATES OF VERTICES ON THE MESH OF A HUMANOID AVATAR

- HASN'T BEEN ATTEMPTED YET
- REQUIRES A LOT OF MOTION CAPTURE DATA
- NOT SURE IF FEASIBLE, BECAUSE OF HUGE NON-DETERMINACY

INDIRECT: VIA INTERMEDIATE REPRESENTATION

- POSSIBLE INTERMEDIATE REPRESENTATIONS:
  - GLOSS
  - PHONETIC REPRESENTATIONS

- STEP 1: **ENCODING** (TEXT —> INT. REP.)
- STEP 2: **SYNTHESIS** (INT. REP. —> AVATAR)

- SEVERAL VARIANTS, DIFFERING IN:
  - TYPE OF INT. REPRESENTATION(S)
  - SYNTHESIS METHOD

CHOICE OF SYNTHESIS METHOD: THREE OPTIONS

MOTION CAPTURE

- HIGH-QUALITY LEXICAL SIGNS
- REQUIRES EXPENSIVE EQUIPMENT, & MUCH WORK TO CLEAN DATA
- OPEN QUESTION HOW TO INCORPORATE GRAMMATICAL NON-MANUALS

MANUAL KEYFRAME ANIMATION

- LOWER QUALITY LEXICAL SIGNS
- NO EXPENSIVE EQUIPMENT, BUT A LOT OF MANUAL LABOR
- OPEN QUESTION HOW TO INCORPORATE GRAMMATICAL NON-MANUALS

SCRIPTED KEYFRAME ANIMATION

- GENERATED FROM **PHONETIC REPRESENTATIONS** OF SIGNS
- LOWER QUALITY LEX. SIGNS
- NO EXPENSIVE EQUIPMENT, RELATIVELY LITTLE MANUAL LABOR
- QUITE EASY TO INCORPORATE GRAMMATICAL NON-MANUALS

```
<hamgestural_sign gloss="WAT">
    <sign_manual>
        <handconfig handshape="finger2" thumbpos="across"/>
        <handconfig extfidir="u"/>
        <handconfig palmor="d"/>
        <location_bodyarm location="shoulders" side="right_at"/>
        <wristmotion motion="swinging" size="small"/>
    </sign_manual>
    <sign_nonmanual>
        <mouthing_tier>
            <mouth_gesture movement="L30"/>
        </mouthing_tier>
        <facialexpr_tier>
            <eye_brows movement="FU" speed="0.8"/>
            <eye_lids movement="SB" speed="0.8"/>
            <eye_gaze direction="AD" speed="0.8"/>
        </facialexpr_tier>
        <head_tier>
            <head_movement movement='SL' />
        </head_tier>
    </sign_nonmanual>
</hamgestural_sign>
```

Manual

handshape

location

movement

Non-manual

mouth

face

head

Fig. 1. SiGML encoding of the NGT sign WAT ('what').

<SIGML>
YOU
</SIGML>

<SIGML>
EAT
</SIGML>

<SIGML>
WHAT
</SIGML>

<SIGML>
PALMS-UP
</SIGML>

JASIGNING
AVATAR
ENGINE

WHAT WOULD YOU LIKE TO EAT?

CHOICE OF INTERMEDIATE REPRESENTATION(S)

1. Only gloss
2. Only phonetic representation
3. Hybrid: both gloss and phonetic representations

GLOSS APPROACH

TEXT → GLOSS → ANIMATION

- Example: Brazilian company *HandTalk* (handtalk.me/en)
- Source: English / Brazilian Portugese
- Target: American Sign Language / Brazilian Sign Language
- Machine learning to map input text to corresponding single-tier glosses
- Combination of key-frame animation and motion capture for animations
- No incorporation (yet) of grammatical non-manuals

PHONETIC APPROACH

TEXT → PHONETIC REPRESENTATION → ANIMATION

- Examples: several systems based on SiGML and JASigning
- Machine learning for mapping text to phonetic representations is not possible
  - Would require large parallel corpora of texts and corresponding phonetic representations
  - Not available, very costly to create
- Manually generating phonetic representations is highly time-consuming and requires expert knowledge of SiGML or similar formalism

COMPLEMENTARY PROS AND CONS

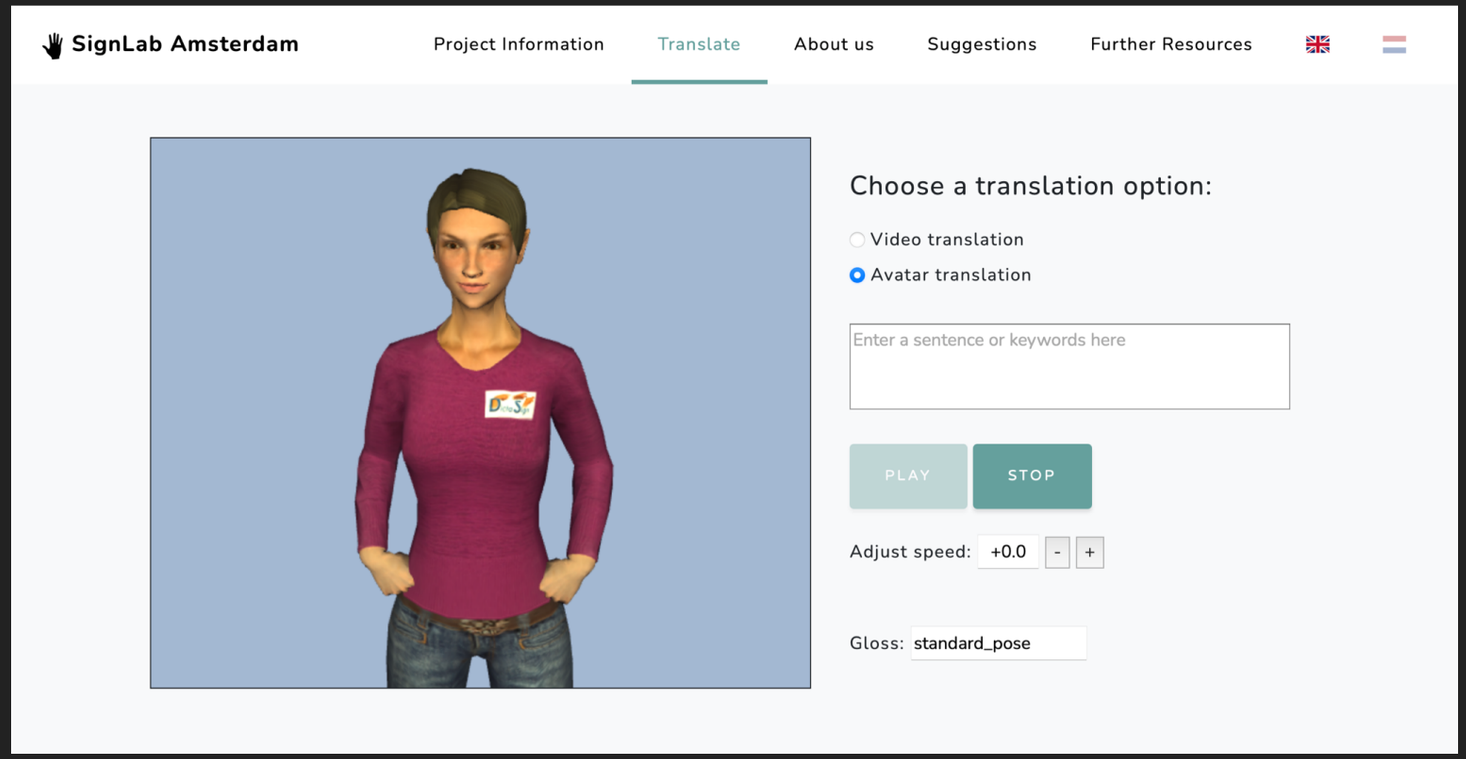
- Gloss approach
  - \+ Enables use of machine learning for Text –> Gloss
  - \- Animation of lexical signs involves heavy manual work/expensive equipment
  - \- Not clear yet how to integrate grammatical non-manuals with lexical signs
  - \- All parts of the system are tailor-made for a particular target sign language
- Phonetic approach
  - \- Initial step, Text –> Phonetic Representation, involves heavy manual work
  - \+ Does not require expensive equipment
  - \+ Grammatical non-manual features can be integrated with lexical signs
  - \+ Synthesis component is not language-specific, can be used for any sign language

HYBRID APPROACH

TEXT —> GLOSS —> PHONETIC REPRESENTATION —> ANIMATION

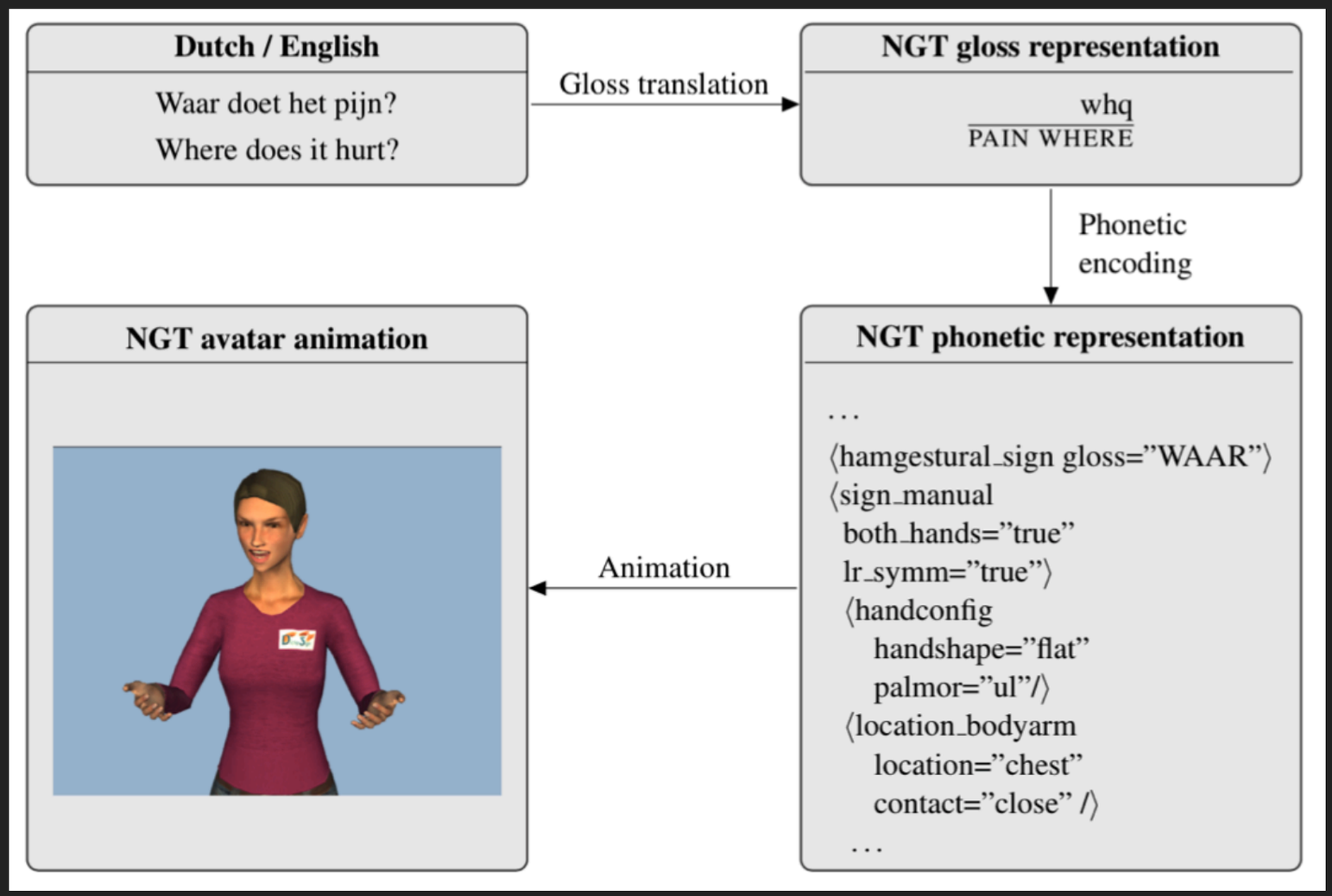
- Example: SignLab Amsterdam COVID application

