AiNed XS: Audio classification

Marijn Schraagen

V&L Lab meeting

January 29th, 2025

Universiteit Utrecht

Low-literacy project

- Understanding low-literacy: what happens while reading?
- Low-literate people: high cognitive load and **stress**
- Experiment: measure reading competency and stress while reading
- Goals:
  - Get insights into the process
  - Build machine learning models to predict reading performance and stress
  - Create tools to make reading experience better and improve reading level

Items

- Variables: easy or difficult text, meaning valid or not
- Age of Acquisition: at what age do children learn a word on average
- 40 words with AoA ≥ 11 years (difficult) and 40 words with AoA ≤ 5 years (easy) from existing AoA dataset
- Find real sentences at OpenSoNaR with these words

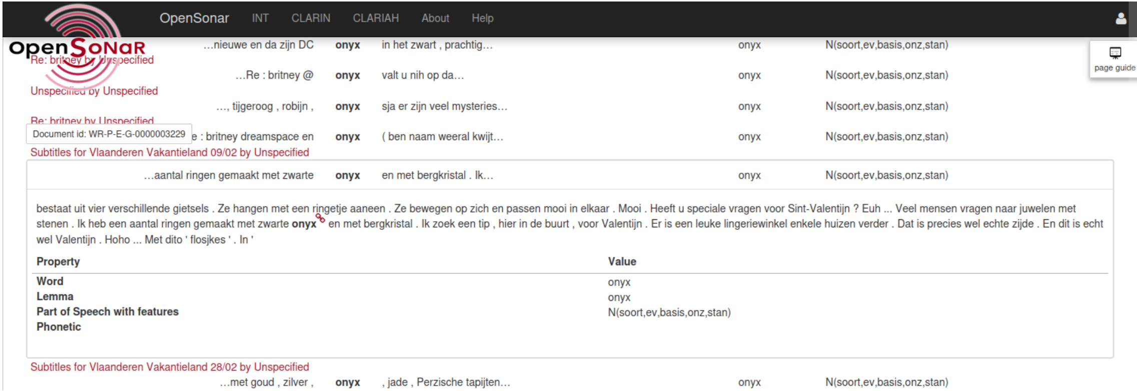
difficulty (AoA)	*word*	*validity*	*sentence*
difficult (15.7)	onyx	valid	Ik heb een aantal ringen gemaakt met zwarte onyx.
difficult (14.7)	atol	invalid	Het atol zelf koken en opeten is geen risico.
easy (4.4)	ijsje	valid	Je mag een keer per week een ijsje eten.
easy (4.5)	ruzie	invalid	Eind mei ging het toen de verkeerde ruzie op.

Items

- Variables: easy or difficult text, meaning valid or not
- Age of Acquisition: at what age do children learn a word on average
- 40 words with AoA ≥ 11 years (difficult) and 40 words with AoA ≤ 5 years (easy) from existing AoA dataset
- Find real sentences at OpenSoNaR with these words

difficulty (AoA)	*word*	*validity*	*sentence*
difficult (15.7)	onyx	valid	I have made a number of rings with black onyx.
difficult (14.7)	atol	invalid	To cook and eat the atol yourself is no risk.
easy (4.4)	ice cream	valid	You can eat ice cream once per week.
easy (4.5)	fight	invalid	At the end of May it went in the wrong fight.

Items



…nieuwe en da zijn DC	**onyx**	in het zwart , prachtig…	onyx	N(soort,ev,basis,onz,stan)

Re: britney by Unspecified

…Re : britney @	**onyx**	valt u nih op da…	onyx	N(soort,ev,basis,onz,stan)

Unspecified by Unspecified

…, tijgeroog , robijn ,	**onyx**	sja er zijn veel mysteries…	onyx	N(soort,ev,basis,onz,stan)

Re: britney by Unspecified

Document id: WR-P-E-G-0000003229

…e : britney dreamspace en	**onyx**	(ben naam weeral kwijt…	onyx	N(soort,ev,basis,onz,stan)

