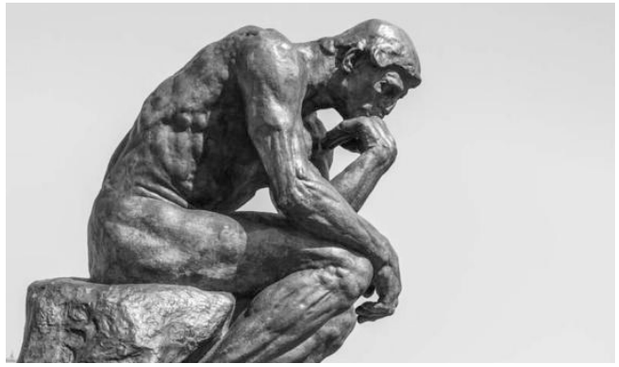
Content analysis and coding procedures

Methods 1, E2021 - Lecture 12
Tuesday 30/11/2021
Fabio Trecca

Methodological approaches to cognitive science

- 1st person methods: Introspective reflection



- 2nd person methods: Asking people (e.g., interview)

- 3rd person methods: Observation (empiricism)

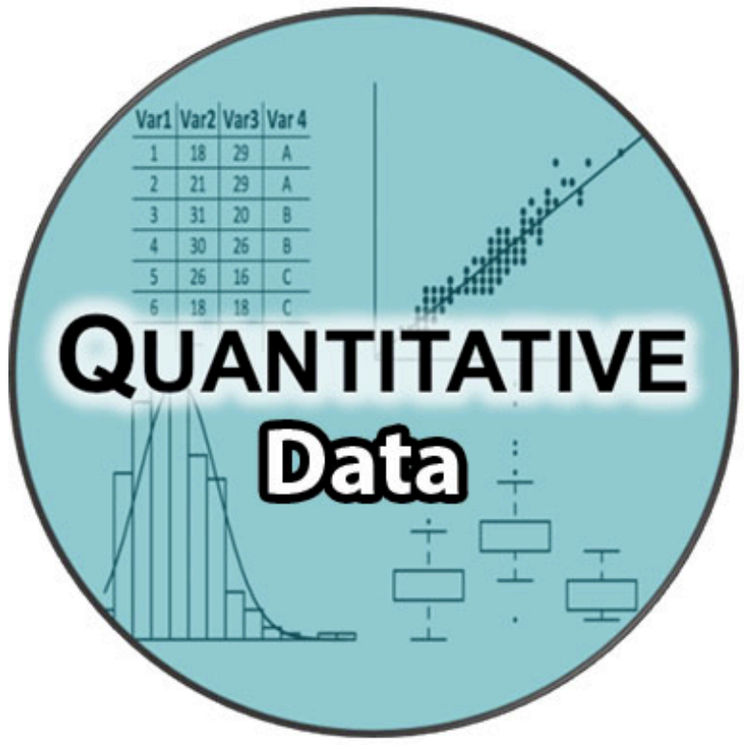
Observation:
rich but noisy human behavior

Experimentation:
control, causality, generalisability

Modeling:
generative mechanisms of human behavior

Quantitative data

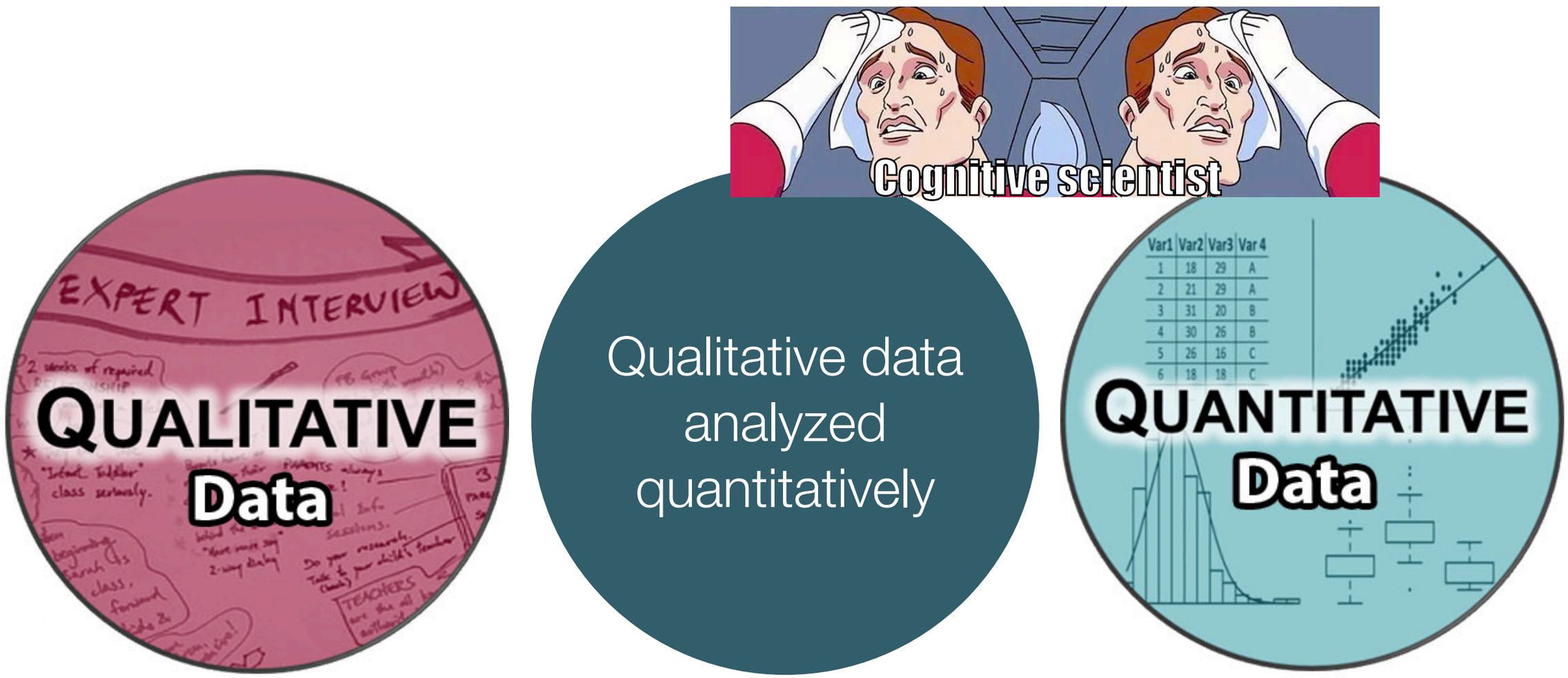
- Accuracy
- Reaction times
- Brain activity
- Eye movements
- Galvanic response
- Heart rate
- Ratings
- Grades
- … etc.
- Systematic, generalizeable, reliable



Qualitative data

- Capture human phenomenological experiences
- E.g. meaning in:
  - (complex) texts
  - social interactions
  - non-verbal communication / gestures
  - emotional expressions
  - visual experiences
- Systematic, generalizeable, reliable??