www.signlab-amsterdam.nl



HYBRID APPROACH

TEXT —> GLOSS —> PHONETIC REPRESENTATION —> ANIMATION



HYBRID APPROACH

TEXT —> GLOSS —> PHONETIC REPRESENTATION —> ANIMATION

- Text –> Gloss
  can be done with rule-based grammar or with machine learning, depending on use case requirements and data availability
- Gloss –> Phonetic Representation
  can be fully automated, rule-based, integrating grammatical non-manuals
- Gloss –> Phonetic Representation –> Animation
  This part of the pipeline is not language-specific, can be applied universally

HYBRID APPROACH

TEXT —> GLOSS —> PHONETIC REPRESENTATION —> ANIMATION

- Generating phonetic representations for lexical signs and grammatical markers is still time-consuming and requires expert knowledge of SIGML
- The SIGML formalism and the JASigning avatar engine need to be improved in several ways, adding more control over movements and timing, making transitions more natural and smooth

MOST SUITABLE APPROACH DEPENDS ON

- Use case requirements
  - High precision (rule-based) vs broad coverage (ML)
  - Visual quality (motion capture) vs scalability (scripted KFA)
- Available resources
  - Parallel data for ML?
  - Descriptive grammar for rule-based translation?
  - Motion capture equipment?
  - Timeframe for development (e.g. COVID)?

IMPLEMENTATION

- Implementation choices were informed by practical considerations:
  - System had to be developed within a short time-frame
  - High accuracy more important than broad approximate coverage
- Aim has *not* been to automate the entire translation process
- Focus on automating the mapping from glosses to phonetic representations
  - Has not been done previously (in non-modular approaches, these two levels do not co-exist)

COLLECTING PHRASES FOR TRANSLATION

- Phrases commonly used during the diagnosis and treatment of COVID-19
  - Based on consultations with healthcare professionals and direct experience

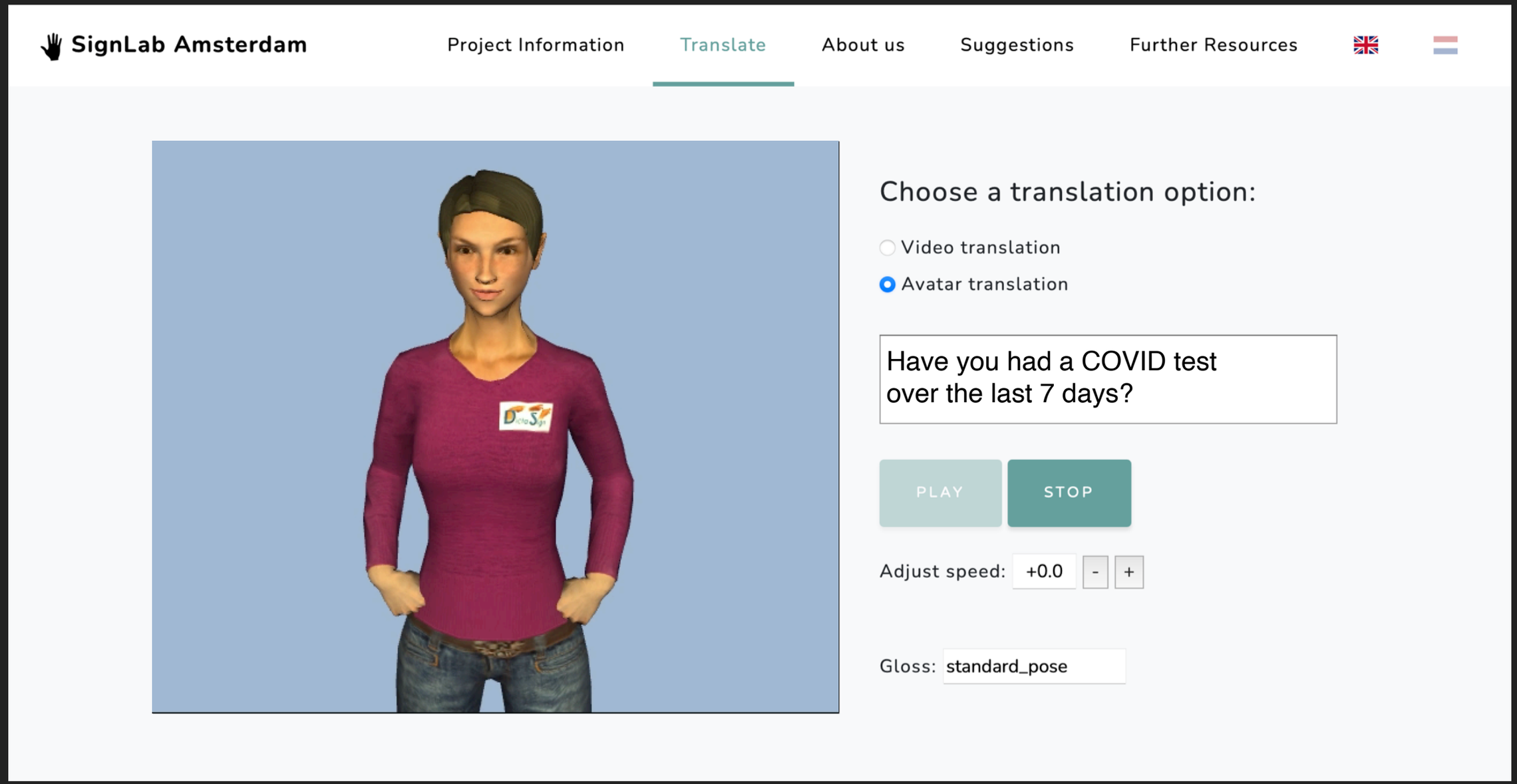
VIDEO ONLY	AVATAR ONLY	HYBRID
▸ EMOTIONAL SENTENCES ▸ COMPLEX SENTENCES ▸ INFORMED CONSENT	▸ SENTENCES WITH MANY VARIATIONS (E.G., TIME OF DAY)	▸ SENTENCES THAT DO NOT FALL INTO OTHER CATEGORIES ▸ BOTH VIDEO AND AVATAR TRANSLATIONS OFFERED

- 139 sentences for video translation
- 7720 sentences for avatar translation

CONSTRUCTING PHONETIC REPRESENTATIONS

- System needs to operate fast at run-time
  - Pre-processed all sentences and stored their SiGML representations in a database
  - System queries database at run-time, does not compute SiGML representations on the fly
- During pre-processing, the Gloss → SiGML encoding has been automated
  - Given a gloss, we first retrieve the SiGML encodings of the lexical signs in that gloss
  - And then adapt this code to integrate non-manual grammatical elements
  - Post-editing is needed in some cases, but automation saves a lot of time

COVID-19 PROJECT



EVALUATION

EVALUATION

- No generally accepted methodology for evaluating comprehensibility of avatars for text-to-sign translation
- Previous methods usually involve on-site experiments (Gibet et al. 2011; Smith and Nolan 2016; Ebling and Glauert 2016; David and Bouillon 2018; Huenerfauth 2006; Kacorri et al. 2015)
- COVID-19 pandemic calls for online evaluation (Quandt et al. 2021; Schnepp et al. 2011)

GOALS

Main goals

1. Individual sign recognition: To what extent do deaf NGT users recognise the individual signs that the avatar translations consist of?
2. Sentence comprehension: To what extent do deaf NGT users understand the avatar translations as intended at sentence level?
3. Clarity: How clear are the avatar translations that the system produces?

Secondary goals

1. Attitude: How do members of the deaf community in the Netherlands view avatar technology for text-to-sign language translation
2. Use cases: What do they see as potentially beneficial use cases for such technology?

METHODOLOGY

- Two groups
  - 14 participants each
    - Supervised
    - Unsupervised
- Online survey

SET-UP

- Supervised group
  - Participants sign in on Zoom
  - Survey opened on computer controlled by researcher
  - Screen is shared
  - Communication with participant solely in NGT through qualified sign language interpreter
- Unsupervised group
  - Received same survey to complete at home
  - No interpreter or researchers present

MAIN LESSONS LEARNED

Three lessons concerning the design of the questionnaire

1. Structure: Crucial to include video items to obtain baseline
2. Format: Ideally all questions and instructions are presented both in NGT and in text, so that participants can choose their preferred format
3. Length: Both the number of items and the length of each item have to be restricted, see below for more specifics

Two lessons concerning the online execution of the questionnaire

4. Individual sign recognition: Task is not straightforwardly understood --- needs to be clarified with examples and structured response form
5. Transcription of responses: Important to include feedback loop --- participants check textual transcription of their signed responses

STRUCTURE

- Introduction, informed consent
- Background questions (mother tongue, demographic data,...)
- Comprehension of avatar translations
- Comprehension of video translations by deaf signer
- General perception of avatar technology and potential use cases

- Important:
  - Including video translations is crucial to obtain baseline
  - Video items should follow avatar items to avoid learning effect
  - Learning effect in the other direction (avatar => video) is not forestalled

FORMAT

- All questions and instructions were formulated both in NGT (by means of pre-recorded videos) and Dutch text
- Participants chose preferred format
  - Most preferred videos, some text
  - Choice of format was greatly appreciated
  - Signer in videos was deaf, this was also appreciated
- Participants reported that questions and instructions were very clear

LENGTH

- Aim was to keep the overall duration of sessions under 45 minutes
  - 10 minutes for introduction, consent, background questions, and example items illustrating the task
  - 10 minutes for questions about perception of technology and potential use cases at the end
    - So: 25 minutes for actual test items, both avatar and video
- As a consequence, the number of test sentences had to be limited:
  - 12 avatar translations
  - 12 corresponding video translations
- The length of test sentences also had to be restricted to avoid short term memory overload (esp. in the individual sign recognition task)
  - Around 7 signs per sentence

RECOGNITION OF INDIVIDUAL SIGNS

What are the individual signs in this sentence?