Subtitles for Vlaanderen Vakantieland 09/02 by Unspecified

…aantal ringen gemaakt met zwarte	**onyx**	en met bergkristal . Ik…	onyx	N(soort,ev,basis,onz,stan)

bestaat uit vier verschillende gietsels . Ze hangen met een ringetje aaneen . Ze bewegen op zich en passen mooi in elkaar . Mooi . Heeft u speciale vragen voor Sint-Valentijn ? Euh ... Veel mensen vragen naar juwelen met stenen . Ik heb een aantal ringen gemaakt met zwarte **onyx** en met bergkristal . Ik zoek een tip , hier in de buurt , voor Valentijn . Er is een leuke lingeriewinkel enkele huizen verder . Dat is precies wel echte zijde . En dit is echt wel Valentijn . Hoho ... Met dito ' flosjkes ' . In '

Property	Value
Word	onyx
Lemma	onyx
Part of Speech with features	N(soort,ev,basis,onz,stan)
Phonetic	

Subtitles for Vlaanderen Vakantieland 28/02 by Unspecified

…met goud , zilver ,	**onyx**	, jade , Perzische tapijten…	onyx	N(soort,ev,basis,onz,stan)

Procedure

1. Demographic questionnaire
2. Set up physiological measurements
    - Galvanic Skin Response (GSR) and photoplethysmograph (Pleth)
3. Reading task
4. Reading level test (online)
5. Interview

Welcome

Thank you for participating in this research.

An image will be shown now. Sit, relax and watch the image.

Click on the arrow up or down to continue.

Universiteit Utrecht

Neutral stimulus (2 minutes)

Instructions

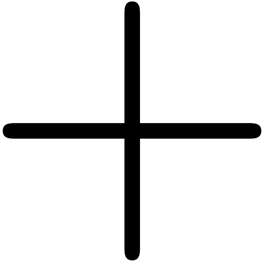
You will see a sentence.

Click on the left arrow if you think the sentence is valid.
Click on the right arrow if you think the sentence is invalid.

Then read the sentence out loud.

After reading, click on the up or down arrow. Then you continue.
Now click on the up or down arrow to practise.

Fixation (1.5sec)



Trial: choice task (50% of participants)

Het aantal spanten op de kiel groeide.

The number of trusses on the keel grew
Omurgadaki kiriş sayısı arttı
Il numero di capriate sulla chiglia è aumentato
Число ферм на киле выросло
龍骨上面嘅桁架數量增加咗

✓ Goed

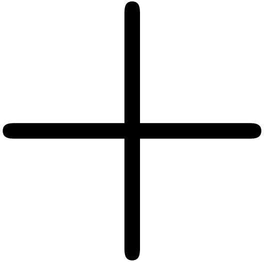
✗ Fout

Trial: reading task (all participants)

Het aantal spanten op de kiel groeide.

The number of trusses on the keel grew
Omurgadaki kiriş sayısı arttı
Il numero di capriate sulla chiglia è aumentato
Число ферм на киле выросло
龍骨上面嘅桁架數量增加咗

Fixation (2.0sec)



Trial

More practise sentences, six in total

Instructions repeated

Same instructions again

Trial

80 trials, randomized order

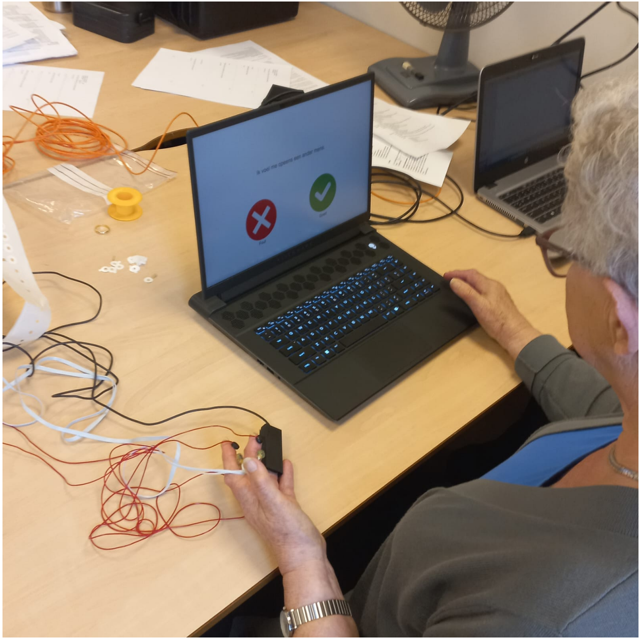
End

This is the end of the experiment.

