ZuCo 2.0: A Dataset of Physiological Recordings During Natural Reading and Annotation

Nora Hollenstein[1], Marius Troendle[2], Ce Zhang[1], Nicolas Langer[2]

[1] Department of Computer Science, ETH Zurich
[2] Department of Psychology, University of Zurich
{noraho,ce.zhang}@inf.ethz.ch
{m.troendle,n.langer}@psychologie.uzh.ch

Abstract

We recorded and preprocessed ZuCo 2.0, a new dataset of simultaneous eye-tracking and electroencephalography during natural reading and during annotation. This corpus contains gaze and brain activity data of 739 English sentences, 349 in a normal reading paradigm and 390 in a task-specific paradigm, in which the 18 participants actively search for a semantic relation type in the given sentences as a linguistic annotation task. This new dataset complements ZuCo 1.0 by providing experiments designed to analyze the differences in cognitive processing between natural reading and annotation. The data is freely available here: `https://osf.io/2urht/`.

Keywords: annotation, cognitive methods, corpus, EEG, eye-tracking, human language processing, naturalistic reading, physiological data

1. Introduction

How humans process language has become increasingly relevant in natural language processing (NLP) since physiological data during language understanding is more accessible and recorded with less effort. In this work, we focus on eye-tracking and electroencephalography (EEG) recordings to capture the reading process. On one hand, eye movement data provides millisecond-accurate records about where humans look when they are reading, and is highly correlated with the cognitive load associated with different stages of text processing. On the other hand, EEG records electrical brain activity across the scalp and is a direct measure of physiological processes, including language processing. The combination of both measurement methods enables us to study the language understanding process in a more natural setting, where participants read full sentences at a time, in their own speed. Eye-tracking then permits us to define exact word boundaries in the timeline of a subject reading a sentence, allowing the extraction of brain activity signals for each word.

Human cognitive language processing data is immensely useful for NLP: Not only can it be leveraged to improve NLP applications (e.g. Barrett et al. (2016) for part-of-speech tagging, or Klerke et al. (2016) for sentence compression), but also to evaluate state-of-the-art machine learning systems. For example, Hollenstein et al. (2019) evaluate word embeddings, or Schwartz et al. (2019) fine-tune language models with brain-relevant bias.

Additionally, the availability of labelled data plays a crucial role in all supervised machine learning applications. Physiological data can be used to understand and improve the labelling process (e.g. Tokunaga et al. (2017)), and, for instance, to build cost models for active learning scenarios (Tomanek et al., 2010). Is it possible to replace this expensive manual work with models trained on physiological activity data recorded from humans while reading? That is to say, can we find and extract the relevant aspects of text understanding and annotation directly from the source, i.e.
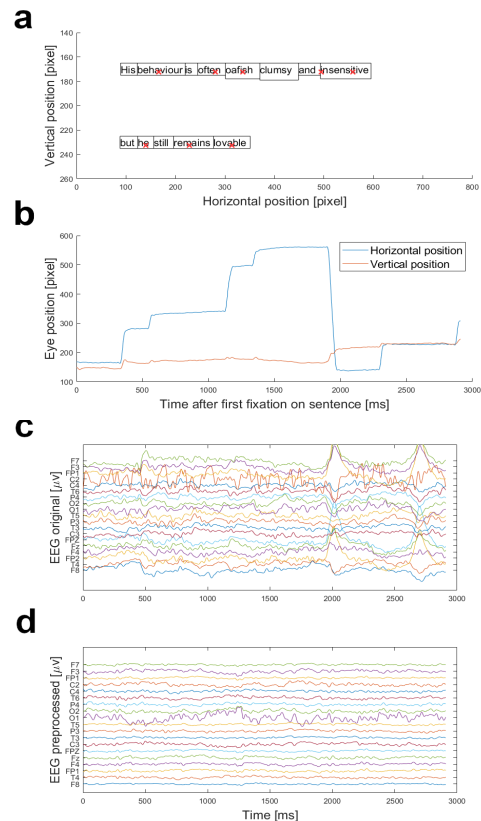

Figure 1: Visualization of eye-tracking and EEG data for a single sentence. (a) Prototypical sentence fixation data. Red crosses indicate fixations; boxes around the words indicate the wordbounds. (b) Fixation data plotted over time. (c) Raw EEG data during a single sentence. (d) Same data as in (c) after preprocessing.

eye-tracking and brain activity signals during reading? Motivated by these questions and our previously released dataset, ZuCo 1.0 (Hollenstein et al., 2018), we developed this new corpus, where we specifically aim to collect