What is the meaning of this sentence?

How clearly was this sentence signed?

Not clear — Very clear

0 1 2 3 4 5 6 7 8 9 10

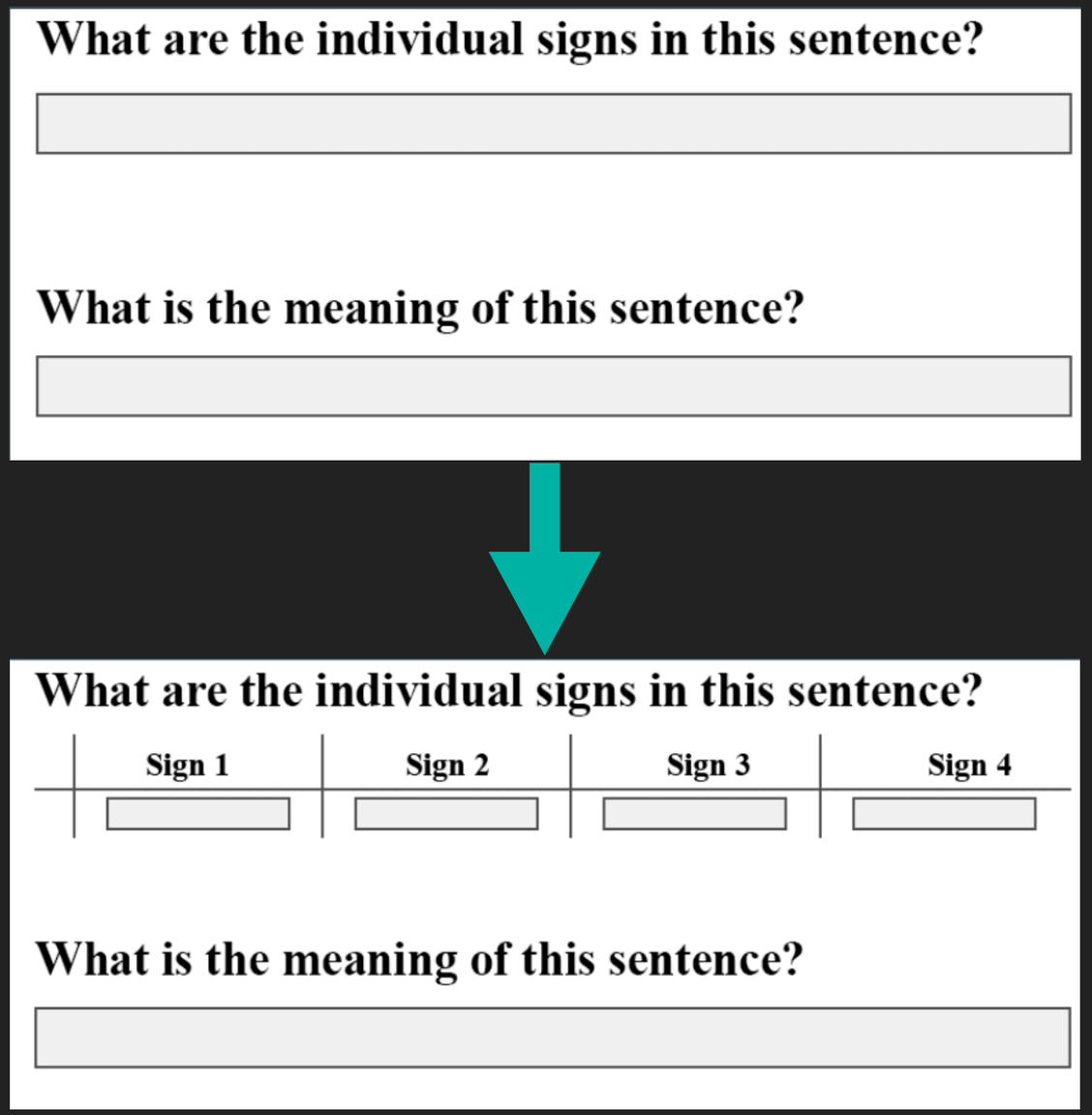
←

→

- In pilot experiment, participants did not understand the first task

RECOGNITION OF INDIVIDUAL SIGNS

- Two adjustments
  - More structured response form
  - Example videos



RECOGNITION OF INDIVIDUAL SIGNS

- Two adjustments
  - More structured response form
  - Example videos

RECOGNITION OF INDIVIDUAL SIGNS

- Two adjustments
  - More structured response form
  - Example videos

What are the individual signs in this sentence?

Sign 1	Sign 2	Sign 3	Sign 4
YOU	EAT	WHAT	QUESTION

What is the meaning of this sentence?

What would you like to eat?

TRANSCRIBING RESPONSES: FEEDBACK LOOP IS CRUCIAL

What are the individual signs in this sentence?

Sign 1	Sign 2	Sign 3	Sign 4

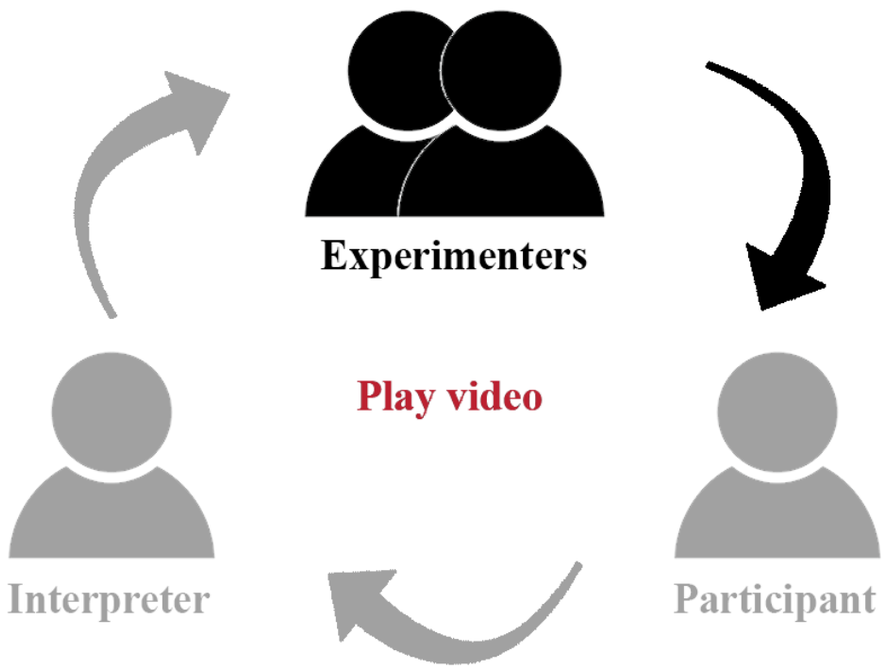
What is the meaning of this sentence?

How clearly was this sentence signed?

