Datasheet: An EEG dataset of word-level brain responses for semantic text relevance

Dataset version: 1.1.0

June 11, 2024

Contents

1 Motivation

- **For what purpose was the dataset created?** The dataset was created to enable research on predicting semantic text relevance from brain recordings – i.e., given brain recordings for a single word or a sequence of words (e.g., sentence) predict whether a word or a sequence of words is semantically relevant or semantically irrelevant to a topic of a document. The dataset was created with a specific goal in mind: capturing semantic text relevance. The dataset contains brain responses of 15 participants who read Wikipedia documents that were either semantically relevant or semantically irrelevant to self-selected topics.
- **Who created the dataset and on behalf of which entity?** The dataset was created by Tuukka Ruotsalo and Michiel Spapé at the University of Helsinki.
- **Who funded the creation of the dataset?** The research was partially funded by the Academy of Finland (grants 322653, 328875, 336085, 328

Table 1: An example of data pertained to each "raw" word-level brain recording instance. `Event` is a unique ID of brain recording, `word` is a word read by the participant, `topic` is the topic of the document to which the `word` belongs to, `relevant topic` is the topic selected by the participant as relevant, `sentence number` represents the sentence number in the document the `word` belongs to, `participant` is the participant's ID.

Event	Word	Topic	Selected topic	Sentence number	Participant
13380340	and	automobile	cat	1	TRPB101

350323, and 352915), the Horizon 2020 FET programme of the European Union (grant CHIST-ERA- 329 20-BCI-001) Computing resources were provided by the Finnish Grid and Cloud Infrastructure 330 (persistent id: urn:nbn:fi:research-infras-2016072533).

2 Composition

- **What do the instances that comprise the dataset represent?** The instances are word-level brain responses of humans reading text documents, together with additional information. Table 1 shows an example of additional information accompanying each "raw" word-level brain recording instance. Here, we describe the "raw" instances. The preprocessed ("cleaned") instances are explained in Section 4. All "raw" instances that belong to one participant are stored in a BrainVision format: a single participant-specific directory contains the header (`.vhdr`) extension), marker (`.vmrk`) extension), EEG data files (`.eeg`) extension), a text file (`.txt`) extension) containing the logs detailing the stimuli displayed to each participant. Thus, the "raw" directory (accessible at `https://doi.org/10.17605/OSF.IO/RPQ68`) contains 15 folders, each folder for a separate participant. The `mindir-marker-protocol5.csv` file in the "raw" directory contains the human-readable markers to interpret the marker files located in each participant's directory.
- **How many instances are there in total?** The "raw" (not preprocessed) data contains 29,080 instances.
- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** A dataset contains word-level brain recordings acquired from 15 participants, each of them reading 16 documents word by word presented on a screen.
- **What data does each instance consist of?** Each "raw" instance consists of "raw" (not preprocessed) brain recordings and accompanying data, as shown in Table 1 that are extracted from the log file.

- **Is there a label or target associated with each instance?** "Raw" instances are not directly associated with labels. The preprocessed ("cleaned") instances are associated with labels, as described in Section 4.
- **Is any information missing from individual instances?** No data are missing.
- **Are relationships between individual instances made explicit ?** Each instance is related to a word. The relationships between instances are explicit. Each word belongs to a sentence, and each sentence belongs to a document. These relationships can be seen from the event and sentence number in a document.
- **Are there recommended data splits (e.g., training, development/validation, testing)?** Yes, there are some recommended strategies for data splits (participant-independent and participant-dependent) that we have provided. See our paper.
- **Are there any errors, sources of noise, or redundancies in the dataset?** Yes, EEG data are characterised by a low signal-to-noise ratio. We have provided a preprocessed ("cleaned") version of the data that can be accessed at the following URL: `https://doi.org/10.17605/OSF.IO/RPQ68`.
- **Is the dataset self-contained, or does it link to or otherwise rely on external resources?** Yes, the dataset is self-contained.
- **Does the dataset contain data that might be considered confidential?** All data are anonymised.
- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** No.
- **Does the dataset identify any subpopulations (e.g., by age, gender)?** No, we do not provide demographics data and thus the subpopulations cannot be identified.
- **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** No, it is not possible. Personally identifiable information are not released.
- **Does the dataset contain data that might be considered sensitive in any way?** The data released do not contain sensitive information and we do not foresee that these sensitive information could be deduced from our data.