ID	Age	Gender	LexTALE	Score NR	Score TSR	Speed NR	Speed TSR
YAC	32	female	76.25%	82.61%	83.85%	5.27	4.96
YAG	47	female	93.75%	91.30%	56.92%	7.64	8.73
YAK	31	female	100.00%	74.07%	96.41%	3.83	5.89
YDG	51	male	100.00%	91.30%	96.67%	4.97	3.93
YDR	25	male	85.00%	78.26%	96.92%	4.32	2.32
YFR	27	male	85.00%	89.13%	94.36%	6.48	4.79
YFS	39	male	90.00%	91.30%	96.15%	3.96	2.85
YHS	31	male	90.00%	78.26%	97.69%	3.30	2.40
YIS	52	male	97.50%	89.13%	98.46%	5.82	2.58
YLS	34	female	93.75%	91.30%	92.31%	5.57	5.85
YMD	31	female	100.00%	86.96%	95.64%	7.50	6.24
YMS	36	female	86.25%	89.13%	95.38%	7.68	3.35
YRH	28	female	81.25%	86.96%	95.64%	5.14	4.32
YRK	29	female	85.00%	97.83%	96.15%	7.35	7.70
YRP	23	female	82.50%	78.26%	90.00%	7.14	8.37
YSD	34	male	95.00%	93.48%	94.36%	5.01	2.87
YSL	32	female	71.25%	84.78%	83.85%	6.73	6.14
YTL*	36	male	81.25%	80.43%	94.10%	7.48	3.23
mean	**34**	**44% male**	**88.54%**	**86.36%**	**91.94%**	**5.84**	**4.81**

Table 1: Subject demographics, LexTALE scores, and control scores and reading speed (i.e. seconds per sentence) for each task. The * next to the subject ID marks a bilingual subject.

recordings during natural reading as well as during annotation.

We provide the first dataset of simultaneous eye movement and brain activity recordings to analyze and compare normal reading to task-specific reading during annotation. The **Zurich Cognitive Language Processing Corpus (ZuCo) 2.0**, including raw and preprocessed eye-tracking and electroencephalography (EEG) data of 18 subjects, as well as the recording and preprocessing scripts, is publicly available at `https://osf.io/2urht/`. It contains physiological data of each subject reading 739 English sentences from Wikipedia (see example in Figure 1). We want to highlight the re-use potential of this data. In addition to the neuro- and psycho-linguistic motivation, this corpus is especially tailored for training and evaluating machine learning algorithms for NLP purposes. It allows to conduct experiments for different semantic tasks, such as information extraction, including entity and relation discovery.

In this work, we describe the corpus construction, including experimental design and data acquisition. Further, we explain which preprocessing and feature extraction steps were applied. And finally, we conduct a detailed technical validation of the data as proof of the quality of the recordings.

2. Related Work

Some eye-tracking corpora of natural reading (e.g. the Dundee corpus (Kennedy et al., 2003), Provo corpus (Luke and Christianson, 2017) and GECO corpus (Cop et al., 2017)), and a few EEG corpora (for example, the UCL corpus by Frank and Willems (2017)) are available. It has been shown that this type of cognitive processing data is useful for improving and evaluating NLP methods (e.g. Long et

al. (2017), Barrett et al. (2018), Hale et al. (2018), Hollenstein et al. (2019)). However, before the Zurich Cognitive Language Processing Corpus (ZuCo 1.0), there was no available data for simultaneous eye-tracking and EEG recordings of natural reading. Dimigen et al. (2011) studied the linguistic effects of eye movements and EEG co-registration in natural reading and showed that they accurately represent lexical processing. Moreover, the simultaneous recordings are crucial to extract word-level brain activity signals.

While the above mentioned studies analyze and leverage natural reading, some NLP work has used eye-tracking during annotation (but, as of yet, not EEG data). Mishra et al. (2016a) and Joshi et al. (2014) recorded eye-tracking during binary sentiment annotation (positive/negative). This data was used to determine the annotation complexity of the text passages based on eye movement metrics (Mishra et al., 2017) and for sarcasm detection (Mishra et al., 2016b). Moreover, eye-tracking has been used to analyze the word sense annotation process in Hindi (Joshi et al., 2013), named entity annotation in Japanese (Tokunaga et al., 2017), and to leverage annotator gaze behaviour for English coreference resolution (Cheri et al., 2016). Finally, Tomanek et al. (2010) used eye-tracking data during entity annotation to build a cost model for active learning. However, until now there was no available data or research that analyzes the differences in the human processing of normal reading versus annotation.

ZuCo1.0 In previous work, we recorded a first dataset of simultaneous eye-tracking and EEG during natural reading (Hollenstein et al., 2018). ZuCo 1.0[1] consists of three reading tasks, two of which contain very similar reading mate-

---

[1]Data available here: `https://osf.io/q3zws/`

	NR	TSR
sentences	349	390
sent. length	mean (SD), range 19.6 (8.8), 5-53	mean (SD), range 21.3 (9.5), 5-53
total words	6828	8310
word types	2412	2437
word length	mean (SD), range 4.9 (2.7), 1-29	mean (SD), range 4.9 (2.7), 1-21
Flesch score	55.38	50.76

Table 2: Descriptive statistics of reading materials (SD = standard deviation), including Flesch readibility scores.

Relation type	Sentences
Political affiliation	45 (9)
Education	72 (10)
Wife	54 (12)
Job title	65 (11)
Employer	54 (10)
Nationality	60 (8)
Founder	40 (8)
total	**390 (68)**

Table 3: Distribution of relation types in the task-specific reading. The right column contains the number of sentences, and the number control sentences without a relation in brackets.

rial and experiments as presented in the current work. However, the main difference and reason for recording ZuCo 2.0 consists in the experiment procedure. For ZuCo 1.0 the normal reading and task-specific reading paradigms were recorded in different sessions on different days. Therefore, the recorded data is not appropriate as a means of comparison between natural reading and annotation, since the differences in the brain activity data might result mostly from the different sessions due to the sensitivity of EEG. This, and extending the dataset with more sentences and more subjects, were the main factors for recording the current corpus. We purposefully maintained an overlap of some sentences between both datasets to allow additional analyses (details are described in Section 3.2.).