Quantitative vs qualitative data (1)



Purely qualitative (Arts)

Purely quantitative (STEM)

Quantitative vs qualitative data (2)

- Some aspects of human cognition are not directly quantifiable / do not come in numeric form
- We need an intermediate process of **coding** = categorizing qualitative phenomena in a format that can be analyzed numerically
  - Automated computer coding: for categories that are well defined / not overlapping (eg. text)
  - Manual human coding: for phenomena that are more nuanced
- How do we overcome subjective biases in coding?

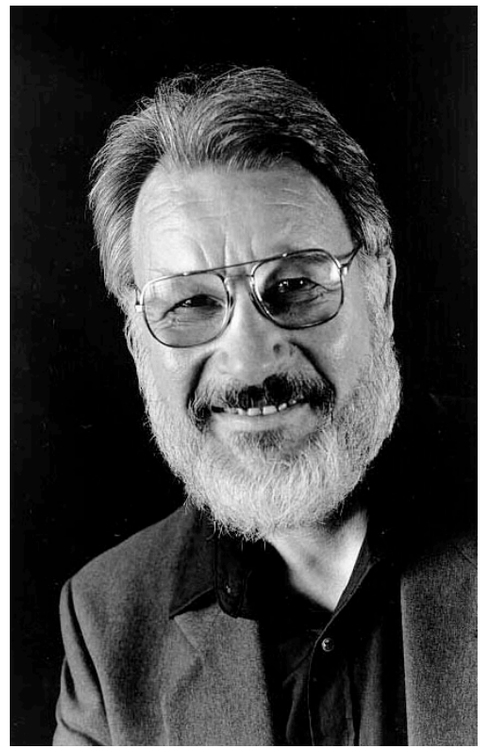
Enter “content analysis”

- “*[A] research technique for making replicable and valid inferences from texts (or other meaningful matter) to the context of their use*” – Krippendorff (2004:18)

- replicability and validity of data acquisition and coding methods

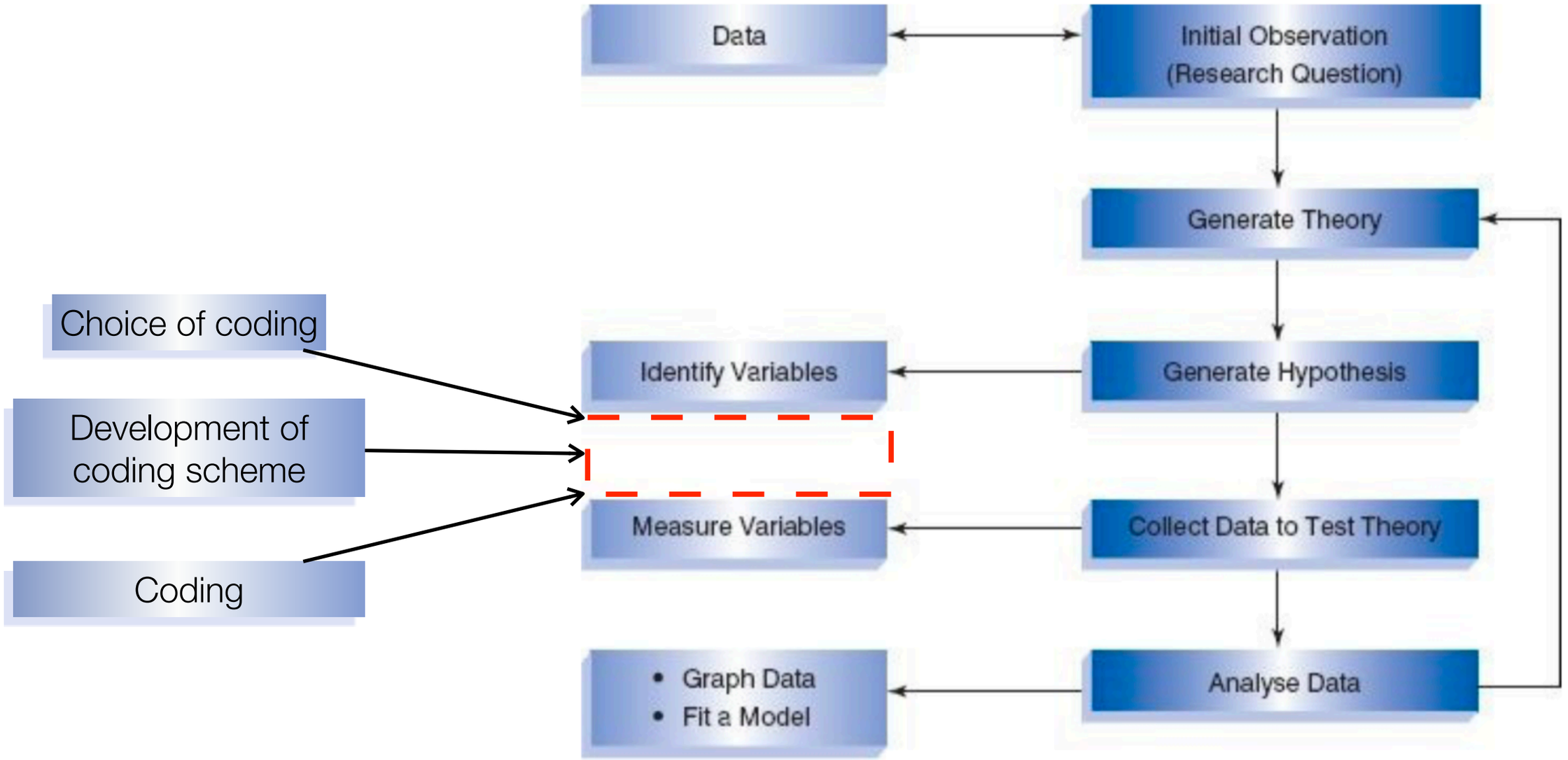
- inferences: meanings are not in the data but in human (conscious) experience

- context: meaning is not absolute, therefore the interpretation criteria must be explicit and systematic



Klaus Krippendorff
(1932)
Professor
of Cybernetics, Language,
and Culture at UPenn

The content-analytical process



Manual coding (1)