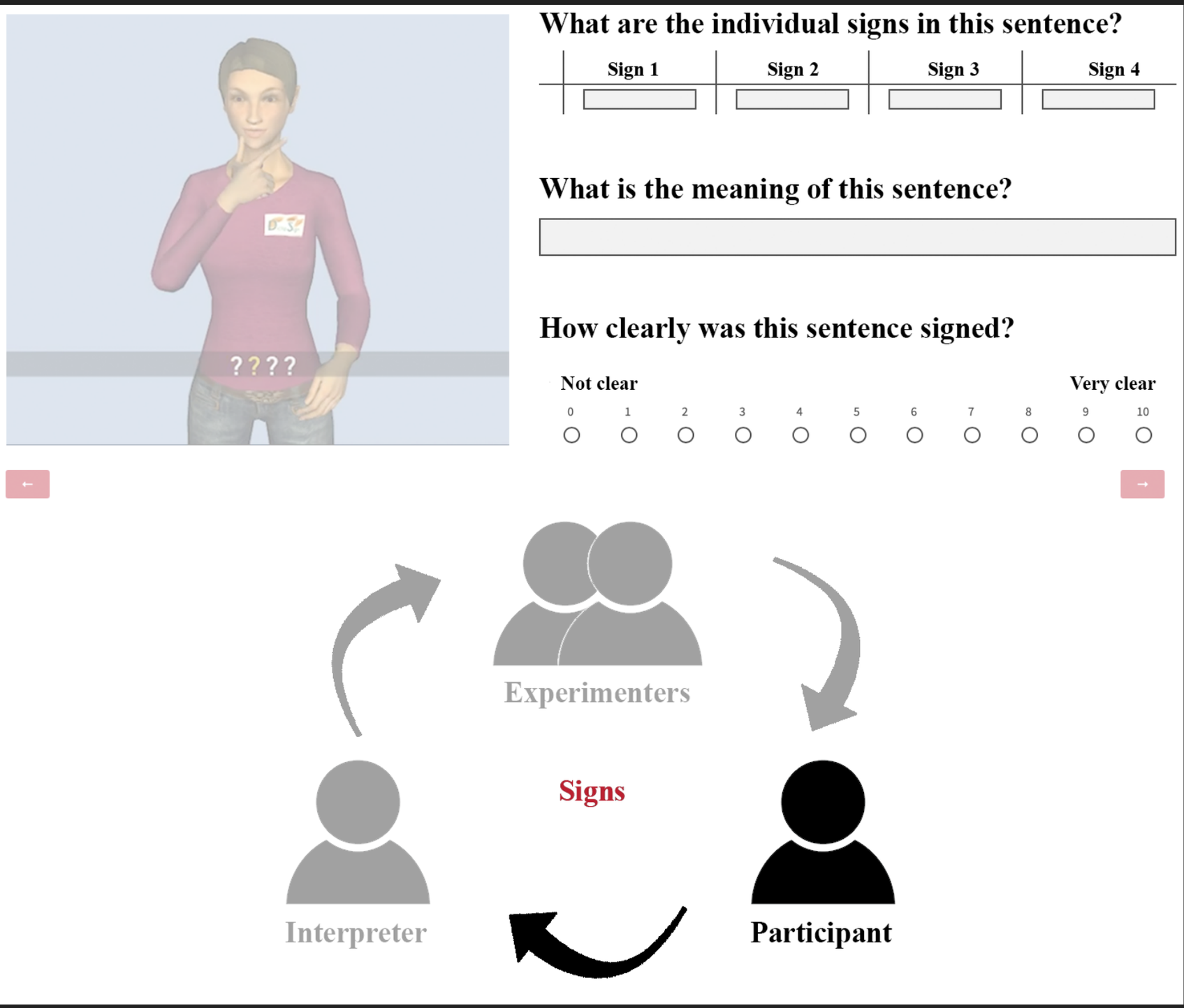
Not clear — Very clear

0 1 2 3 4 5 6 7 8 9 10

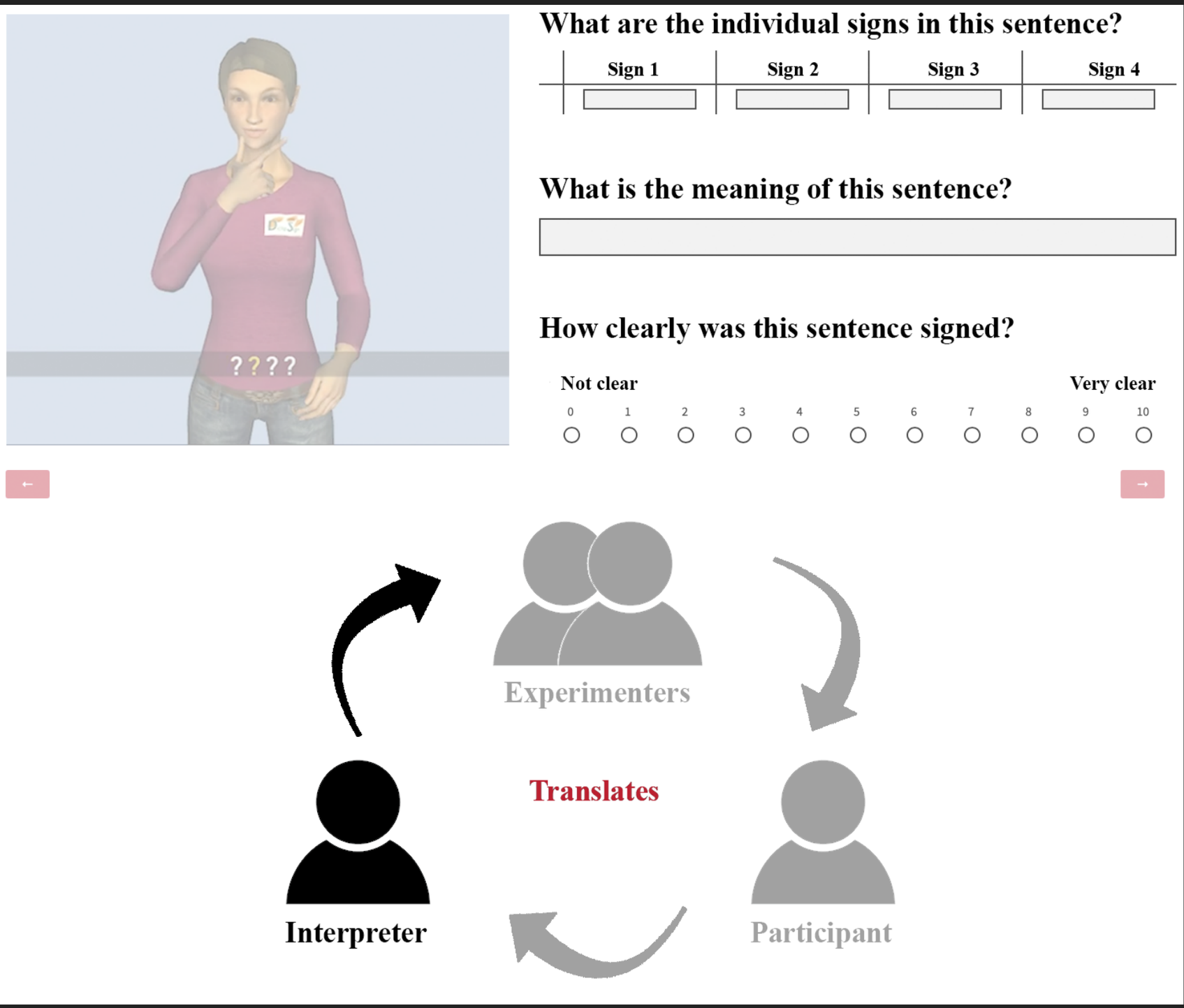
← →



TRANSCRIBING RESPONSES: FEEDBACK LOOP IS CRUCIAL



TRANSCRIBING RESPONSES: FEEDBACK LOOP IS CRUCIAL



TRANSCRIBING RESPONSES: FEEDBACK LOOP IS CRUCIAL

What are the individual signs in this sentence?

Sign 1	Sign 2	Sign 3	Sign 4
YOU	PILLS	USE	QUESTION

What is the meaning of this sentence?

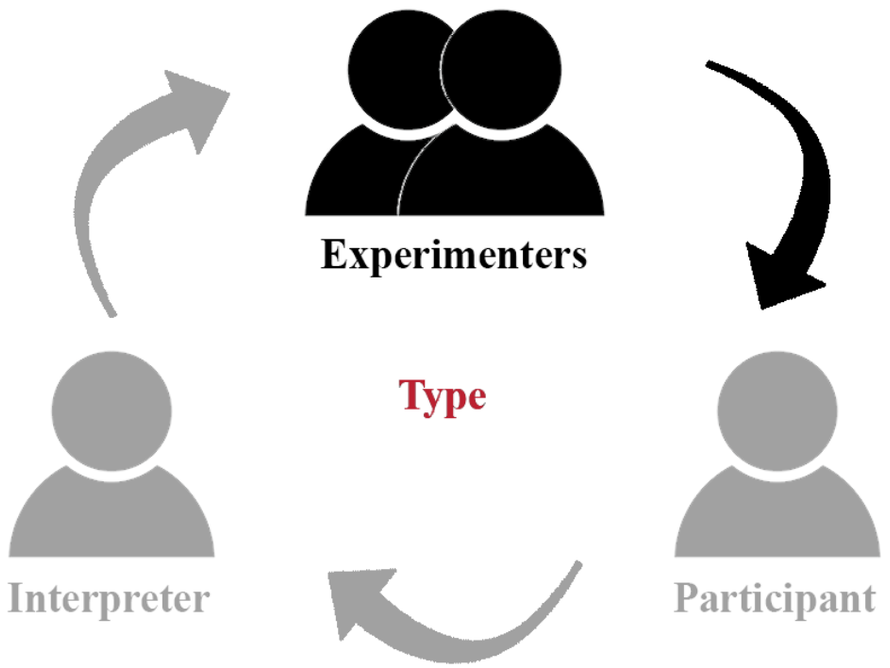
Do you take any pills?

How clearly was this sentence signed?

Not clear — Very clear

0 1 2 3 4 5 6 7 8 9 10

← →



TRANSCRIBING RESPONSES: FEEDBACK LOOP IS CRUCIAL

What are the individual signs in this sentence?

Sign 1	Sign 2	Sign 3	Sign 4
YOU	PILLS	USE	QUESTION

What is the meaning of this sentence?

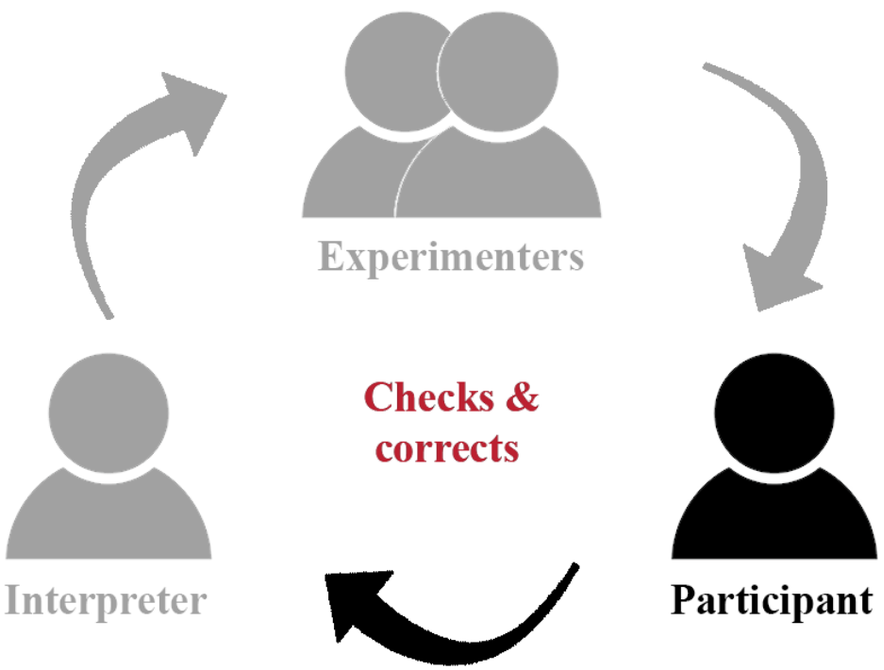
Do you take any pills?

How clearly was this sentence signed?

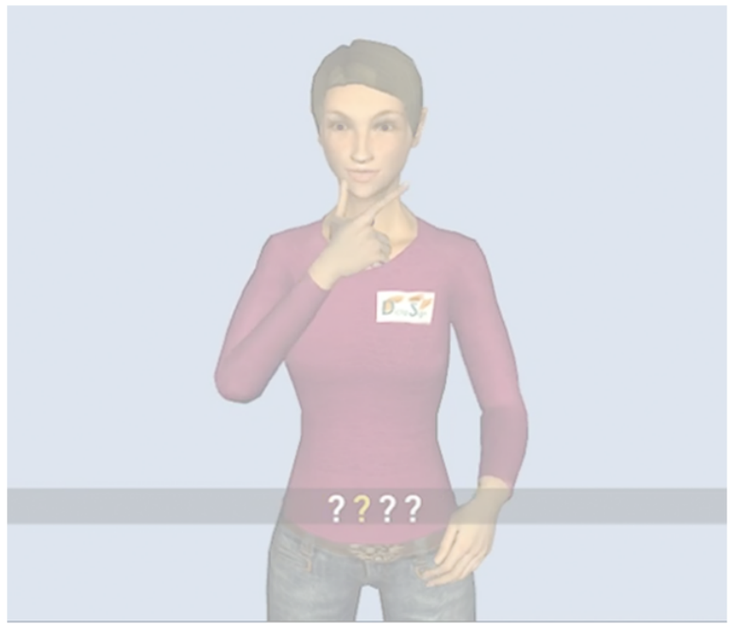
Not clear — Very clear

0 1 2 3 4 5 6 7 8 9 10

← →



TRANSCRIBING RESPONSES: FEEDBACK LOOP IS CRUCIAL



What are the individual signs in this sentence?

Sign 1	Sign 2	Sign 3	Sign 4
YOU	PILLS	USE	QUESTION

What is the meaning of this sentence?

