Documentaion

Participants: demographics (demographics.csv)

The data is collected for 32 German native speakers. The demographics includes:

- **num**: participant's unique id
- **gender**: participant gender (m - male; f or w - female)
- **age**: participant's age
- **dominant_eye**: participant's dominant eye (r - right; l - left)
- **list_nr**: experimental list that was given to a participant (ranges from 1-4)
- **handedness**: participant's dominant hand (r - right; l - left)
- **subjectID**: participant's ID incorporating the above information (excluding the handedness)

So, **subjectID** represents numGenderAgeDominant_eyeList_nr, e.g., 10m23r2 means that it is the 10th subject, male, 23 y.o., with the right dominant eye, who completed list 2 in the experiment.

However, there are 2 subjects with mistakes in **subjectID** encoding (it holds across all experiment-related files):

- 28m3r4: a correct age is 33
- 6m129r2: a correct age is 19

Literacy level (literacy_scores.csv)

Literacy level is measured using two tests (each test has its own score):

- Vocabulary test
  - Score (**VocabAcc**) = proportion of correctly answered questions
- Author recognition test
  - Score (**ARTscore**) = number of correctly answered questions *minus* number of incorrectly answered questions
  - Note that for some subjects, there was no ART score collected

Note: column **subjectID** is identical to subjectID from **demographics.csv**.

In both tests, higher scores mean higher literacy.

Vocabulary test is implemented as in *Van Os, M. (2023). Rational Speech Comprehension: Effects of Predictability and Background Noise. PhD thesis, Saarland University. doi: http://dx.doi.org/10.22028/D291-40555*

Author recognition test is implemented as in *Grolig, L., Tiffin-Richards, S.P. & Schroeder, S. Print exposure across the reading life span. Read Writ 33, 1423–1441 (2020). doi: https://doi.org/10.1007/s11145-019-10014-3*

Materials (materials.xlsx)

Critical words (*words* Sheet)

There were 12 real words chosen, and for each word (column **control word**); a pseudoword counterpart was generated (column **pseudoword**). The frequency for control words was taken from CELEX database[1] as

[1] *Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). The CELEX Lexical Database. Release 2 (CD-ROM). Linguistic Data Consortium.*

frequency per Mln. It was further transformed as follows: log10(CELEX frequency + 1). The resulting value is stored in column **frequency**.

Sentences and experimental lists(*List1-List4* Sheets)

Consequently, for each pair of control/pseudo words, there were generated 12 critical sentences. These critical sentences were distributed among 4 experimental lists (Sheets *List1-List4*).

Description of columns in Sheets *List1-List4*:

- ITEM: sentence ID (1-12 for critical sentences; 13-28 for filler sentences)
- CONDITION: either pseudo, control, or filler
  - pseudo: a sentence contains a pseudo word
  - control: a sentence contains a control word
  - filler (13-18): a sentence contains a low-frequency word
  - filler (19-28): no experimental manipulation
- SENTENCE: a sentence that subjects were presented with (plain text, no highlights)
- QUESTION: 1 - after reading a sentence, there was a YES/NO question; 0 - no question
- QUTEXT: a text of the question
- CORRECT: a correct answer
- WRONG: a wrong answer
- LIST: list number
- TYPE: item - sentence contains control/pseudo word manipulation; filler - no control/pseudo manipulation

Each list is generated such that there are 3 sentences with a pseudoword, and 9 sentences with a control word. Each sentence appears only once in the pseudo condition across the lists.

Each list also contained 16 filler sentences (ITEM 13-28). They were identical across the lists. Fillers with ITEM 13-18 contained a low frequency word (in red font) at approx. the same position where control/pseudo words occurred. Fillers with ITEM 19-28 did not contain any manipulation. Thus, each subject read in total 28 different sentences.

Note: Item 9 was excluded from the analysis due to a typo

Eye-tracking data (ia_seminar_project_tc.csv)

The eye-tracking measures were collected per word in each sentence for each subject.

The file contains the following columns:

- **RECORDING_SESSION_LABEL**: participant's ID (identical to **subjectID** in **demographics.csv**)
- **trial**: in which order a given sentence was presented to subject (range 1-27)
- **IA_ID**: a word ID in a given sentence
- **item**: item id that is identical to ITEM in **materials.xlsx**
- **IA_LABEL**: a word for which the measures are collected
- **wordlength**: a word's length
- **condition**: word condition
  - control - this is a control word
  - pseudo - this is a pseudoword
  - none - this is neither a control, nor a pseudo word, but the word belongs to a critical sentence
  - filler - this is a word from a filler sentence
- **is_critical**: a binary indicator for critical words
  - 1 for control/pseudo
  - 0 for none/filler
- **sentenceCondition**: sentence (aka item) condition
  - control: a sentence contains a control word

  - pseudo: a sentence contains a pseudoword
  - filler: a filler sentence
- **is_spill1**: a binary indicator if the word followed the critical word (0 - no; 1 - yes)
- **is_spill2**: a binary indicator if the word followed the critical word + 1 (0 - no; 1 - yes)
- **is_spill3**: a binary indicator if the word followed the critical word + 2 (0 - no; 1 - yes)
- **filler**: a binary indicator if the word is filler (0 - no; 1 - yes)
- **LF**: an indicator for low-frequency words in filler sentences (0 - no; 1 - yes)
- **HF**: an indicator for high-frequency words in filler sentences (0 - no; 1 - yes)
- **function_word**: an indicator for function words in filler sentences (0 - no; 1 - yes)
- **other_filler**: an indicator for other words in filler sentences (0 - no; 1 - yes)
- **composite**: a literacy composite score for a given subject (combines ARTscore and VocabAcc)

Eye-tracking measures (durations are in ms):

- **fixation_duration**: fixation duration (sum of all fixation durations within a word)
- **duration_firstpass**: first pass fixation duration (sum of all fixation durations that occurred in the first pass of the word, i.e., from the time when the word was first entered until it was left for the first time)
- **duration_firstfixation**: first fixation duration (duration of the first fixation that occurred within a word)
- **fix_count**: total number of fixations
- **avg_pupil**: average pupil size
- **IA_REGRESSION_IN_COUNT**: number of times the word was entered from the right
- **IA_REGRESSION_OUT_COUNT**: number of times the word was exited to the left before a left word was fixated in the trial
- **saccade_length**: length of the first saccade from the current word to a word on the right (in pixel)
- **saccade_duration**: a corresponding saccade duration
- **go_past_time**: go-past time (sum of fixation durations from the time the word was first entered until it was left to the right)

What pre-processing steps were taken (the currect dataset is already pre-cleaned)

The following fixations/trials/subjects were removed:

- fixations shorter than 120 ms, which occurred before or after a blink
- fixations not within an interest area (IA)
- trials where more than four fixations occurred outside the IA
- trials where the critical word was not fixated, were removed (similar for filler words that were selected for model training)
- all trials with item 9, as it had a typo

Author recognition test exclusion:

- subjects with more than 10% of timeouts on ART questions