- Coders/raters go through the material and count occurrences/categorize events/ annotate

- Advantages:

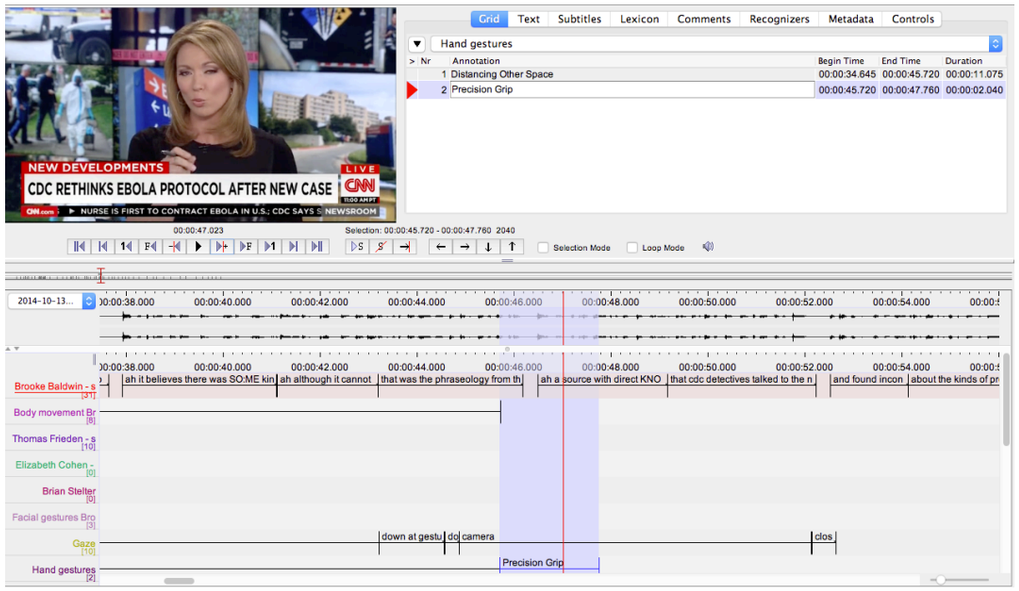
  - nuanced (contextualized meanings, emotions, gestures, multimodal expressions)

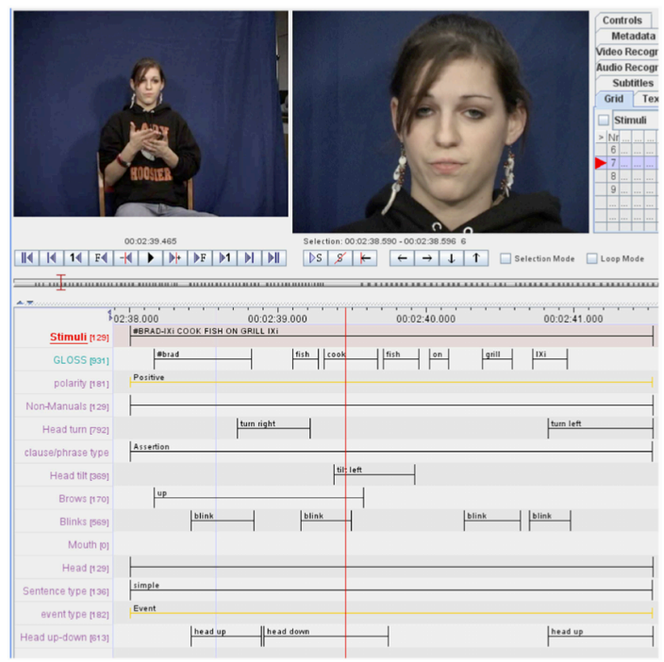
  - flexible

- Disadvantages:

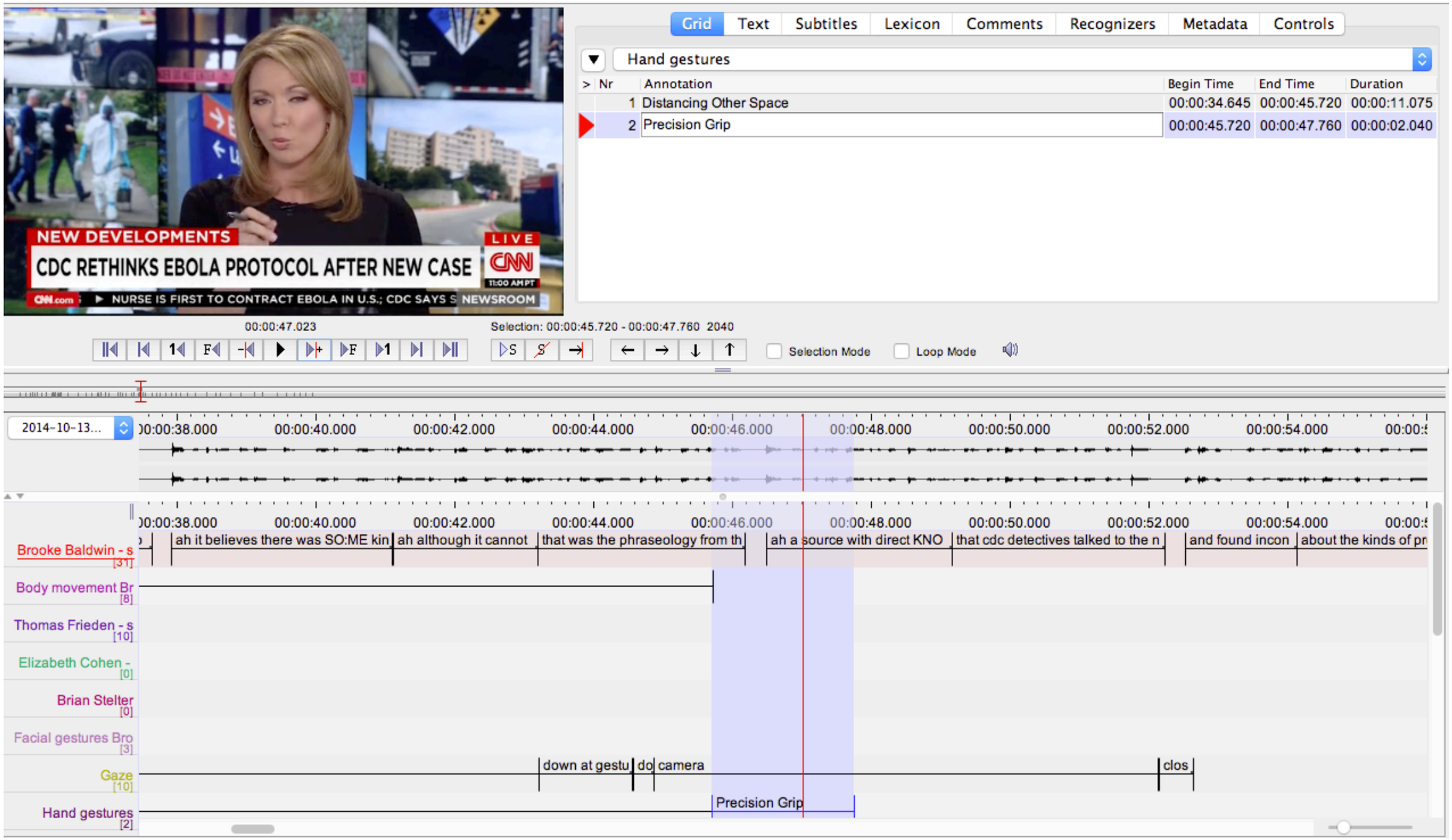
  - resourceful (multiple individuals)