3 Collection Process

- **How was the data associated with each instance acquired?** Each participant performed eight reading tasks. Before each reading task, a participant had to select a topic (a relevant topic). Each reading task consisted of six readings trials, and each trial contained one sentence from the document belonging to the relevant topic and another sentence from the document belonging to the topic that was not selected by the participant as relevant. During each reading trial, the participant read two sentences one after another presented on a screen, word by word. We have collected word-level brain responses while participants were reading documents.
- **What mechanisms or procedures were used to collect the data? How were these mechanisms or procedures validated?** A Brain Products QuickAmp USB was used to acquire brain recordings from the electrodes placed at the 32 relatively equidistant sites of the 10/10 system of Fp1, Fp2, F7, F3, Fz, F4, 140 F8, FT9, FC5, FC1, FC2, FC6, FT10, T7, C3, Cz, C4, T8, TP9, CP5, CP1, CP2, CP6, TP10, P7, P3, 141 Pz, P4, P8, O1, Iz, and O2. We verified the correct placement of electrodes for each participant. Additionally, we checked the quality of acquired brain recordings. At the end of each reading trial, the participant had to answer which topic was selected as relevant and how many words the participant encountered are semantically relevant to the selected topic, to ensure that they have kept a focus on a semantically relevant topic (a selected topic).
- **If the dataset is a sample from a larger set, what was the sampling strategy?** We recruited participants via a convenience sampling and by advertisement on the university mailing list.
- **Who was involved in the data collection process, and how were they compensated?** Students (eight female, seven male) were involved in data collection. As a compensation for their participation (up to two hours) they received a film ticket.
- **Over what timeframe was the data collected?** The EEG data were originally recorded between November 2014 and January 2015 as part of a larger experimentation. The additional annotations, ground truth data, and other metadata were produced between 2022 and 2024.
- **Were any ethical review processes conducted?** Yes, our study was conducted in accordance to the principles of the Aalto University Research Ethics Committee.
- **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources?** Yes, we collected data directly from the individuals.
- **Were the individuals in question notified about the data collection?** Yes, the data were recorded in laboratory environment and

participants signed an informed consent. All individuals who participated in the EEG data collection were informed about the collection procedure and their rights. The participants had the right to abandon the EEG acquisition procedure at any time without any negative consequences.

In detail, the consent contained the following information:

The Experiment

The aim of this experiment is to investigate EEG measurements during an information retrieval task.

What You Will Be Asked To Do: Before the Experiment

EEG is a non-invasive technique by which brain activity is measured using electrodes that are "clipped" on a wearable EEG cap (like a shower cap). Additionally, we use five EMG electrode pairs to measure your eye movements and facial muscle activation, and EDA electrodes to measure your emotional arousal. During the preparations, you will fill out a short form measuring your handedness general behavioural decencies. To improve the conductivity of the electrodes, your skin will be cleaned before the experiment with surgical spirit (alcohol). This can be slightly irritating and can (sometimes) leave red marks. However, these should disappear within a few hours after the experiment. Please inform the experimenter if at any point the cleaning starts to feel uncomfortable. We then used a blunt, hollow needle to apply the hypoallergenic gel between the skin and the electrodes. The needle is used to avoid physical contact with the skin, for hygiene purposes, and to gently scrub the skin. It is not used to break through the skin and should not be uncomfortable. If it is, please inform the experimenter immediately.

What You Will Be Asked To Do: During the experiment.

You will be reading words on a screen while your brain activity is measured using EEG. The words are related to various common topics. Then, you will be asked questions about the topic. Depending on your answer to these questions, you will receive or lose points. Detailed instructions will be presented during the experiment itself. Before the experiment, the experimenter will demonstrate what you will be required to do. You will also be asked to try out the experiment before the real experiment begins. There will be a few breaks in the experiment where you can get something to drink or eat. Please be aware that it is not possible to go to the bathroom during the experiment. EEG is extremely sensitive to eye movements and blinks, so please keep your eyes centred on the screen as much as possible. Also, try to minimise head movements during the experiments.

If something is unclear, please ask questions at any point. Please, remember that you may quit the experiment at any point and that there will be no negative consequences for doing so.

After the experiment.

Some of the participants in this experiment can be asked to return for a second round of the experiment. Please indicate after the experiment if you do not wish this to happen.

Compensation and time exposure

The whole procedure (setup, experiment, finalisation) will take less than 3 hours. You will be compensated with two movie tickets (worth ca. €20).

In addition, the participants were informed about their rights to ask their data to be deleted, the time the data would be anonymised, and they were given the contact information of the principal investigator as well as the personnel responsible for the laboratory experimentation.