3. Corpus Construction

In this section we describe the contents and experimental design of the ZuCo 2.0 corpus.

3.1. Participants

We recorded data from 19 participants and discarded the data of one of them due to technical difficulties with the eye-tracking calibration. Hence, we share the data of 18 participants. All participants are healthy adults (mean age = 34 (SD=8.3), 10 females). Their native language is English, originating from Australia, Canada, UK, USA or South Africa. Two participants are left-handed and three participants wear glasses for reading. Details on subject demographics can be found in Table 1. All participants gave written consent for their participation and the re-use of the data prior to the start of the experiments. The study was approved by the Ethics Commission of the University of Zurich.

3.2. Reading materials

During the recording session, the participants read 739 sentences that were selected from the Wikipedia corpus provided by Culotta et al. (2006). This corpus was chosen because it provides annotations of semantic relations. Relation detection is a high-level semantic task requiring complex cognitive processing. We included seven of the originally defined relation types: *political_affiliation*, *education*, *founder*, *wife/husband*, *job_title*, *nationality*, and *employer*. The sentences were chosen in the same length range as ZuCo 1.0, and with similar Flesch reading ease scores (Kincaid et al., 1975). The dataset statistics are shown in Table 2.

Of the 739 sentences, the participants read 349 sentences in a normal reading paradigm, and 390 sentences in a task-specific reading paradigm, in which they had to determine whether a certain relation type occurred in the sentence or not. Table 3 shows the distribution of the different relation types in the sentences of the task-specific annotation paradigm.

Purposefully, there are 63 duplicates between the normal reading and the task-specific sentences (8% of all sentences). The intention of these duplicate sentences is to provide a set of sentences read twice by all participants with a different task in mind. Hence, this enables the comparison of eye-tracking and brain activity data when reading normally and when annotating specific relations (see examples in Section 4.).

Furthermore, there is also an overlap in the sentences between ZuCo 1.0 and ZuCo 2.0. 100 normal reading and 85 task-specific sentences recorded for this dataset were already recorded in ZuCo 1.0. This allows for comparisons between the different recording procedures (i.e. session-specific effects) and between more participants (subject-specific effects).

3.3. Experimental design

As mentioned above, we recorded two different reading tasks for the ZuCo 2.0 dataset. During both tasks the participants were able to read in their own speed, using a control pad to move to the next sentence and to answer the control questions, which allowed for natural reading. Since each subject reads at their own personal pace, the reading speed between varies between subjects. Table 1 shows the average reading speed for each task, i.e. the average number of seconds a subject spends per sentence before switching to the next one.

All 739 sentences were recorded in a single session for each participant. The duration of the recording sessions was between 100 and 180 minutes, depending on the time required to set up and calibrate the devices, and the personal reading speed of the participants.

We recorded 14 blocks of approx. 50 sentences, alternating between tasks: 50 sentences of normal reading, followed by 50 sentences of task-specific reading. The order of blocks and sentences within blocks was identical for all subjects.
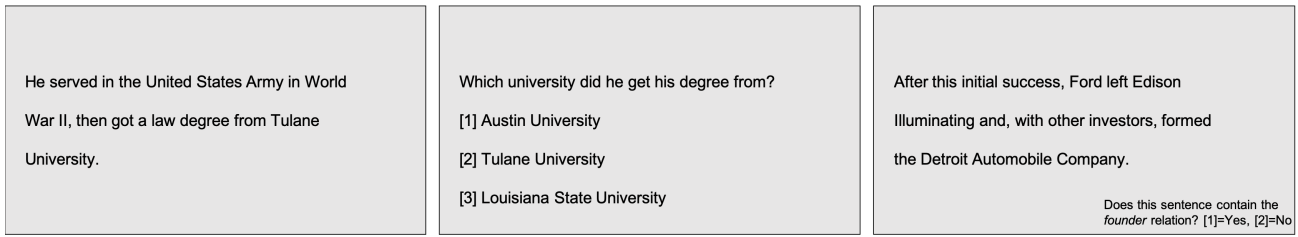
| He served in the United States Army in World War II, then got a law degree from Tulane University. | Which university did he get his degree from?

[1] Austin University
[2] Tulane University
[3] Louisiana State University | After this initial success, Ford left Edison Illuminating and, with other investors, formed the Detroit Automobile Company.