  - error prone (low reliability)

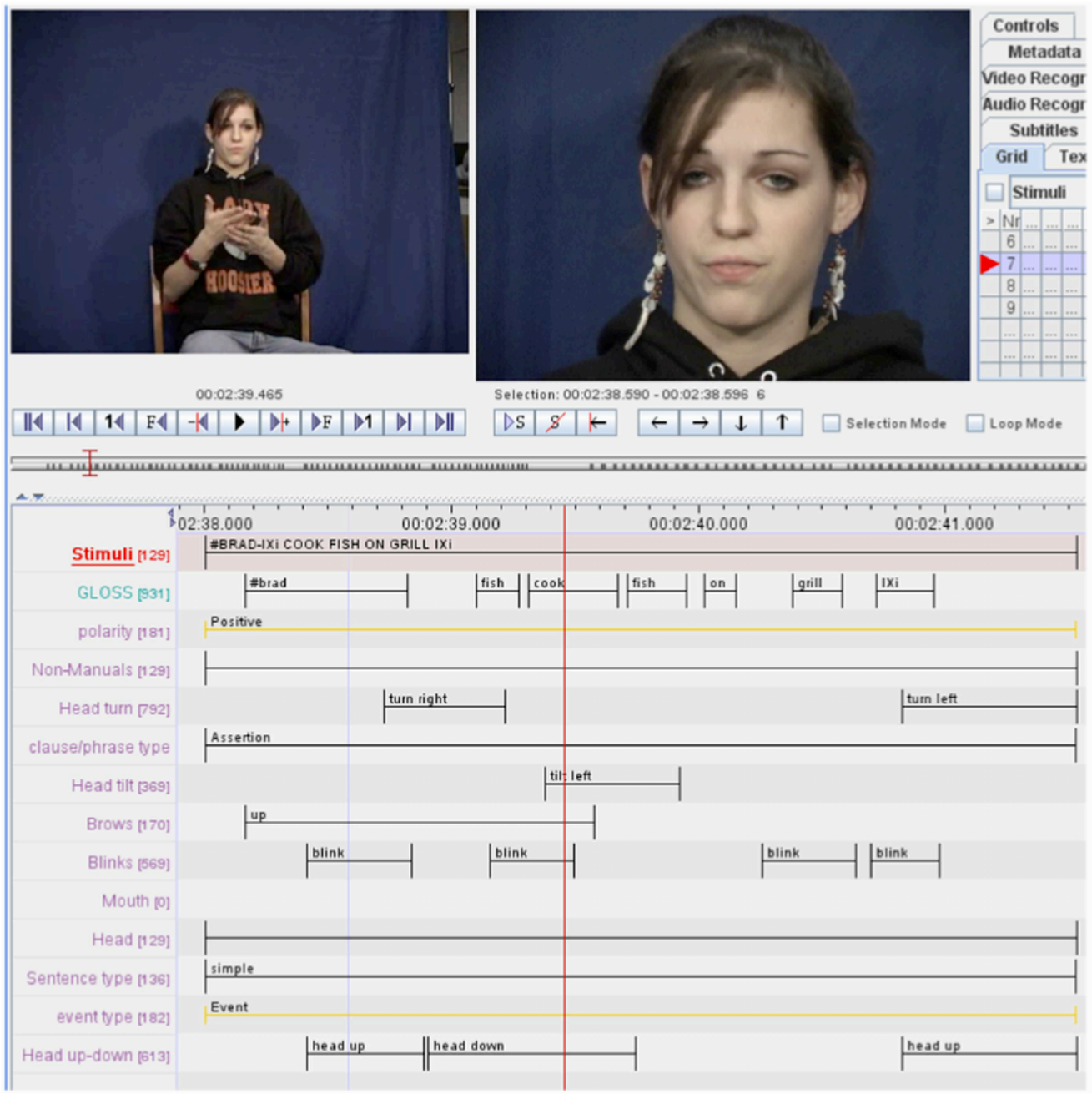




Manual coding (2)



Manual coding (3)

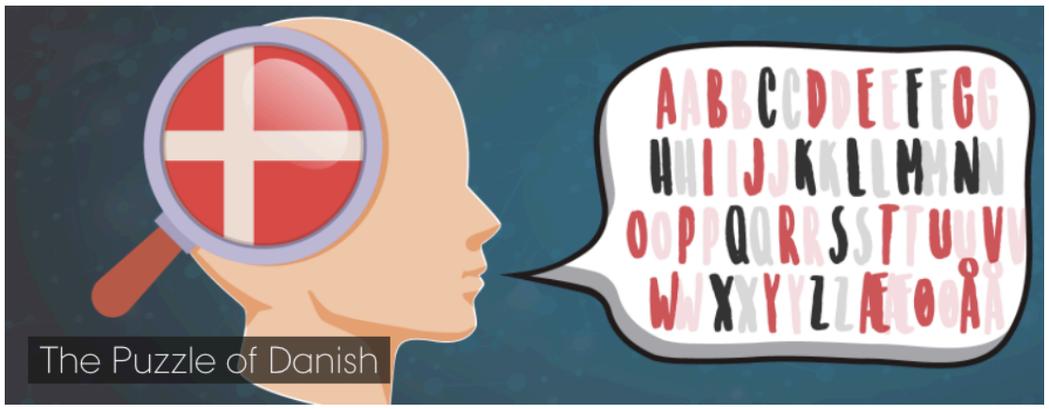


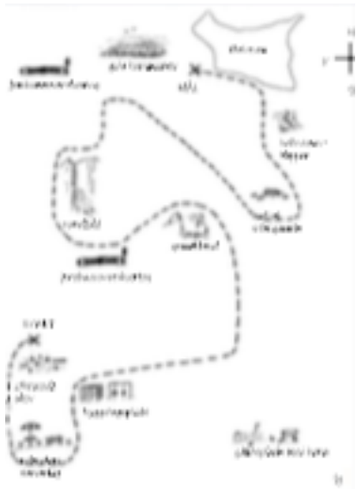
Manual coding (4)

- Systematic process:
  - define a detailed coding scheme
  - recruite one/multiple independent and naive coders
  - train the coders
- Compare the independent codings for how reliable they are (do they select the same categories?
- If reliability is ok, then analyze the data; if not, start again

Example: Analyzing dialogue data (1)

- Task-oriented conversation vs spontaneous conversation
- RQ: How much common ground (backchannel, alignment)?