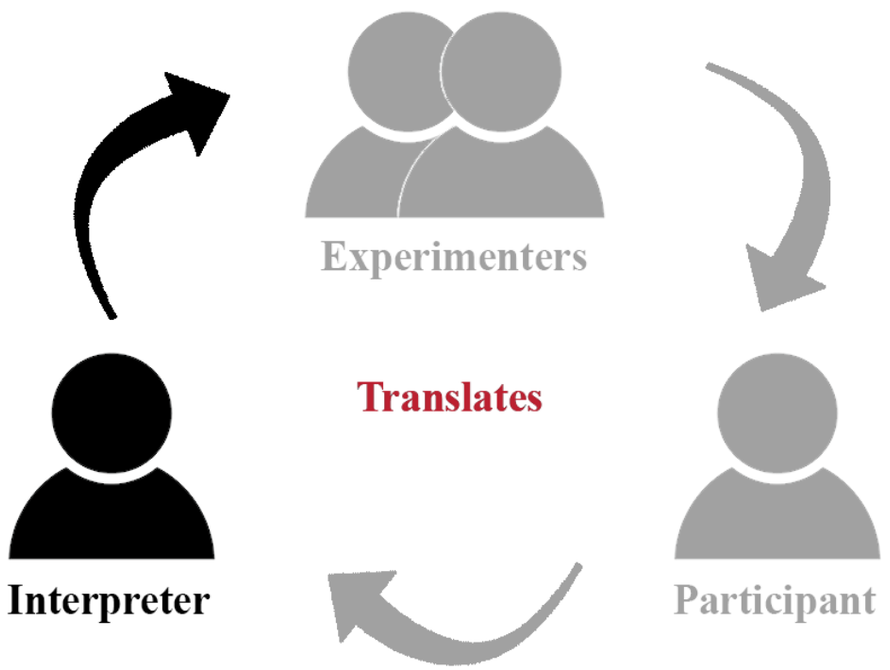
Do you take any pills?

How clearly was this sentence signed?

Not clear — Very clear

0 1 2 3 4 5 6 7 8 9 10

← →



TRANSCRIBING RESPONSES: FEEDBACK LOOP IS CRUCIAL

What are the individual signs in this sentence?

Sign 1	Sign 2	Sign 3	Sign 4
YOU	MEDICINE	USE	QUESTION

What is the meaning of this sentence?

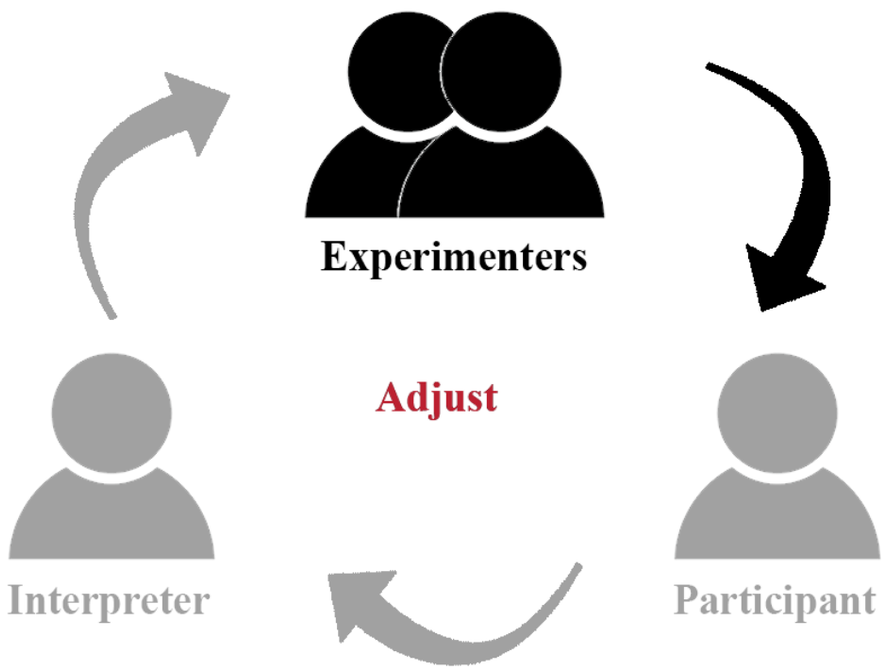
Do you take any **medications**?

How clearly was this sentence signed?

Not clear — Very clear

0 1 2 3 4 5 6 7 8 9 10

← →



TRANSCRIBING RESPONSES: FEEDBACK LOOP IS CRUCIAL

What are the individual signs in this sentence?

Sign 1	Sign 2	Sign 3	Sign 4
YOU	MEDICINE	USE	QUESTION

What is the meaning of this sentence?

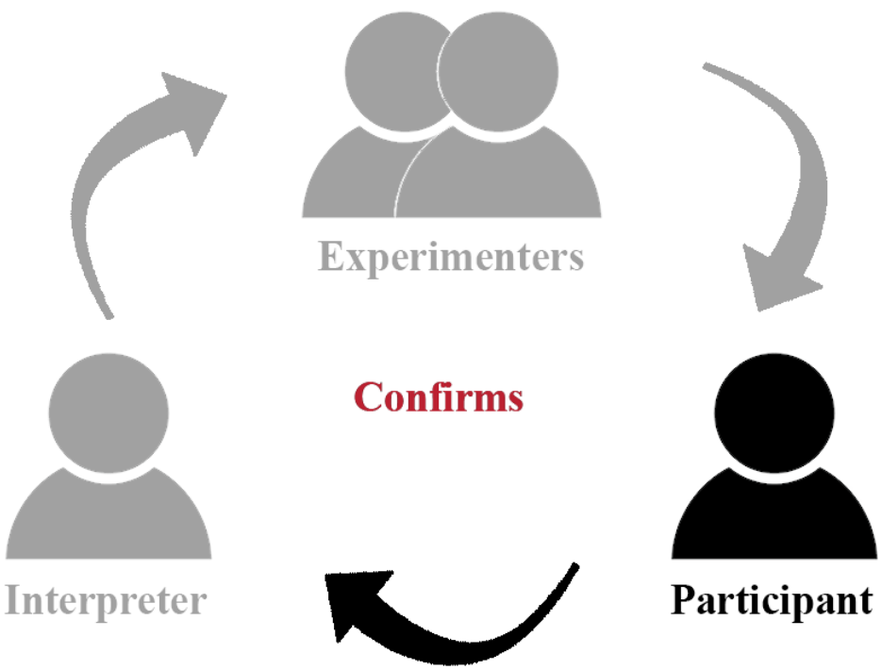
Do you take any medications?

How clearly was this sentence signed?