Does this sentence contain the *founder* relation? [1]=Yes, [2]=No |

Figure 2: Example sentences on the recording screen: (left) a normal reading sentence, (middle) a control question for a normal reading sentence, and (right) a task-specific annotation sentence.
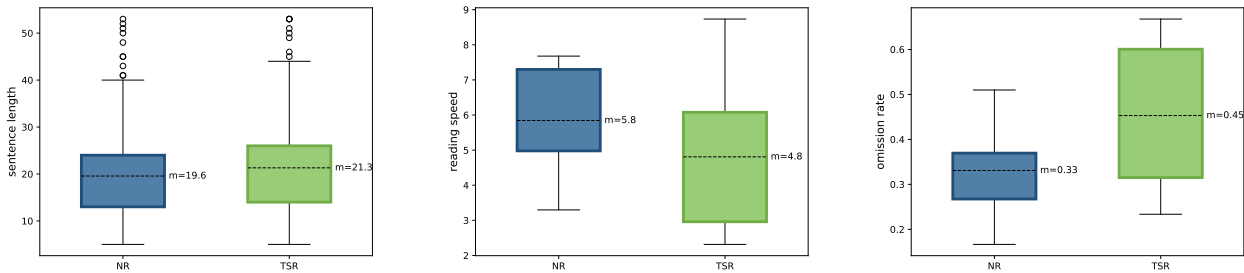
Figure 3: Sentence length (words per sentence), reading speed (seconds per sentence) and omission rate (percentage of words not fixated) comparison between normal reading (NR) and task-specific reading (TSR).

Each sentence block was preceded by a practice round of three sentences and followed by a short break to ensure a clear separation between the reading tasks.

Normal reading (NR) In the first task, participants were instructed to read the sentences naturally, without any specific task other than comprehension. Participants were told to read the sentences normally without any special instructions. Figure 2 (left) shows an example sentence as it was depicted on the screen during recording. As shown in Figure 2 (middle), the control condition for this task consisted of single-choice questions about the content of the previous sentence. 12% of randomly selected sentences were followed by such a comprehension question with three answer options on a separate screen.

Task-specific reading (TSR) In the second task, the participants were instructed to search for a specific relation in each sentence they read. Instead of comprehension questions, the participants had to decide for each sentence whether it contains the relation or not, i.e. they were actively annotating each sentence. Figure 2 (right) shows an example screen for this task. 17% of the sentences did not include the relation type and were used as control conditions. All sentences within one block involved the same relation type. The blocks started with a practice round, which described the relation and was followed by three sample sentences, so that the participants would be familiar with the respective relation type.

3.4. Linguistic assessment

As a linguistic assessment, the vocabulary and language proficiency of the participants was tested with the Lex-TALE test (Lexical Test for Advanced Learners of English, Lemhöfer and Broersma (2012)). This is an unspeeded lexical decision task designed for intermediate to highly pro-
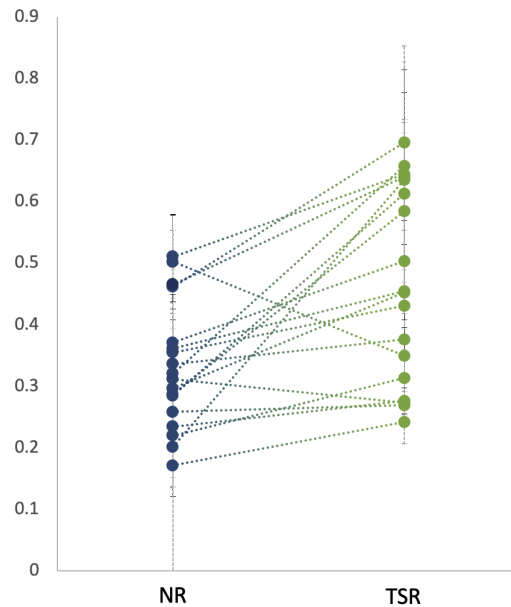

Figure 4: Skipping proportion on word level for both tasks.

ficient language users. The average LexTALE score over all participants was 88.54%. Moreover, we also report the scores the participants achieved with their answers to the reading comprehension control questions and their relation annotations. The detailed scores for all participants are also presented in Table 1.

3.5. Data acquisition

Data acquisition took place in a sound-attenuated and dark experiment room. Participants were seated at a distance of 68cm from a 24-inch monitor with a resolution of 800x600
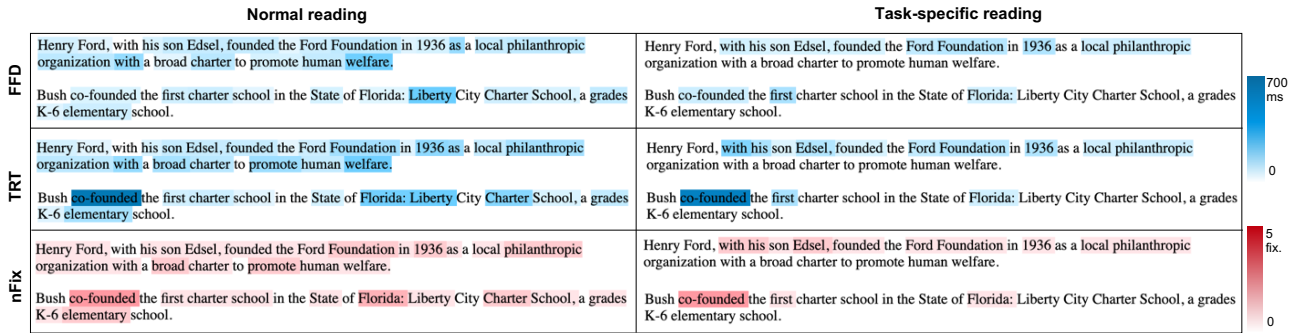
Figure 5: Fixation heatmaps for two sentences containing the relation *founder*, showing a comparison of the eye-tracking features between normal reading and task-specific annotation reading (first fixation duration (FFD), total reading time (TRT), number of fixations (nFix).