Example: Analyzing dialogue data (2)

- A coding scheme needs to be generated as part of the variable identification and measurement process

- Quality of coding scheme determines quality of data

- Quality must be assessed through piloting

- If low quality, then revise the coding scheme and pilot again

General instructions for transcription:

The study aims at describing some of the linguistic properties of Danish as compared to Norwegian. We collected 45 minutes of interaction between two interlocutors in 2 different conditions: spontaneous interaction and task-oriented interaction.
The files are marked with two numbers, e.g. 7_2. The first number (in this case "7") is the pair number, and the second number (here "2") is the session number. Use these numbers to name your transcription-document.

In order to be able to analyse the data we need:

1. An **orthographic transcription** of the interactions. By orthographic we mean that we are interested in writing the single words as if they were written in correct Danish/Norwegian, no matter the actual pronunciation. The transcription should however reflect the syntactic structure used in the recordings. Do not add words that are not said in the recording, nor if it is correct according to dansk/norsk retskrivning/spelling. In case of doubt make a note and consult Christina. If you cannot hear an utterance, a word, or a non-verbal vocalization well enough to transcribe it meaningfully, just write xxx.
If a word is cut of before the speaker finishes it, just write what you hear and add a hyphen when the word ends. E.g.

00:03:57.0 A: Har du egentlig prøvet at bage de der cooki- kokoskugler? 00:04:01.3

Note that it is important to not add words that are not said, even if this introduces syntactic errors. For example: If the speaker says "Det så fin"; it should be transcribed as
A: Det så fint
"Fint" is changed to the correct form, but the verb (er) is not added.

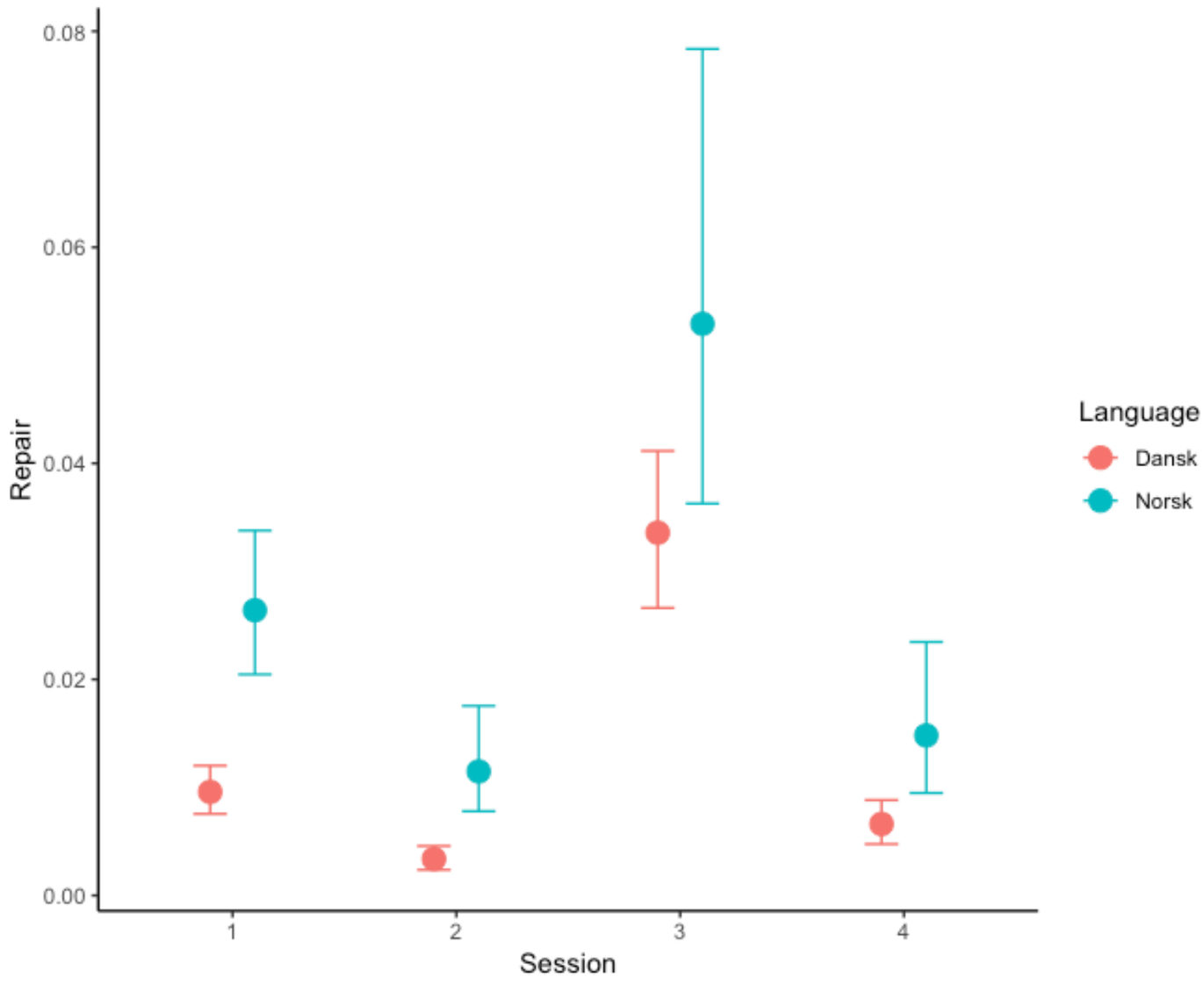
Example: Analyzing dialogue data (3)

- Begin the coding/transcription process

- In this case, dialogues were both transcribed and coded for occurrences of variables of interest (backchannel, repair, suggestions, etc.)

Example: Analyzing dialogue data (4)

- Once coded, data can be analyzed quatitatively/numerically



Dideriksen et al., in prep.