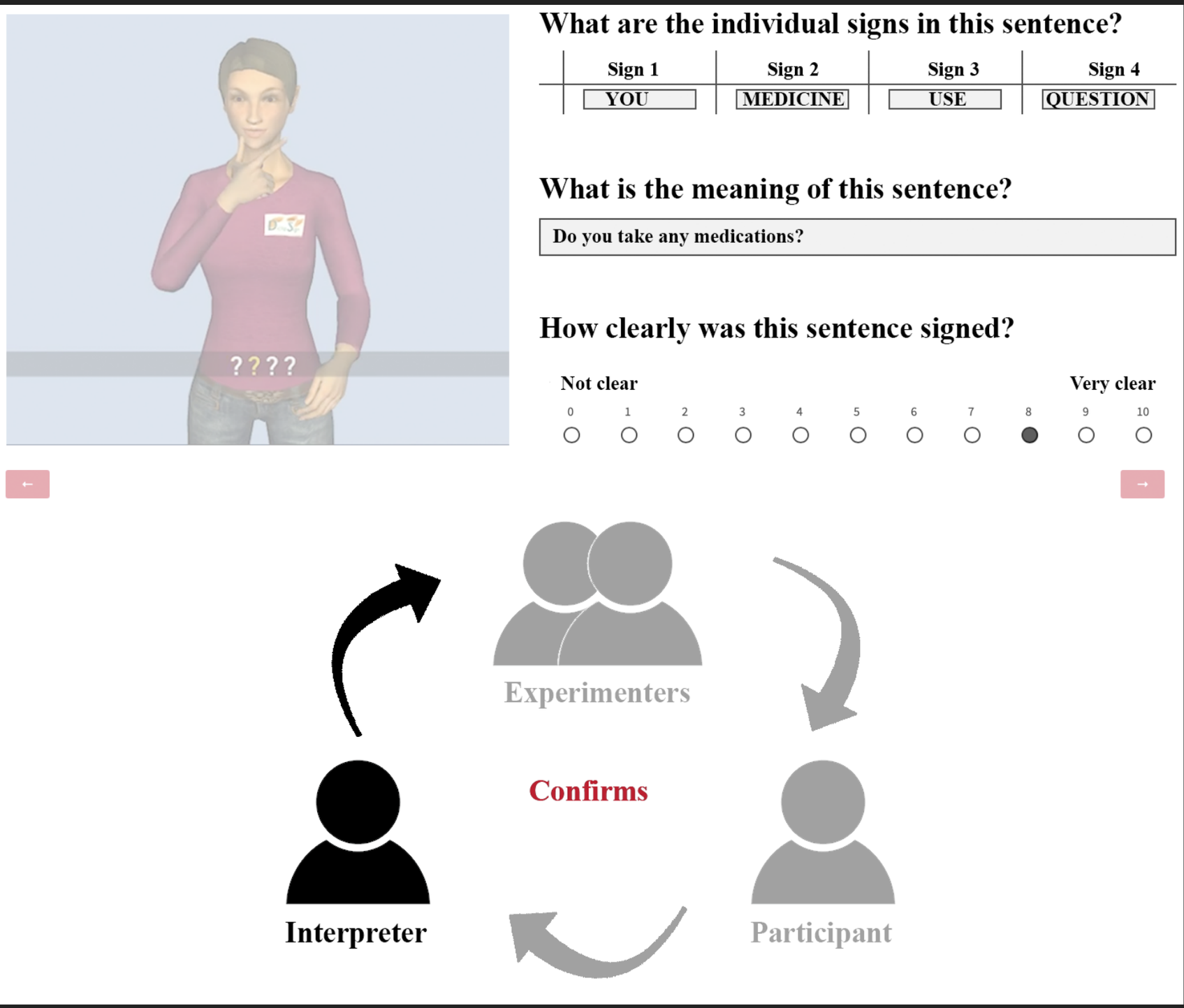
Not clear — Very clear

0 1 2 3 4 5 6 7 8 9 10

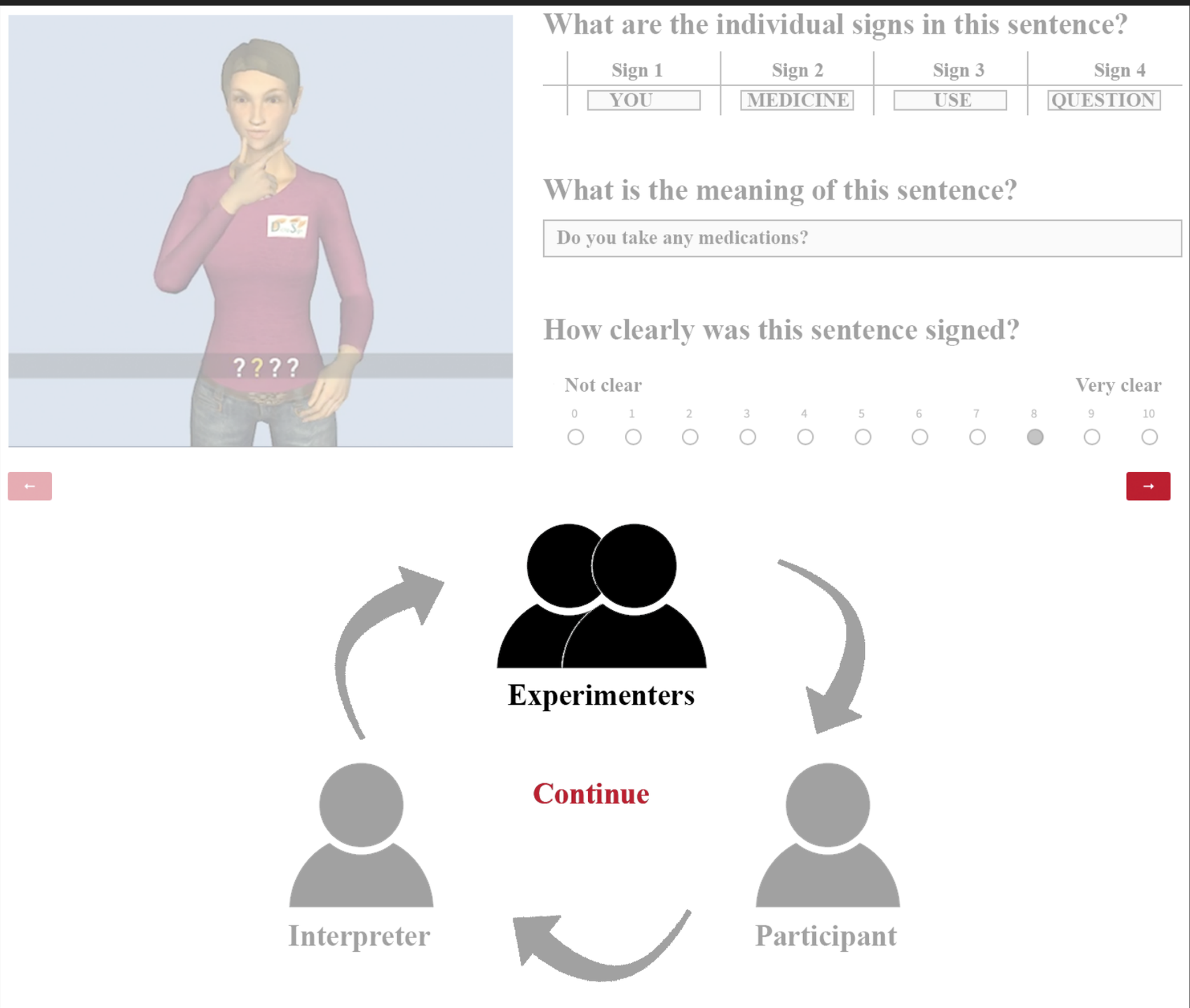
← →



TRANSCRIBING RESPONSES: FEEDBACK LOOP IS CRUCIAL

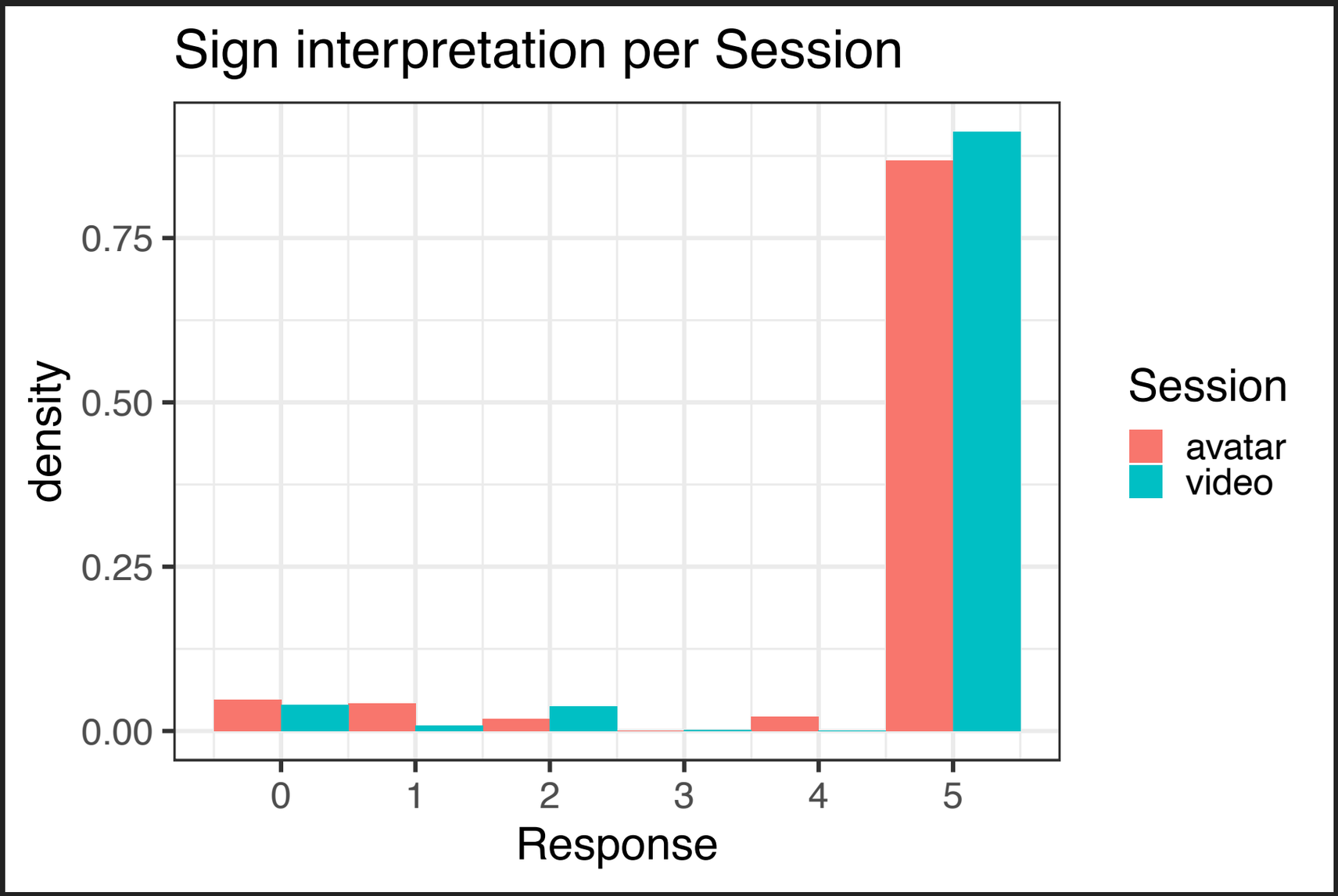


TRANSCRIBING RESPONSES: FEEDBACK LOOP IS CRUCIAL

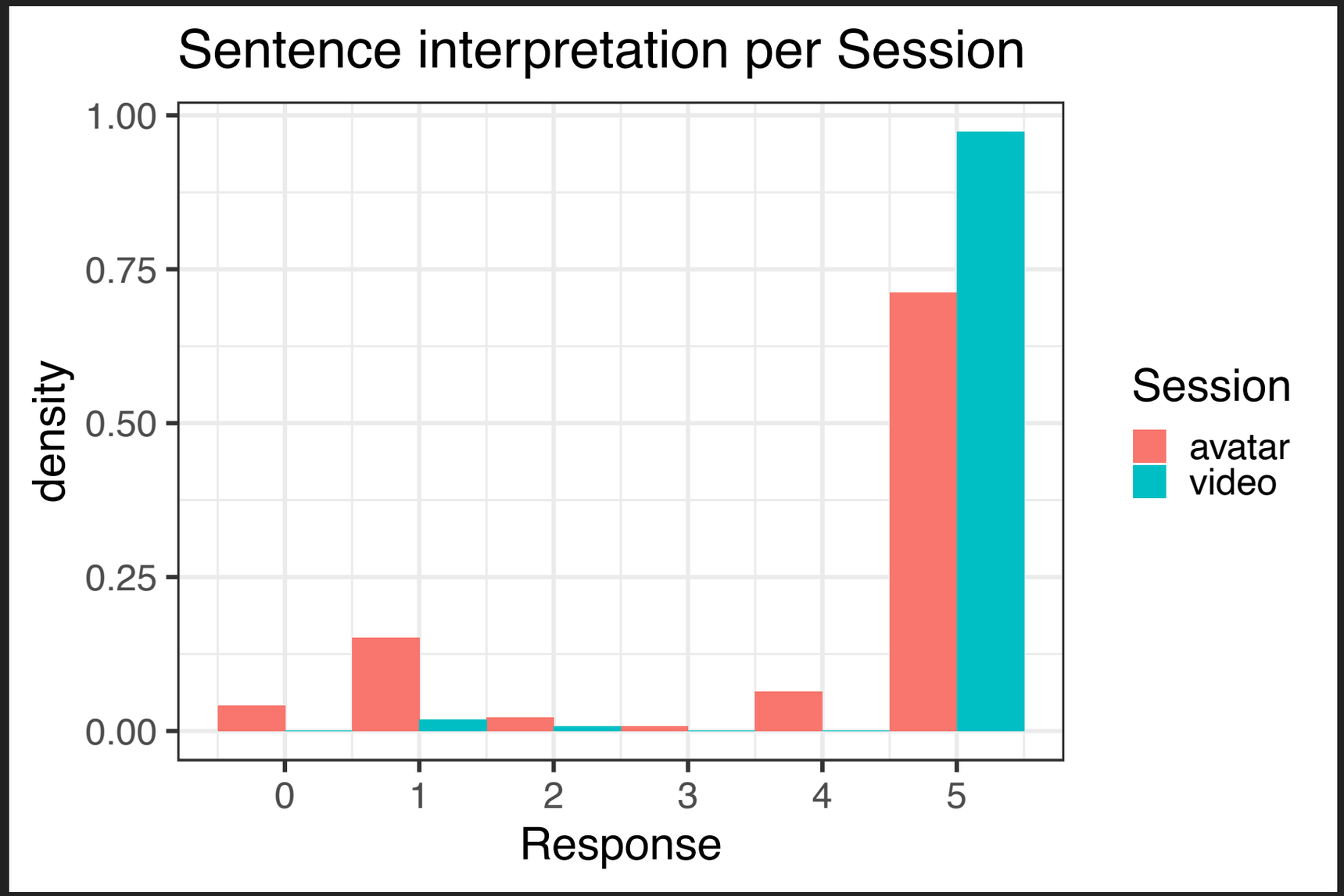


RESULTS

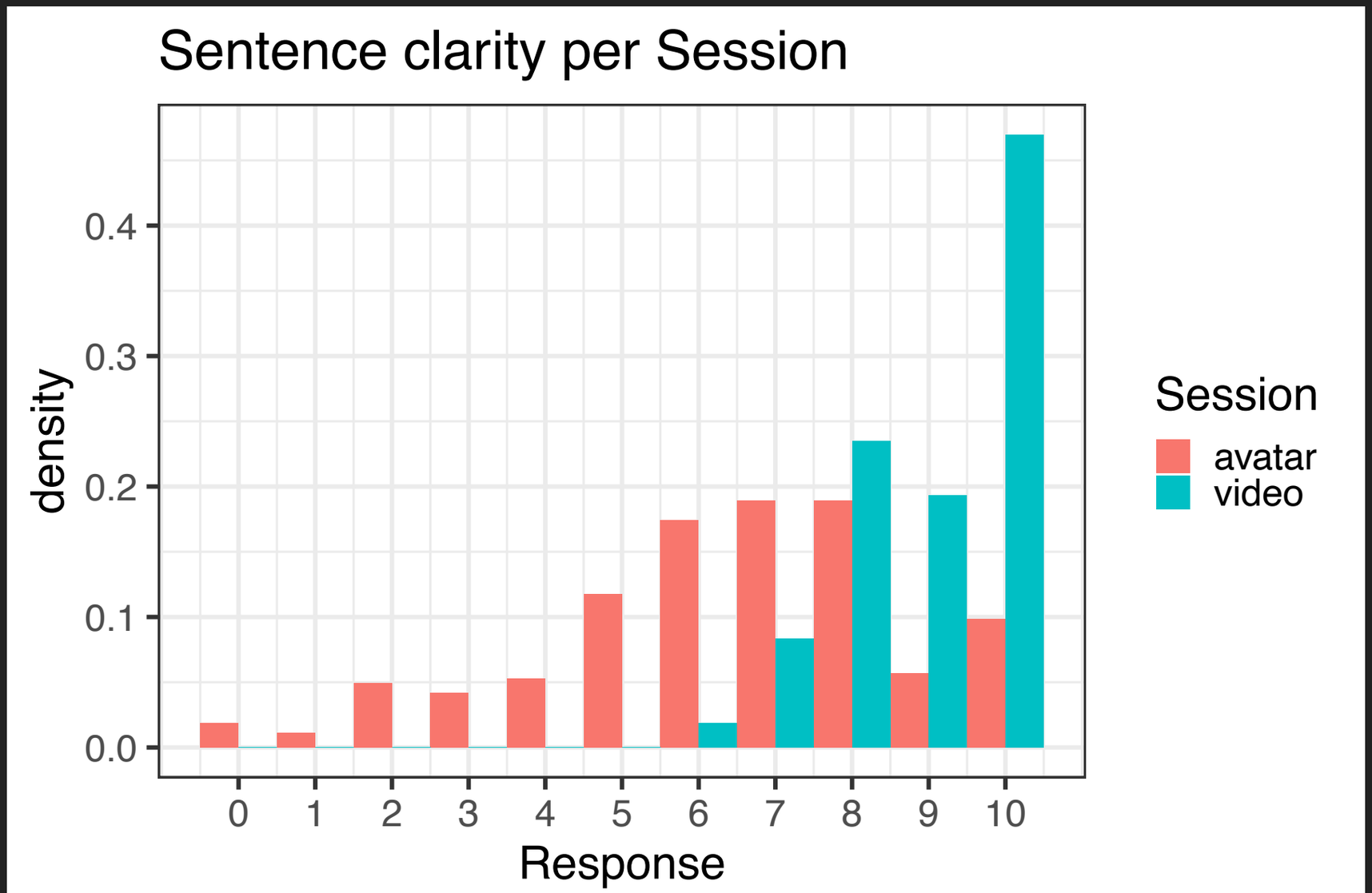
INDIVIDUAL SIGN RECOGNITION



SENTENCE COMPREHENSION



CLARITY


Sentence clarity per Session
density
0.0
0.1
0.2
0.3
0.4
0
1
2
3
4
5
6
7
8
9
10
Response
Session
avatar
video

ATTITUDE

- 82% believed the avatar should be studied further
  - Participants generally found it important and useful