pixels. A stable head position was ensured via a chin rest. Participants were instructed to stay as still as possible during the tasks to avoid motor EEG artifacts. Participants were also offered snacks and water during the breaks and were encouraged to rest. All sentences were presented at the same position on the screen and could span multiple lines. The sentences were presented in black on a light grey background with font size 20-point Arial, resulting in a letter height of 0.8 mm. The experiment was programmed in MATLAB 2016b (MATLAB, MathWorks, 2000), using PsychToolbox (Brainard, 1997). Participants completed the tasks sitting alone in the room, while two research assistants were monitoring their progress in the adjoining room. All recording scripts including detailed participant instructions are available alongside the data.

Eye-tracking acquisition Eye position and pupil size were recorded with an infrared video-based eye tracker (EyeLink 1000 Plus, SR Research) at a sampling rate of 500 Hz. The eye tracker was calibrated with a 9-point grid at the beginning of the session and re-validated before each block of sentences.

EEG acquisition High-density EEG data were recorded at a sampling rate of 500 Hz with a bandpass of 0.1 to 100 Hz, using a 128-channel EEG Geodesic Hydrocel system (Electrical Geodesics). The recording reference was set at electrode *Cz*. The head circumference of each participant was measured to select an appropriately sized EEG net. To ensure good contact, the impedance of each electrode was checked prior to recording, and was kept below 40 kOhm. Electrode impedance levels were checked after every third block of 50 sentences (approx. every 30 mins) and reduced if necessary.

3.6. Preprocessing and feature extraction

Eye-tracking The eye-tracking data consists of (x, y) gaze location entries for all individual fixations (Figure 1b). Coordinates were given in pixels with respect to the monitor coordinates (the upper left corner of the screen was $(0, 0)$ and down/right was positive). We provide this raw data as well as various engineered eye-tracking features. For this feature extraction only fixations within the boundaries of each displayed word were extracted. Data points distinctly not associated with reading (minimum distance
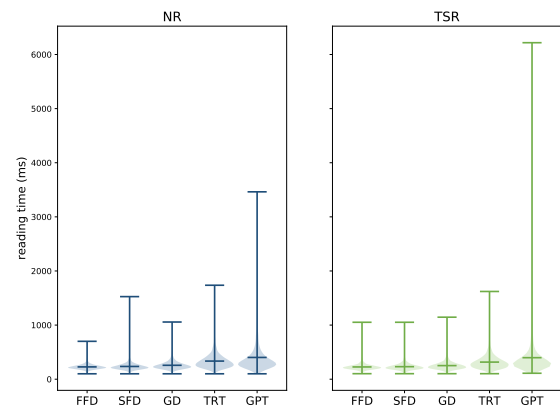

Figure 6: Violin plots showing means, distributions, and ranges of the reading time measures per word for each task and each eye-tracking feature (x-axis) in milliseconds.

of 50 pixels to the text) were excluded. Additionally, fixations shorter than 100 ms were excluded from the analyses, because these are unlikely to reflect fixations relevant for reading (Sereno and Rayner, 2003). On the basis of the GECO and ZuCo 1.0 corpora, we extracted the following features: (i) *gaze duration* (GD), the sum of all fixations on the current word in the first-pass reading before the eye moves out of the word; (ii) *total reading time* (TRT), the sum of all fixation durations on the current word, including regressions; (iii) *first fixation duration* (FFD), the duration of the first fixation on the prevailing word; (iv) *single fixation duration* (SFD), the duration of the first and only fixation on the current word; and (v) *go-past time* (GPT), the sum of all fixations prior to progressing to the right of the current word, including regressions to previous words that originated from the current word. For each of these eye-tracking features we additionally computed the pupil size. Furthermore, we extracted the number of fixations and mean pupil size for each word and sentence.

EEG The EEG data shared in this project are available as raw data, but also preprocessed with Automagic (version 1.4.6, Pedroni et al. (2019)), a tool for automatic EEG data cleaning and validation. 105 EEG channels (i.e.
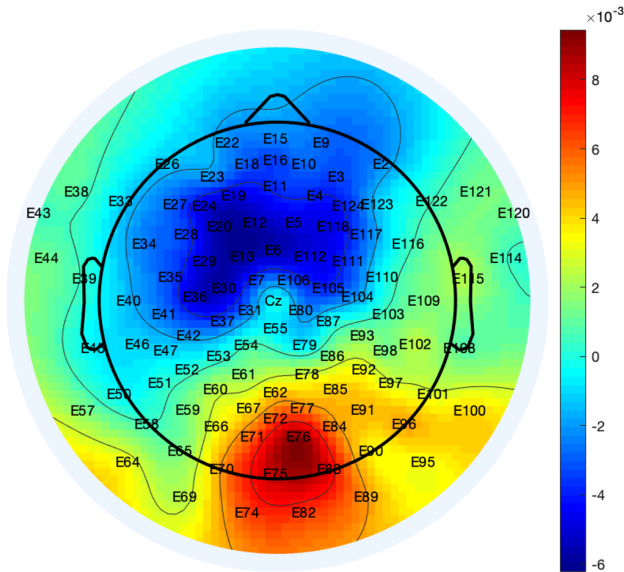
Figure 7: Topography plot of the difference in the EEG activity between both reading tasks, averaged over all sentences all all subjects of the dataset (scalp viewed from above, nose at the top).