Example: Analyzing social eye-gaze cues (1)

Contents lists available at ScienceDirect

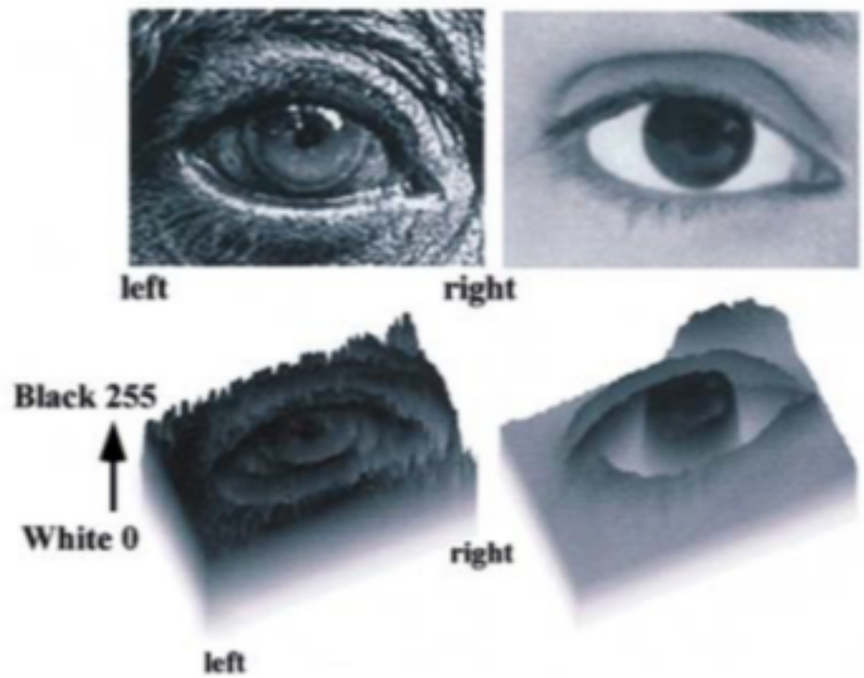
Language & Communication

journal homepage: www.elsevier.com/locate/langcom

Bodily constraints contributing to multimodal referentiality in humans: The contribution of a de-pigmented sclera to proto-declaratives

Juan Olvido Perea García*, Katrine Rosendal Ehlers, Kristian Tylén

Aarhus Universitet, Nordre Ringgade 1, 8000 Aarhus C, Denmark



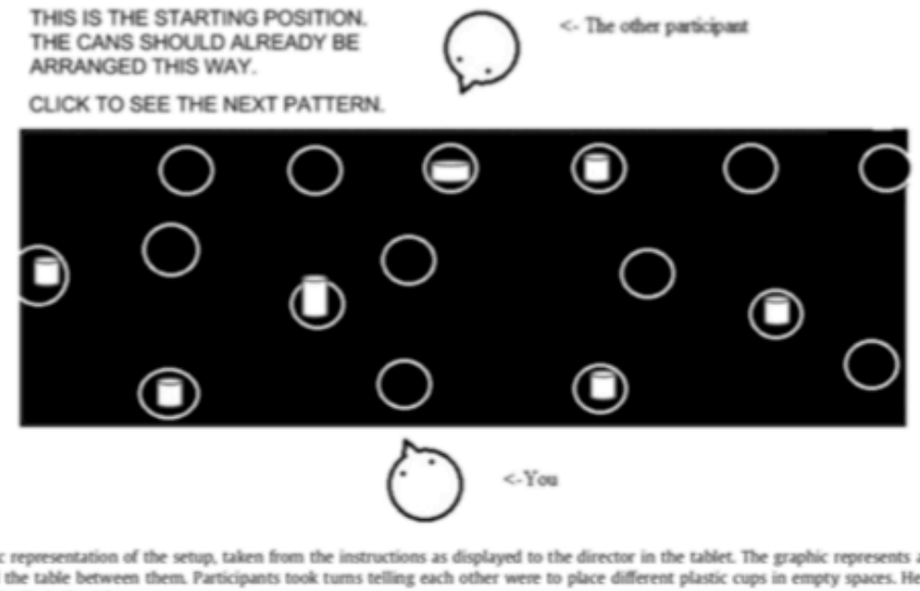


Fig. 1. Schematic representation of the setup, taken from the instructions as displayed to the director in the tablet. The graphic represents a bird-view of the participants and the table between them. Participants took turns telling each other were to place different plastic cups in empty spaces. Here translated into English, but originally in Danish.

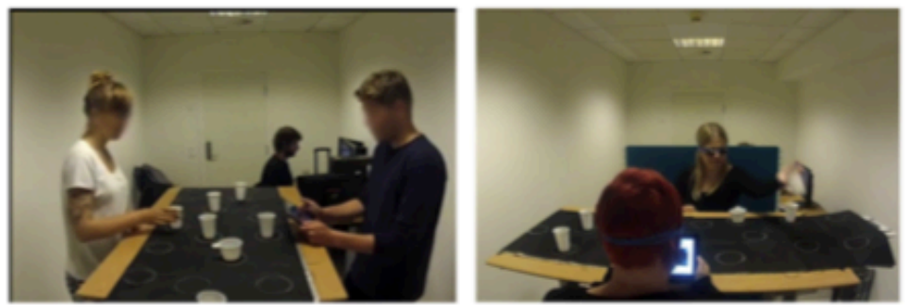


Fig. 2. Screenshots from the footage in both conditions ("gaze" to the left; "no-gaze" to the right). Error bars express 95% confidence intervals.

Example: Analyzing social eye-gaze cues (2)

- H1-a: Lower accuracy if no gaze-cues
- H1-b: Higher freqency of conversational repair if no gaze-cues
- H1-c: Participants will compensate for lack of gaze-cues by relying more on verbal strategies

Coding scheme excerpt

Categories of interest	Definition	Example
Repair	Any expression that give rise to subsequent repetition, clarification, or reformulation	"To which side again?"
Visual-verbal	Multi-modal reference combining a deictic expression (this/that one, here/there) with a non-verbal expression	"that one over there"
Verbal only	Instruction depended on speech alone	"take the one closest to the windows"