  - Current state not satisfactory enough
  - Multiple participants found the avatar scary at times
- 18% believed the avatar should *not* be studied further
  - Mimicry impossible
  - Jobs from interpreters and deaf people

USE CASES

- Majority
  - Travel information and announcements
  - Government and organisations
  - Shopping and groceries
- Divided opinions
  - Learning sign language
  - Medical environment

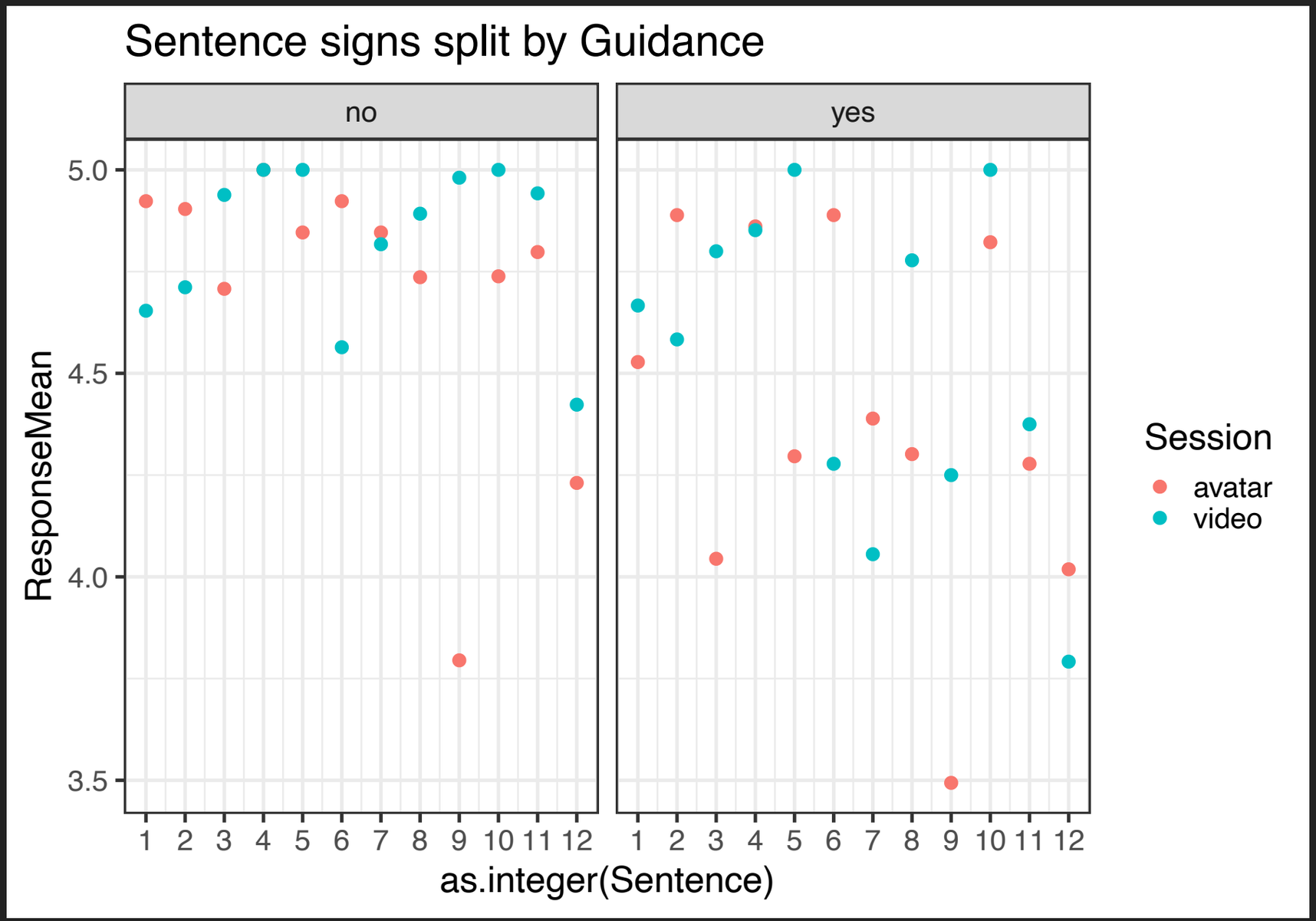
FEEDBACK ON METHODOLOGY

- Questions were easy to answer
  - Supervised slightly more so than unsupervised
- Example question
- Participants felt taken seriously
  - Feedback loop
- Avatar and video sections