electrodes) were used from the scalp recordings. 9 EOG channels were used for artifact removal and additional 14 channels lying mainly on the neck and face were discarded before data analysis. Bad channels were identified and interpolated. We used the Multiple Artifact Rejection Algorithm (MARA), a supervised machine learning algorithm that evaluates ICA components, for automatic artifact rejection. MARA has been trained on manual component classifications, and thus captures a wide range of artifacts. MARA is especially effective at detecting and removing eye and muscle artifact components. The effect of this preprocessing can be seen in Figure 1d.

After preprocessing, we synchronized the EEG and eye-tracking data to enable EEG analyses time-locked to the onsets of fixations. To compute oscillatory power measures, we band-pass filtered the continuous EEG signals across an entire reading task for five different frequency bands resulting in a time-series for each frequency band. The independent frequency bands were determined as follows: $theta_1$ (4–6 Hz), $theta_2$ (6.5–8 Hz), $alpha_1$ (8.5–10 Hz), $alpha_2$ (10.5–13 Hz), $beta_1$ (13.5–18 Hz), $beta_2$ (18.5–30 Hz), $gamma_1$ (30.5–40 Hz), and $gamma_2$ (40–49.5 Hz). We then applied a Hilbert transformation to each of these time-series. We specifically chose the Hilbert transformation to maintain the temporal information of the amplitude of the frequency bands, to enable the power of the different frequencies for time segments defined through the fixations from the eye-tracking recording. Thus, for each eye-tracking feature we computed the corresponding EEG feature in each frequency band. Furthermore, we extracted sentence-level EEG features by calculating the power in each frequency band, and additionally, the difference of the power spectra between frontal left and right homologue electrodes pairs. For each eye-tracking based EEG feature, all channels were subject to an artifact rejection criterion of

$90\mu V$ to exclude trials with transient noise.

4. Data Validation

The aim of the technical validation of the data is to guarantee good recording quality and to replicate findings of previous studies investigating co-registration of EEG and eye movement data during natural reading tasks (e.g. Dimigen et al. (2011)). We also compare the results to ZuCo 1.0 (Hollenstein et al., 2018), which allows a more direct comparison due to the analogous recording procedure.

Eye-tracking We validated the recorded eye-tracking data by analyzing the fixations made by all subjects through their reading speed and omission rate on sentence level. The omission rate is defined as the percentage of words that is not fixated in a sentence. Figure 3 (middle) shows the mean reading speed over all subjects, measured in seconds per sentence and Figure 3 (right) shows the mean omission rates aggregated over all subjects for each task. Clearly, the participants made less fixations during the task-specific reading, which lead to faster reading speed.

Moreover, we corroborated these sentence-level metrics by visualizing the skipping proportion on word level (Figure 4). The skipping proportion is the average rate of words being skipped (i.e. not being fixated) in a sentence. As expected, this also increases in the task-specific reading.

Although the reading material is from the same source and of the same length range (see Figure 3 (left)), in the first task (NR) passive reading was recorded, while in the second task (TSR) the subjects had to annotate a specific relation type in each sentence. Thus, the task-specific annotation reading lead to shorter passes because the goal was merely to recognize a relation in the text, but not necessarily to process the every word in each sentence. This distinct reading behavior is shown in Figure 5, where fixations occur until the end of the sentence during normal reading, while during task-specific reading the fixations stop after the decisive words to detect a given relation type. Finally, we also analyzed the average reading times for each of the extracted eye-tracking features. The means and distributions for both tasks are shown in Figure 6. These results are in line with the recorded data in ZuCo 1.0, as well as with the features extracted in the GECO corpus (Cop et al., 2017).

EEG The differences between normal reading and task-specific annotation reading is also evident in the brain activity data (see Figure 7).

As a first validation step, we extracted fixation-related potentials (FRPs), where the EEG signal during all fixations of one task are averaged. Figure 8 shows the time-series of the resulting FRPs for two electrodes (*PO8* and *Cz*), as well as topographies of the voltage distributions across the scalp at selected points in time. The five components (for which the scalp topographies are plotted) are highly similar in the time-course of the chosen electrodes to Dimigen et al. (2011) as well as to ZuCo 1.0.