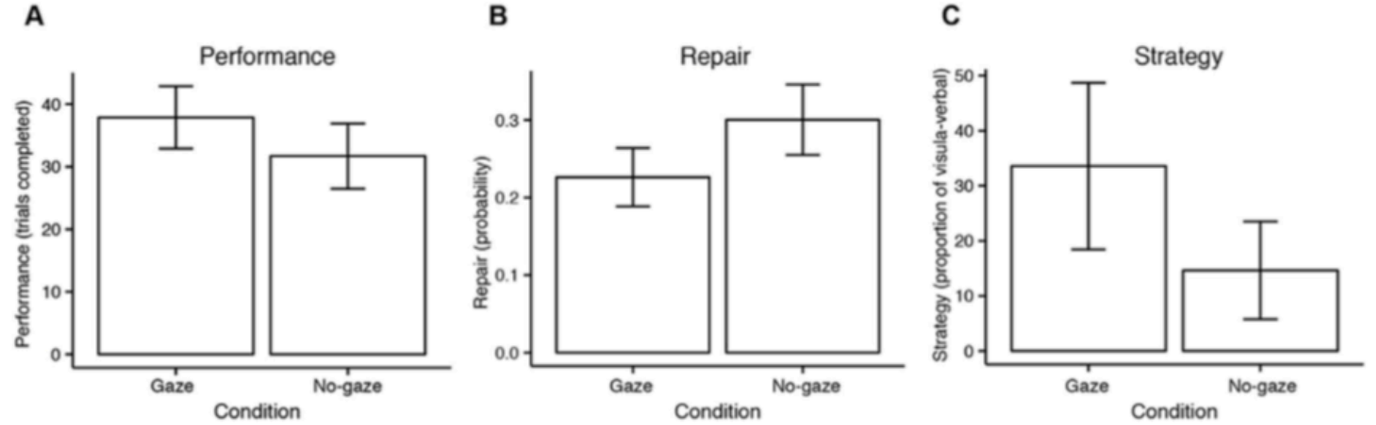
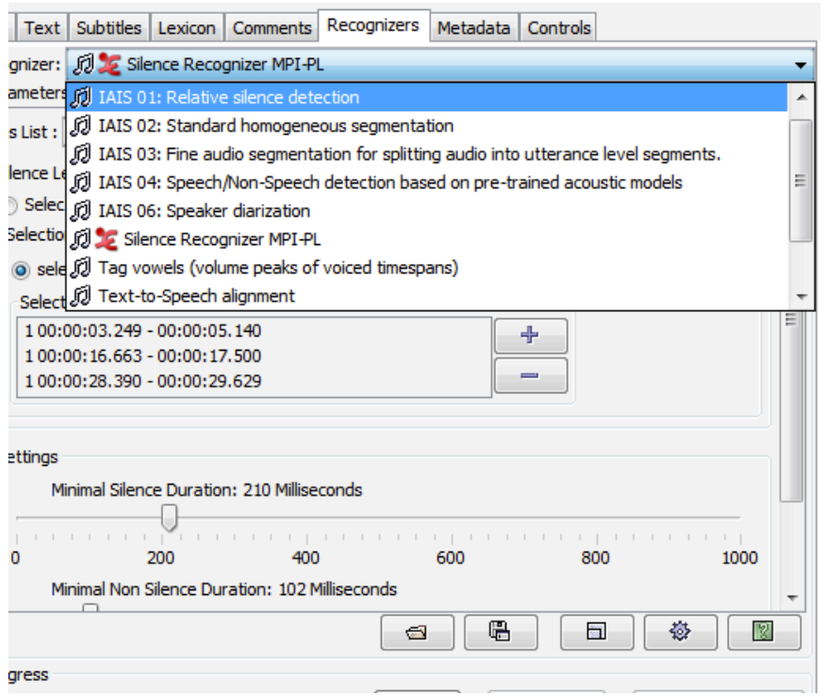
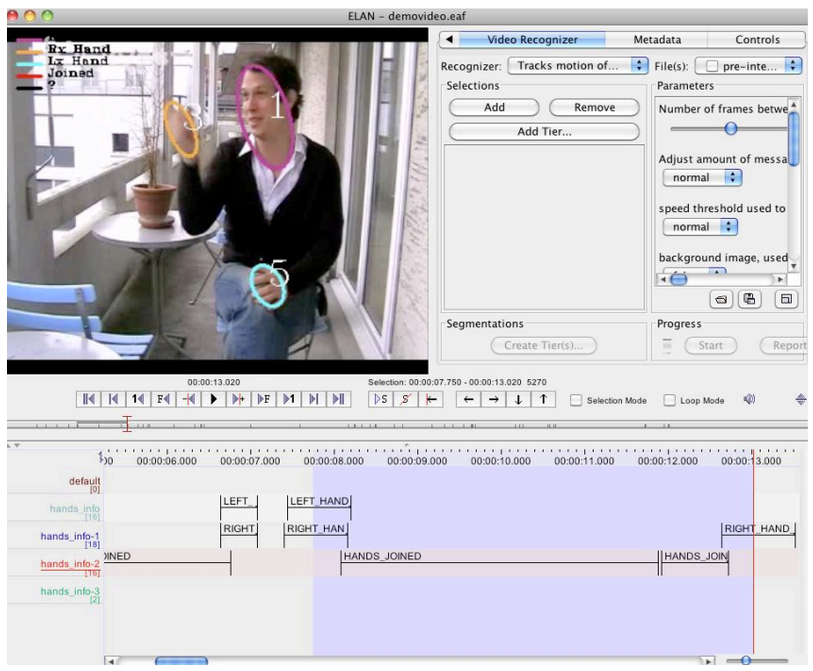


Fig. 3. Results. Effect of the two conditions, that is the "gaze" group (who had unrestricted access to gaze-cues), and the "no-gaze" group (whose access to gaze-cues was obstructed by goggles) on the three dependent variables. A: The relation between condition and performance. B: The relation between condition and the probability of experiencing episodes of conversational repair. C: The relation between condition and relative frequency of strategies (the proportion of visual–verbal relative to verbal-only).

Automated computer coding

- A/V processing
- Corpus methods:
  - define a finite set of key variables of interest
  - run automatic search/count in test corpus
- Advantages:
  - objective, unbiased, efficient
- Disadvantages:
  - limited in scope, unflexible, bad at overlapping categories





Inter-rater reliability (IRR)

- A measure of the degree to which independent coders agree in categorizing the data
- High reliability:
  - → Analyze the data
- Low reliability:
  - Bad coding scheme (target categories not sufficiently well defined?)
  - One or more raters are not sufficiently trained/competent
  - → Discard data, revise coding scheme, re-train raters
- How do we quantify inter-rater reliability?