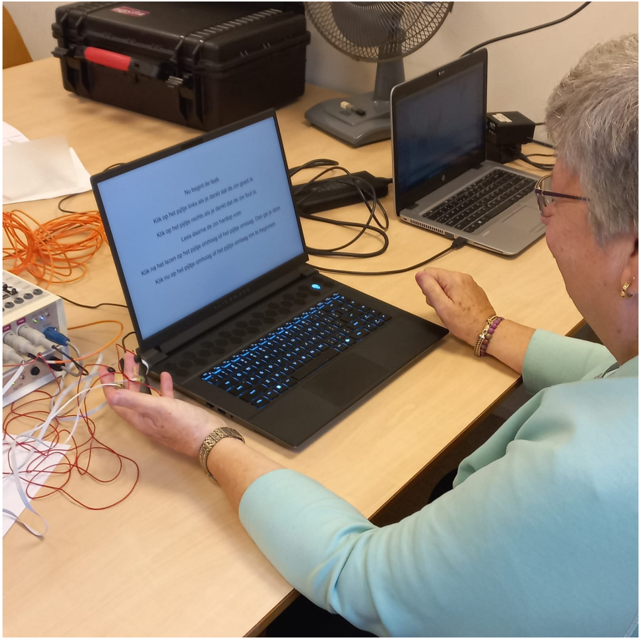
Thank you for participating.

Click on the up or down arrow to finish.

Impression of the experiment



Impression of the experiment



Nu begint de taak

Klik op het pijltje links als je denkt dat de zin goed is.

Klik op het pijltje rechts als je denkt dat de zin fout is.

Lees daarna de zin hardop voor

Klik na het lezen op het pijltje omhoog of het pijltje omlaag. Dan ga je door

Klik nu op het pijltje omhoog of het pijltje omlaag om te beginnen

Result data

- 110 participants, 55 low-literate and 55 high-literate
- Choice task data and response times (50% of participants)
- **Microphone recordings for reading task**
- Reading times
- GSR and Pleth
- Interview data
- Language proficiency test results
- Automatic transcriptions of audio files by ASR
- **Machine learning prediction results**

Variables to classify on

- Session level
    - Participant low or high literate
    - Participant native or non-native (low literate)
    - Choice task or not (50% each)
    - Regular or counterbalanced button order
- Item level
    - Sentence easy or difficult
    - Sentence valid or invalid
    - Correct button pressed or not

Variables to classify on

- Participant low or high literate
- Train a classifier that given the audio signal predicts if the participant is high or low literate
- Content-independent: high and low literate participants read the same sentences

Variables to classify on

- Participant native or non-native
- Train a classifier that given the audio signal predicts if the participant is native or non-native
- Content-independent: native and non-native participants read the same sentences

Variables to classify on

- Choice task or not
- Train a classifier that given the audio signal predicts if the participant performed the choice task or not
- Reminder choice task:
  1. Participant sees the sentence
  2. Does not read the sentence out loud yet
  3. Participant presses button “valid” or “invalid”
  4. Buttons disappear
  5. Participant reads sentence out loud, which is recorded with the microphone
  6. Participant reads sentence twice: first silent reading to make the choice, the again to make the recording
- Idea: second reading more confident, classifier may be able to pick up on this
- Content-independent: choice task and non-choice task participants read the same sentences

Variables to classify on

- Sentence easy or difficult, sentence valid or invalid
- Train a classifier that given the audio signal predicts if the sentence as spoken by this participant is easy/difficult resp. valid/invalid
- **Content-dependent**: classifier could learn the contents of the sentences from training data instead of taking participant information into account
- Partial solution: make sure training and test sentences do not overlap

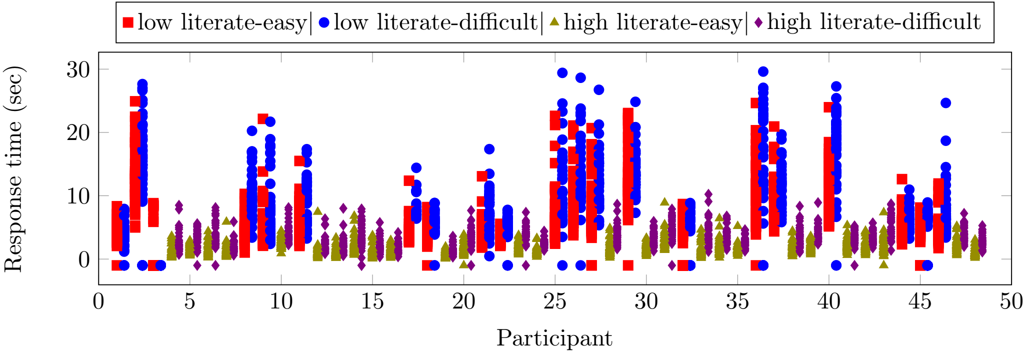
Variables to classify on

- Correct button pressed or not
- Train a classifier that given the audio signal predicts if the participant pressed the correct button
- Partly content-independent: correct and incorrect button is a participant characteristic
- However, correlation between easy/difficult, valid/invalid, and correctness of button press