- **Did the individuals in question consent to the collection and use of their data?** Yes, the participants signed the informed consent prior to participation.
- **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** The data were initially pseudonymized by storing a separate link identifier for the purposes of the individuals to revoke or delete their data according to General Data Protection Regulation (GDPR). After the period of 3 years, the data was anonymized and link codes were destroyed. Please also see more detailed information below.
- **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?** Data were collected in a larger experiment, for which a data management plan was created. The data management plan included secure storage for the period of pseudonymized data, anonymisation procedure, and data deletion policy that are outlined elsewhere in this datasheet. As per the data management protocol (DMP), we verified the likelihood of a breach of privacy by use of the most advanced techniques available at that stage as negligible, given the sparse global representation of publicly available, personally identified EEG datasets. That situation, and hence the risk, remains as now, and since our data are anonymised, we do not contribute to this potential danger. However, as per the potential impact of the data here presented, it should be noted that EEG/ERP data are far more related to the data acquisition context than interindividual factors, and that the present context seems of little value for potential malevolent actors. In sum, the release of our dataset does, upon analysis, presents low likelihood of limited danger in terms of data protection.

4 Preprocessing

- **Was any preprocessing of the data done?** Yes, data preprocessing was applied to "raw" data. EEG data were filtered by applying 35 Hz

low-pass and 0.25 Hz high-pass filters. Invalid word-level brain recordings were estimated by calculating word-level variances ($< 0.5\mu V$) and the max-min criterium ($40\mu V$) and consequently removed.

- **Was the "raw" data saved in addition to the preprocessed data?** Yes, we provide "raw" as well as preprocessed ("cleaned") data. Data can be accessed at the following URL: `https://osf.io/rpq68/`. The `cleaned` folder contains 15 files of the *fif-file* format each containing participant's word-level brain responses and the corresponding metadata.
- **Is the software used to preprocess the instances available?** Yes, the software to preprocess data is written in Python and is available at the following ULR: `https://github.com/VadymV/EEG-dataset-for-semantic-text-relevance`.

5 Uses

- **Has the dataset been used for any tasks already?** Yes, the dataset has been used to predict word relevance and sentence relevance. See our paper for details.
- **Is there a repository that links to any or all papers or systems that use the dataset?** The GitHub repository (`https://github.com/VadymV/EEG-dataset-for-semantic-text-relevance`) contains the code to reproduce the benchmark results performed on our dataset as well as the link to the dataset.
- **What (other) tasks could the dataset be used for?** The data could be used to (1) predict document relevance, (2) whether participant's interest influences the prediction of word/sentence/document relevance, (3) whether participant's pre-knowledge about the topic of a document influences the prediction of word/sentence/document relevance.
- **Is there anything about the composition of the dataset or the way it was collected and preprocessed that might impact future uses?** Not that we are ware of.
- **Are there tasks for which the dataset should not be used?** Not that we are ware of.

6 Distribution

- **Will the dataset be distributed to third parties outside of the entity on behalf of which the dataset was created?** The dataset is accessible at the following URL: `https://osf.io/rpq68/`. We also plan to create a dataset website accessible at the following URL: `https://vadymv.github.io/EEG-dataset-for-semantic-text-relevance/`.

- **Does the dataset have a digital object identifier (DOI)?** Yes, `https://doi.org/10.17605/OSF.IO/RPQ68`.
- **When will the dataset be distributed?** The dataset is available at `https://doi.org/10.17605/OSF.IO/RPQ68`.
- **Will the dataset be distributed under a copyright or other intellectual property (IP) licence, and/or under applicable terms of use (ToU)?** The dataset is distributed under Apache Licence 2.0.
- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** No.
- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** No.

7 Maintenance

- **Who will be supporting/hosting/maintaining the dataset?** The authors of the paper that introduced the dataset.
- **How can the owner/curator/manager of the dataset be contacted?**

  Questions about the code:
  Vadym Gryshchuk (vagr@di.ku.dk).

  Questions about the EEG data collection procedure:
  Michiel Spapé (mspape@um.edu.mo) and Tuukka Ruotsalo (tr@di.ku.dk).

  Other inquiries:
  Tuukka Ruotsalo (tr@di.ku.dk or tuukka.ruotsalo@lut.fi).
- **Is there an erratum?** The erratum will be provided at the dataset website: `https://vadymv.github.io/EEG-dataset-for-semantic-text-relevance/`.
- **Will the dataset be updated?** Any changes to the dataset will be immediately communicated via the dataset website: `https://vadymv.github.io/EEG-dataset-for-semantic-text-relevance/`.
- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances?** The pseudonymization and anonymisation were carried out and communicated to the participants as follows. Printed material with identifying information (consent form) was kept in a locked room. Data collected with any electronic tool is stored in anonymised form, this means that the identity of participants can only be identified using a code (pseudonymized data). The code, as well as the consent form, was destroyed once data collection and processing was complete or the given period of 3 years was over (whichever is sooner).

- **Will older versions of the dataset continue to be supported, hosted, maintained?** Yes, older versions will be supported, maintained, and hosted. Any changes will be communicated on the dataset website and the changelog. Our dataset has a dataset version (see at the right top corner of the page).
- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** Our dataset is released under Apache License 2.0 which allows to reuse the dataset as long as the authors of the dataset get the credit of attribution (please cite our publication).