Percent agreement (simple agreement)

- Proportion of agreement of coded units between two independent coders
- Best suited for categorical variables
- For continuous variables, use Pearson's correlation

Var#	Raters		Difference
	Mark	Susan	
1	1	1	0
2	1	0	1
3	1	1	0
4	0	1	–1
5	1	1	0
6	0	0	0
7	1	1	0
8	1	1	0
9	0	0	0
10	1	1	0
Number of Zeros			8
Number of Items			10
Percent Agreement			80

	Pearson Correlations	
Scoring Procedure	Intra-Rater	Inter-Rater
Doyon	0.81	0.91
Shulman	0.9	0.83
Tuokko	0.97	0.99
Watson	0.84	0.98
Wolf-Klein	0.96	0.81

Cohen's kappa (κ) (1)

- Better estimates percent agreement by taking into account the agreement that occurs by chance
- $$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$$

Cohen's kappa (κ) (1)

- Better estimates percent agreement by taking into account the agreement that occurs by chance
- $$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$$

Observed agreement between raters

Expected agreement between raters

Cohen's kappa (κ) (2)

- $\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$

- $p_o = \frac{a + d}{a + b + c + d} = \frac{10 + 5}{10 + 15 + 20 + 5} = 0.3$

		B	
		Yes	No
A	Yes	a 10	b 15
	No	c 20	d 05

- $p_e = p_{Yes} + p_{No}$, where:
  - $p_{Yes} = (\frac{a + b}{a + b + c + d}) \times (\frac{a + c}{a + b + c + d}) = (\frac{25}{50}) \times (\frac{30}{50}) = 0.3$
  - $p_{No} = (\frac{c + d}{a + b + c + d}) \times (\frac{b + d}{a + b + c + d}) = (\frac{25}{50}) \times (\frac{20}{50}) = 0.2$
- $p_e = 0.3 + 0.2 = 0.5$

- $\kappa = \frac{0.3 - 0.5}{1 - 0.5} = 1 - \frac{1 - 0.3}{1 - 0.3} = -0.4$

Cohen's kappa (κ) (2)

- $\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$

The ratings on which the two raters agree

- $p_o = \frac{a + d}{a + b + c + d} = \frac{10 + 5}{10 + 15 + 20 + 5} = 0.3$

All ratings

		B	
		Yes	No
A	Yes	a 10	b 15
	No	c 20	d 05

- $p_e = p_{Yes} + p_{No}$, where:
  - $p_{Yes} = (\frac{a + b}{a + b + c + d}) \times (\frac{a + c}{a + b + c + d}) = (\frac{25}{50}) \times (\frac{30}{50}) = 0.3$
  - $p_{No} = (\frac{c + d}{a + b + c + d}) \times (\frac{b + d}{a + b + c + d}) = (\frac{25}{50}) \times (\frac{20}{50}) = 0.2$
- $p_e = 0.3 + 0.2 = 0.5$

- $\kappa = \frac{0.3 - 0.5}{1 - 0.5} = 1 - \frac{1 - 0.3}{1 - 0.3} = -0.4$

Cohen's kappa (κ) (2)

- $\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$

The ratings on which the two raters agree

- $p_o = \frac{a + d}{a + b + c + d} = \frac{10 + 5}{10 + 15 + 20 + 5} = 0.3$

All ratings

- $p_e = p_{Yes} + p_{No}$, where:
  - $p_{Yes} = (\frac{a + b}{a + b + c + d}) \times (\frac{a + c}{a + b + c + d}) = (\frac{25}{50}) \times (\frac{30}{50}) = 0.3$
  - $p_{No} = (\frac{c + d}{a + b + c + d}) \times (\frac{b + d}{a + b + c + d}) = (\frac{25}{50}) \times (\frac{20}{50}) = 0.2$

Times rater A codes Yes

Times rater B codes Yes

Times rater B codes No

Times rater A codes No

- $p_e = 0.3 + 0.2 = 0.5$

- $\kappa = \frac{0.3 - 0.5}{1 - 0.5} = 1 - \frac{1 - 0.3}{1 - 0.3} = -0.4$

		B	
		Yes	No
A	Yes	a: 10	b: 15
	No	c: 20	d: 05

Cohen's kappa (κ) (3)

- $\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$

- $p_o = \frac{a + d}{a + b + c + d} = \frac{10 + 5}{10 + 15 + 20 + 5} = 0.3$

- $p_e = p_{Yes} + p_{No}$, where:
  - $p_{Yes} = (\frac{a + b}{a + b + c + d}) \times (\frac{a + c}{a + b + c + d}) = (\frac{25}{50}) \times (\frac{30}{50}) = 0.3$
  - $p_{No} = (\frac{c + d}{a + b + c + d}) \times (\frac{b + d}{a + b + c + d}) = (\frac{25}{50}) \times (\frac{20}{50}) = 0.2$
- $p_e = 0.3 + 0.2 = 0.5$

- $\kappa = \frac{0.3 - 0.5}{1 - 0.5} = 1 - \frac{1 - 0.3}{1 - 0.3} = -0.4$

		B	
		Yes	No
A	Yes	a 10	b 15
	No	c 20	d 05

Value of Kappa	Level of Agreement	% of Data that are Reliable
0-.20	None	0-4%
.21-.39	Minimal	4-15%
.40-.59	Weak	15-35%
.60-.79	Moderate	35-63%
.80-.90	Strong	64-81%
Above .90	Almost Perfect	82-100%

Cohen's kappa (κ) (4)

- Negative value of κ?
  - Number of agreements observed is fewer than what would be observed by chance

Value of Kappa	Level of Agreement	% of Data that are Reliable
0-.20	None	0-4%
.21-.39	Minimal	4-15%
.40-.59	Weak	15-35%
.60-.79	Moderate	35-63%
.80-.90	Strong	64-81%
Above .90	Almost Perfect	82-100%

- $\kappa = -0.4$:
  - weak agreement between raters below what would be expected by change

Other metrices

- Scott's $\boldsymbol{\pi}$

  equivalent to Cohen's κ, calculates p_e using squared arithmetics means of marginal proportions instead of squared geometric means

- Fleiss' κ

  generalization of Scott's $\boldsymbol{\pi}$ for more than two raters

- Krippendorff's α

  applicable to any number of coders, missing data, any levels of measurement, and adjusts itself to small sample sizes

Summary

- Some aspects of human cognition can’t be quantified directly
- = Not all experiments produce numeric outputs
- Qualitative data → Coding → Quantitative data:
  - human coding
  - computer coding
- Overcome subjective bias in coding by:
  - having 2+ independent coders that are naive to the task
  - statistically measuring their agreement

Thursday

- We will practice how to turn a qualitative material into measurable units via coding

- We will learn how to assess interrater reliability in R

- Jonathan will make some files available for the exercise prior to class