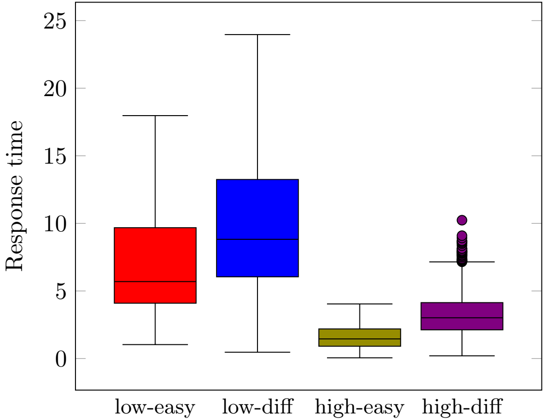
Stimulus validity

- Assumption: experimental variables are capable of showing differences between participants
- If response times and accuracy are the same, then participants are **not** influenced by the variables
- Then an audio classifier is less likely to separate groups
- Validity check: statistical tests for response times and accuracy

Choice task response time: easy vs. difficult

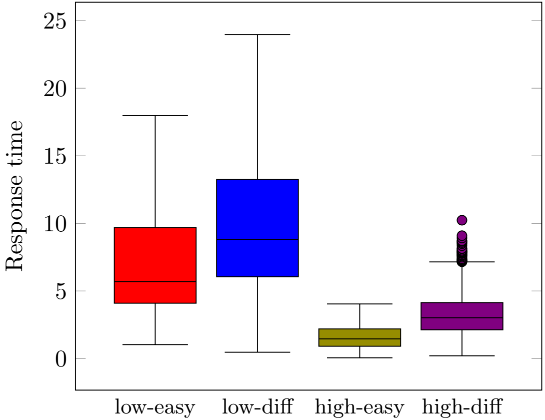


Choice task response time: easy vs difficult



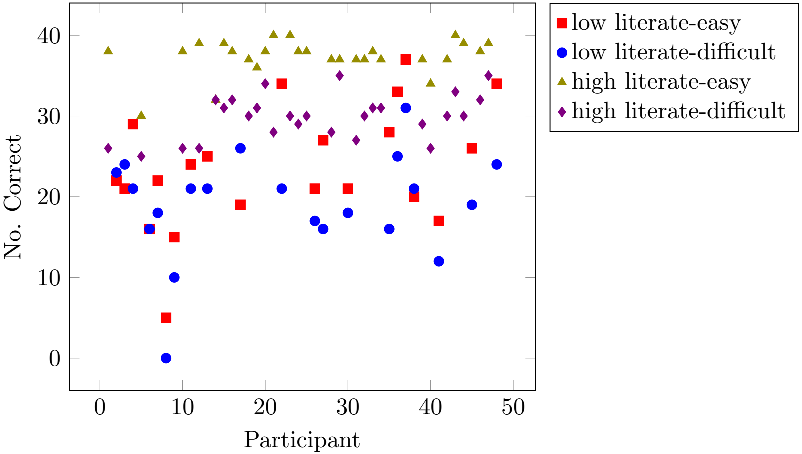
- Difference of response time between low and high literate: t-value=51.4, **p=0.0**
- Difference of response time between easy and difficult: t-value=-13, **p=2e-40**
- Low literate, RT difference between easy and difficult: t-value=-11, **p=8e-28**
- High literate, RT difference between easy and difficult: t-value=-26, **p=2e-136**