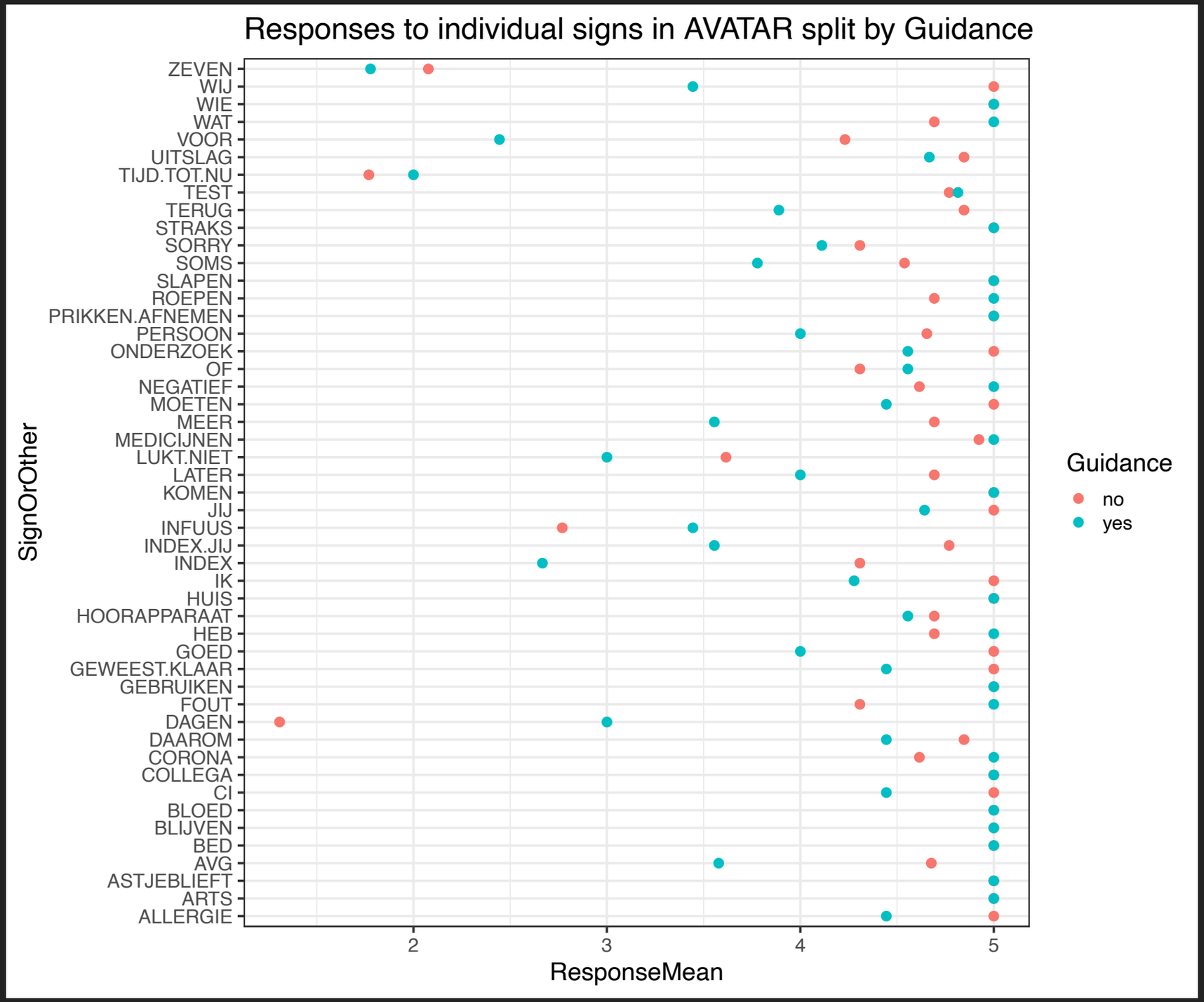
FEEDBACK ON METHODOLOGY

- Ability to provide feedback and give suggestions during the testing process
- Opinions on the presence of an interpreter were divided
  - Supervised participants appreciated it
  - Unsupervised participants indicated it would not have made it easier for them
  - Participation bias?

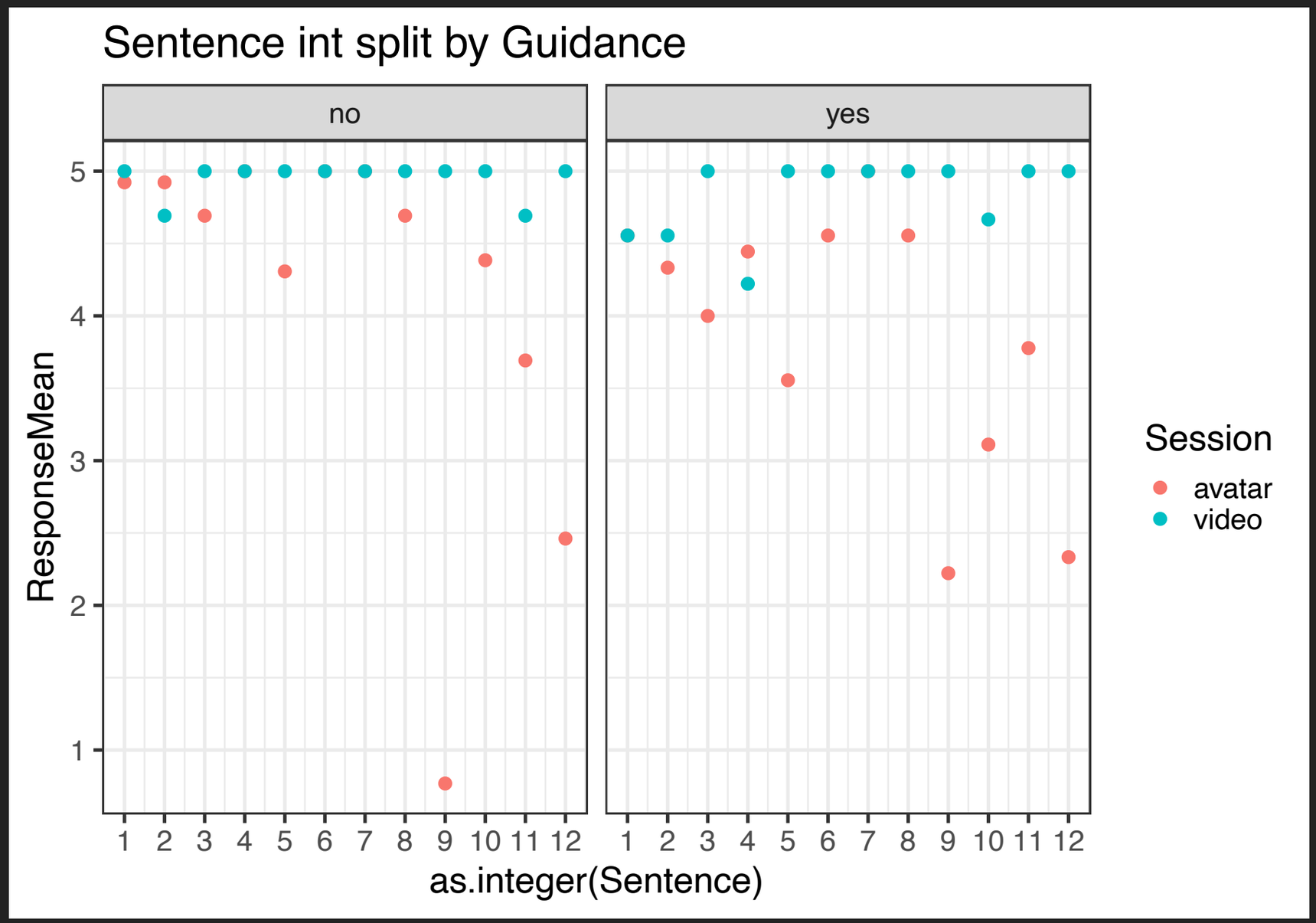
EFFECT OF SUPERVISION



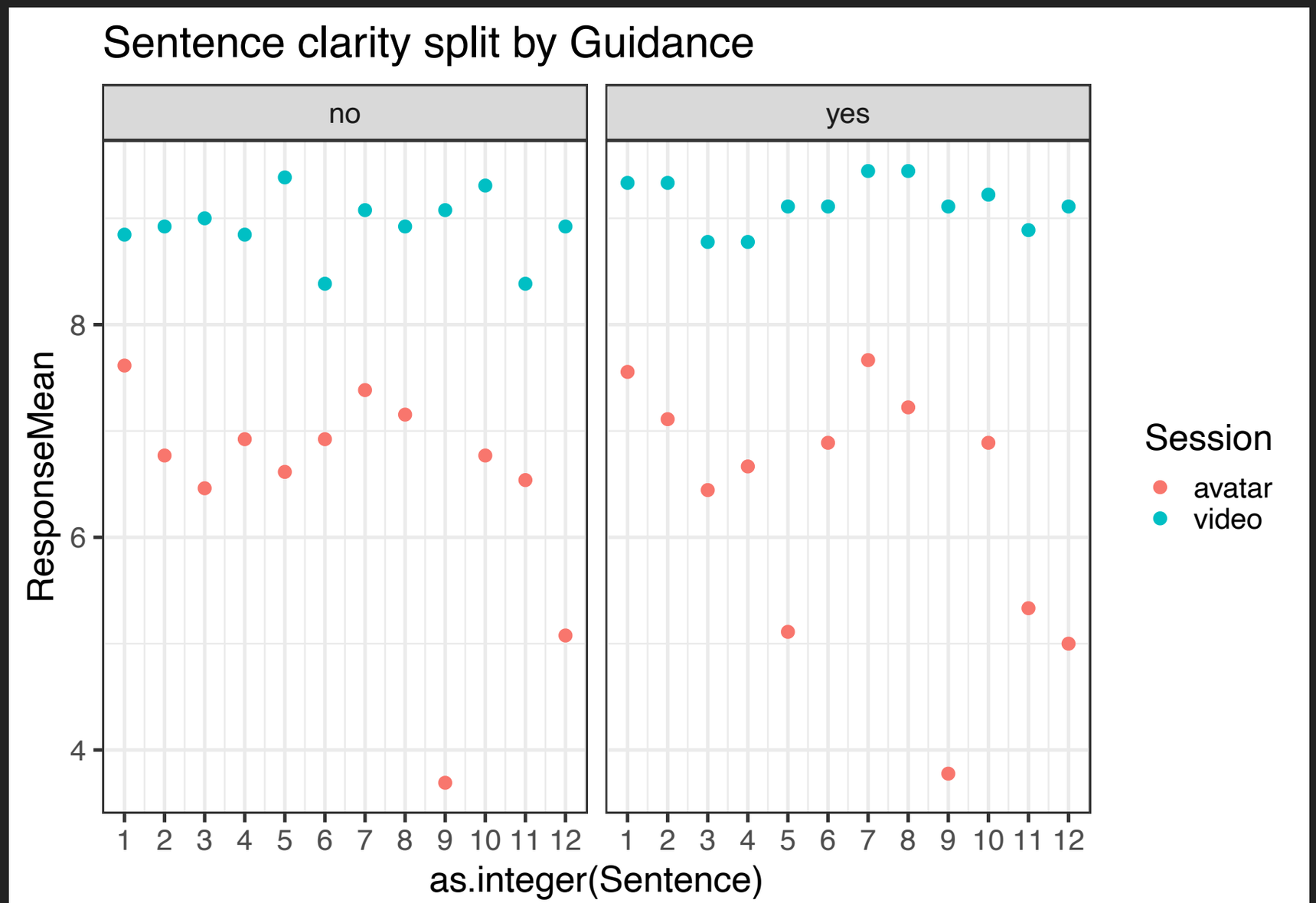
EFFECT OF SUPERVISION



EFFECT OF SUPERVISION



EFFECT OF SUPERVISION



DISCUSSION, CONCLUSION & FUTURE WORK

DISCUSSION

- Avatar technology
  - \- Can be less natural and more difficult to comprehend than video
  - \+ Scales up more easily than video translations
- Video translation
  - \+ Better in terms of naturalness and comprehensibility
  - \+ More likely to make patients feel comfortable
  - \- Does not scale up efficiently
- General advantage of machine translation (video/avatar) over human translation: privacy
- System does currently not support complex dialogue
- Unknown how big the learning effect from avatar –> video was

CONCLUSION

- Why research text-to-sign translation
- Investigated application through COVID-19 use case
- Evaluated our prototype
- Shared lessons learned from online evaluation
- Looked at preliminary results
- Discussed various prospects and limitations of the system

FUTURE WORK

- Apply modular approach in other domains (trains, airports)
- Improve avatar visualisation
- Expand database
- Investigate possibility of hybrid motion capture

THANK YOU

UNIVERSITEIT VAN AMSTERDAM

SIGNLAB AMSTERDAM

nederlandse publieke omroep

AURIS

erc

ZonMw

NWO