Moreover, these previous studies were able to show an effect of fixation duration on the resulting FRPs. To show this dependency we followed two approaches. First, for each reading task, all single-trial FRPs were ordered by fixation duration and a vertical sliding time-window was used
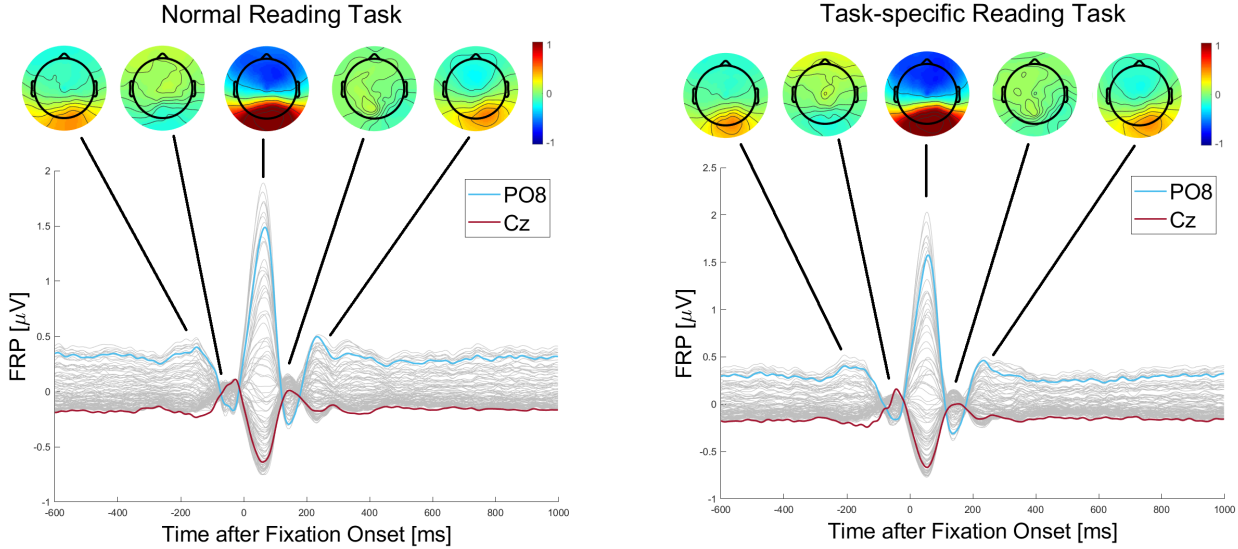
Figure 8: Fixation-related potentials (FRPs) during both task conditions with selected scalp level potential distributions. Topographies show color-coded amplitudes in microvolt.
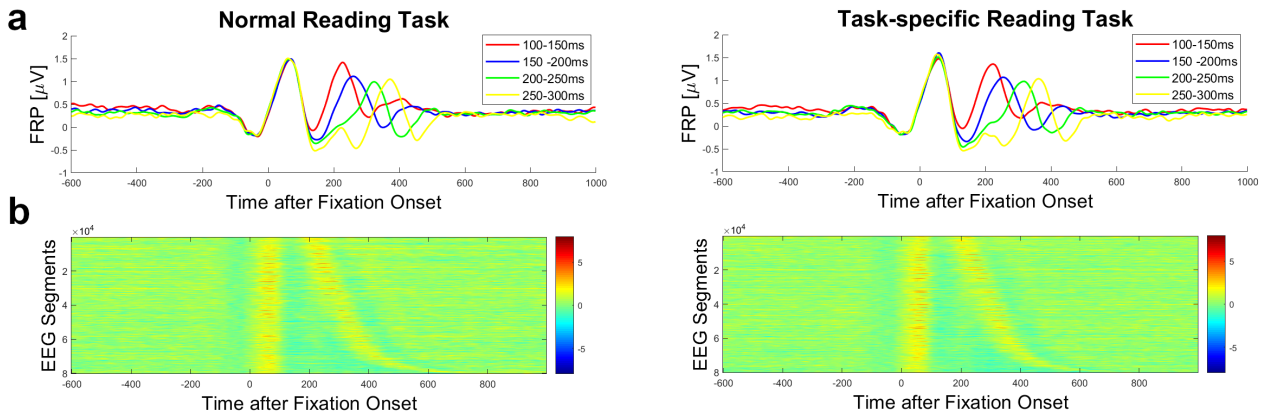

Figure 9: Clustered EEG segments. (a) FRPs of electrode *Cz*, clustered by duration of the fixation. (b) Each horizontal line represents the mean of the current and 50 adjacent EEG epochs, segmented on fixation onset. Segments are ordered by fixation duration (top: shortest fixation, bottom: longest fixation). Color represents the amplitude of the signal in microvolt.

to smooth the data (Dimigen et al., 2011). Figure 9 (bottom) shows the resulting plots. In line with this previous work, a first positivation can be identified at 100 ms post-fixation onset. A second positive peak is located dependent on the duration of the fixation, which can be explained by the time-jittered succeeding fixation. The second approach is based on Henderson et al. (2013) in which single trial EEG segments are clustered by the duration of the current fixation. As shown in Figure 9 (top), we chose four clusters and averaged the data within each cluster to four distinct FRPs, depending on the fixation duration. Again, the same positivation peaks become apparent. Both findings are consistent with the previous work mentioned and with our findings from ZuCo 1.0.