Choice task response time: easy vs difficult



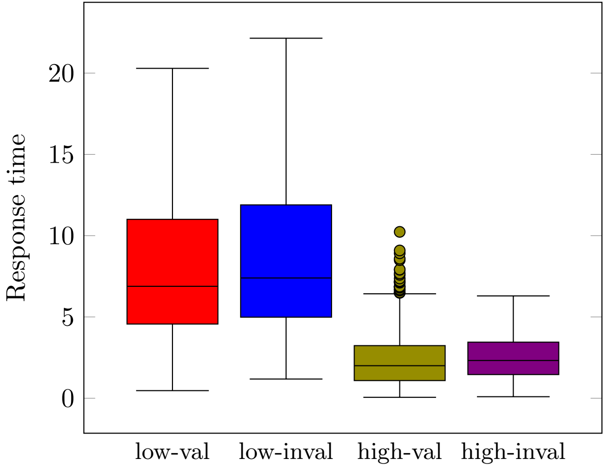
- Difference of response time between low and high literate: t-value=51.4, **p=0.0**
- **Note:** Difficult sentences are longer (61 characters, 10.9 words) than easy sentences (40 characters, 7.9 words)
- Difference in *response time per word* between easy and difficult: t-value=-2.9, **p=0.003**

Choice task correct



- Difference of correct answers between easy and difficult: t-value=3.8, **p=0.0002**
- Low literate only, difference between easy and difficult: t-value=2.1, p=0.04
- High literate only, difference between easy and difficult: t-value=10.8, **p=5e-15**
- Difference no. correct between low and high literate: t-value=-9.2, **p=5e-12**

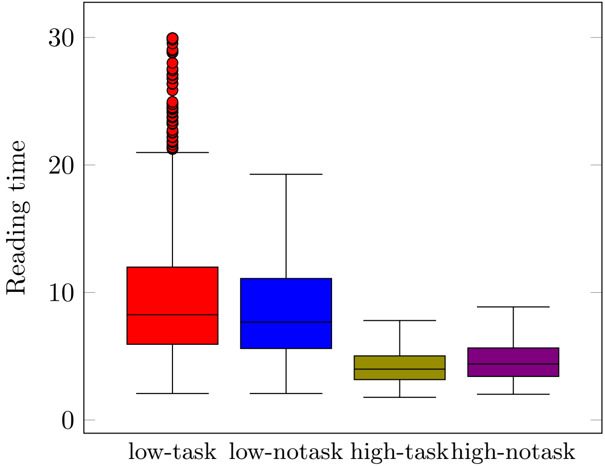
Choice task response time: valid vs. invalid



- Difference of response time between valid and invalid: t-value=-2.7, **p=0.007**
- Low literate, RT difference between valid and invalid: t-value=-2.3, p=0.022
- High literate, RT difference between valid and invalid: t-value=-4.5, **p=5e-5**

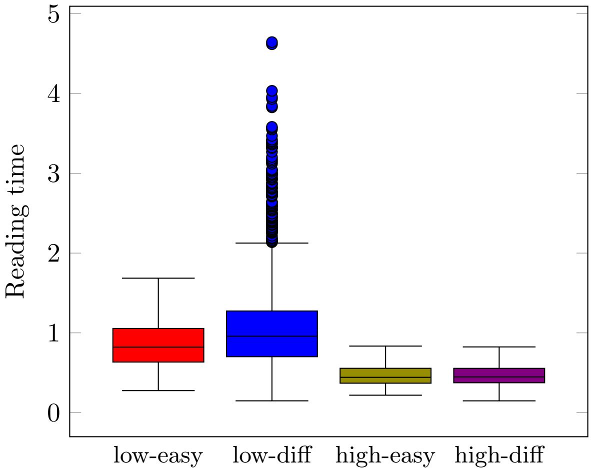
  Note: valid and invalid sentences are pairwise the same length

Reading time: choice task or not



- Difference of reading time between low and high literate: t-value=-55, **p=0.0**
- Difference of reading time with or without choice task: t-value=-0.45, p=0.65
- Low literate only, reading time with or without task: t-value=1.1, p=0.27
- High literate only, reading time with or without task: t-value=-6.8, **p=1e-11**

Reading time (per word): easy or difficult



- Difference reading time per word, easy vs. difficult: t-value=-6.7, **p=1e-11**
- Low literate time per word, easy vs. difficult: t-value=-8.9, **p=6e-19**
- High literate time per word, easy vs. difficult: t-value = 0.6, p=0.56

Validity of stimuli

- Almost all RT and accuracy differences significant
- Conclusion: stimuli are representative for this task
- Exceptions:
  - High literate people read with the same speed (words per minute) for easy or difficult sentences
  - Low literate people read the same speed with or without the button task
  - Low literate people have the same response time for choosing validity of sentences
  - Low literate people make the same amount of validity mistakes for easy and difficult sentences
- All of these make sense