5. Conclusion

We presented a new, freely available corpus of eye movement and electrical brain activity recordings during natural reading as well as during annotation. This is the first dataset

that allows for the comparison between these two reading paradigms. We described the materials and experiment design in detail and conducted an extensive validation to ensure the quality of the recorded data. Since this corpus is tailored to cognitively-inspired NLP, the applications and re-use potentials of this data are extensive. The provided word-level *and* sentence-level eye-tracking and EEG features can be used to improve and evaluate NLP and machine learning methods, for instance, to evaluate linguistic phenomena in neural models via neurolinguistic data. For instance, human language processing recordings can be used to probe the syntactic skills of language models (Toneva and Wehbe, 2019) or to evaluate the cognitive plausibility of word representations (Hollenstein et al., 2019). In addition, because the sentences contains semantic relation labels and the annotations of all participants, it can also be widely used for relation extraction and classification. Finally, the two carefully constructed reading paradigms allow for the comparison between normal reading and read-

ing during annotation, which can be relevant to improve the manual labelling process as well as the quality of the annotations for supervised machine learning.

6. Bibliographical References

Barrett, M., Bingel, J., Keller, F., and Søgaard, A. (2016). Weakly supervised part-of-speech tagging using eye-tracking data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 579–584.

Barrett, M., Bingel, J., Hollenstein, N., Rei, M., and Søgaard, A. (2018). Sequence classification with human attention. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 302–312.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial vision*, 10(4):433–436.

Cheri, J., Mishra, A., and Bhattacharyya, P. (2016). Leveraging annotators' gaze behaviour for coreference resolution. In *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning*, pages 22–26.

Cop, U., Dirix, N., Drieghe, D., and Duyck, W. (2017). Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49(2):602–615.

Culotta, A., McCallum, A., and Betz, J. (2006). Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 296–303.

Dimigen, O., Sommer, W., Hohlfeld, A., Jacobs, A. M., and Kliegl, R. (2011). Coregistration of eye movements and EEG in natural reading: analyses and review. *Journal of Experimental Psychology: General*, 140(4):552.

Frank, S. L. and Willems, R. M. (2017). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, 32(9):1192–1203.

Hale, J., Dyer, C., Kuncoro, A., and Brennan, J. R. (2018). Finding syntax in human encephalography with beam search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2727–2736.

Henderson, J. M., Luke, S. G., Schmidt, J., and Richards, J. E. (2013). Co-registration of eye movements and event-related potentials in connected-text paragraph reading. *Frontiers in systems neuroscience*, 7:28.

Hollenstein, N., Rotsztejn, J., Troendle, M., Pedroni, A., Zhang, C., and Langer, N. (2018). ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific Data*.

Hollenstein, N., de la Torre, A., Langer, N., and Zhang, C. (2019). CogniVal: A framework for cognitive word embedding evaluation. In *Proceedings of the 23nd Conference on Computational Natural Language Learning*.

Joshi, S., Kanojia, D., and Bhattacharyya, P. (2013). More than meets the eye: Study of human cognition in sense annotation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 733–738.

Joshi, A., Mishra, A., Senthamilselvan, N., and Bhattacharyya, P. (2014). Measuring sentiment annotation complexity of text. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 36–41.

Kennedy, A., Hill, R., and Pynte, J. (2003). The Dundee corpus. In *Proceedings of the 12th European Conference on Eye Movement*.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Klerke, S., Goldberg, Y., and Søgaard, A. (2016). Improving sentence compression by learning to predict gaze. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1528–1533.

Lemhöfer, K. and Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of english. *Behavior Research Methods*, 44(2):325–343.

Long, Y., Lu, Q., Xiang, R., Li, M., and Huang, C.-R. (2017). A cognition based attention model for sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 462–471.

Luke, S. G. and Christianson, K. (2017). The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, pages 1–8.

MATLAB, MathWorks. (2000). The language of technical computing, the MathWorks.

Mishra, A., Kanojia, D., and Bhattacharyya, P. (2016a). Predicting readers' sarcasm understandability by modeling gaze behavior. In *AAAI Conference on Artificial Intelligence*, pages 3747–3753.

Mishra, A., Kanojia, D., Nagar, S., Dey, K., and Bhattacharyya, P. (2016b). Harnessing cognitive features for sarcasm detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1095–1104.

Mishra, A., Kanojia, D., Nagar, S., Dey, K., and Bhattacharyya, P. (2017). Scanpath complexity: Modeling reading effort using gaze information. In *Thirty-First AAAI Conference on Artificial Intelligence*, pages 4429–4436.

Pedroni, A., Bahreini, A., and Langer, N. (2019). Automagic: Standardized preprocessing of big EEG data. *NeuroImage*.

Schwartz, D., Toneva, M., and Wehbe, L. (2019). Inducing brain-relevant bias in natural language processing models. In *Advances in Neural Information Processing Systems*, pages 14100–14110.

Sereno, S. C. and Rayner, K. (2003). Measuring word

recognition in reading: eye movements and event-related potentials. *Trends in cognitive sciences*, 7(11):489–493.

Tokunaga, T., Nishikawa, H., and Iwakura, T. (2017). An eye-tracking study of named entity annotation. *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 758–764.

Tomanek, K., Hahn, U., Lohmann, S., and Ziegler, J. (2010). A cognitive cost model of annotations based on eye-tracking data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1158–1167.

Toneva, M. and Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems*, pages 14928–14938.