Classifier setup

- Supervised classification
- WAv2Vec2 audio transformer pretrained on Dutch
- 80-10-10 train-validation-test split
- 10 epochs
- batch size: 4
- decreasing learning rate: 0.001–0.00001

Native vs. non-native

- Given a recording, predict if the participant is low-literate or high-literate
- 50% of participants is low-literate
- Test Low-literate Class Accuracy: 83.97%
  Test High-literate Class Accuracy: 86.48%
  Test Low-literate Non-native Accuracy: 82.27%
  Test Low-literate Native Accuracy: 87.16%
  Overall Test Accuracy: 86.15%

Native vs. non-native

- Given a recording, predict if the participant is native or non-native
- 37% of participants is non-native
- 74% of low-literate participants
- Test Native Class Accuracy: 88.37%
  Test Non-Native Class Accuracy: 81.46%
  Overall Test Accuracy: 85.82%

Correct answer

- Given a recording, predict if the participant gave the correct answer on the button task
- 67% of answers were correct
- Test Incorrectly Pressed Button Class Accuracy: 20.90%
  Test Correctly Pressed Button Class Accuracy: 88.49%
  Overall Test Accuracy: 66.50%
- Note: does not improve on baseline of always choosing ‘correct’
- But: classifier still learns something to reach 21%

Choice task or not

TODO

but initial results seem to indicate some performance

Sentence easy or difficult

TODO

in combination with sentence-level split to avoid training on contents

Sentence valid or invalid

TODO

Data cleaning

- Trim audio files to 15sec or 5sec
  - Goal: reduce amount of padding
  - Possible confound: longer recordings → low literate

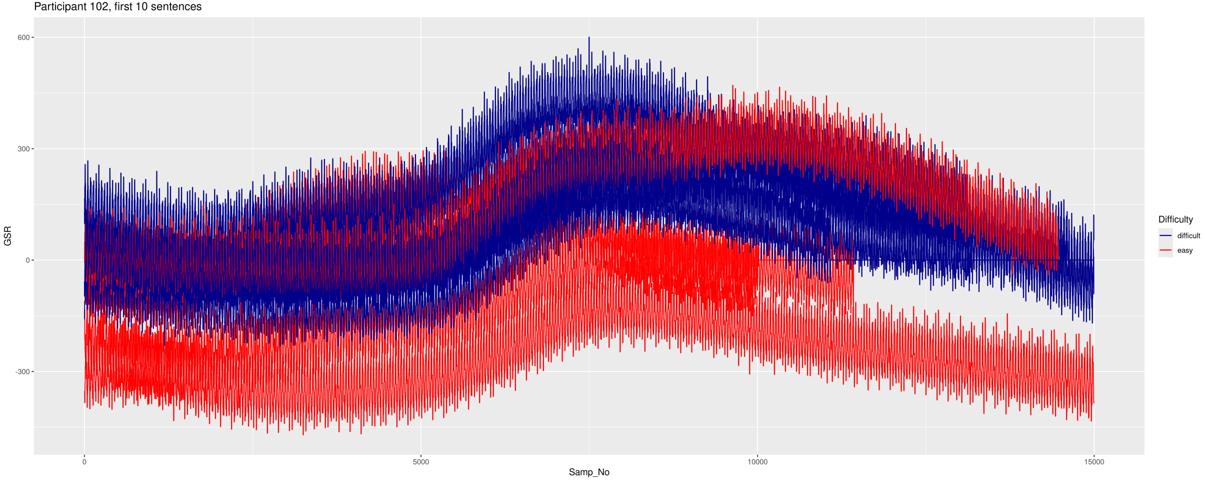
Overview of audio classification results

classifier	class 0	class 1
low vs. high literacy	0.84	0.86
native vs. non-native	0.88	0.81
correct vs. incorrect answer	0.88	0.21
...		

Other results and analysis

- Analysis of Word Error Rate with ASR
  - Reference transcriptions needed, otherwise an error could be either a pronunciation error or an ASR error
- Relation of physiological measures with reading difficulty and literacy
- Qualitative analysis of post-experiment interviews
- Analysis of demographic questionnaire and language proficiency test

Galvanic Skin Response analysis


Participant 102, first 10 sentences
GSR
Samp_No
Difficulty
difficult
easy