The Index of Cognitive Activity Predicts Cognitive Processing Load in Linguistic Task

HONG GAO, University of Tübingen, Germany
ZIJIAN LU, Justus Liebig-University Giessen, Germany
VERA DEMBERG, Saarland University, Germany
ENKELEJDA KASNECI, University of Tübingen, Germany

The Index of Cognitive Activity (ICA) has been shown to index cognitive processing load in language processing tasks in previous work. However, the effect has so far only been shown as an aggregate effect across many subjects and trials. The feasibility and reliability of using ICA as a predictor of language-related cognitive processing load in a single-trial setting have not yet been assessed. Therefore, in this study, we compare the single-trial performance of various classification models, including Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), Random Forest, and Gradient Boosting on two sentence-reading experiments. These algorithms were trained on ICA values to predict semantic and syntactic violations of the sentences read by the participants. The results showed that all trained classifiers performed above the 50% chance level. Of these classifiers, Gradient Boosting showed the best performance with an accuracy of 74.48% for semantic violation detection and 71.61% for syntactic violation detection, respectively. Our results indicate that the ICA is a viable measure for detecting cognitive processing load caused by language violation on a single-trial basis.

CCS Concepts: • **Computing methodologies** → **Classification and regression trees**; • **Human-centered computing** → Empirical studies in HCI.

Additional Key Words and Phrases: Index of Cognitive Activity, cognitive load detection, classification, eye tracking, linguistic reading comprehension

ACM Reference Format:
Hong Gao, Zijian Lu, Vera Demberg, and Enkelejda Kasneci. 2021. The Index of Cognitive Activity Predicts Cognitive Processing Load in Linguistic Task. In *EMICS '21: ACM CHI '21 Workshop on Eye Movements as an Interface to Cognitive State, May 14, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 5 pages.

1 INTRODUCTION

Previous studies have provided evidence that pupil size reflects cognitive load in various human cognitive processes, such as problem solving [3, 12, 16], language comprehension [9–11], attention [2, 5] , and many more. However, pupil dilation is known to be affected by the illumination or subject emotional status [14, 15]. When the illumination changes, dilations induced by cognitive activity and the light reflex may overlay [6]. Even under controlled lighting condition, a potential confound for pupillometric measures of cognitive load may exist due to random pupillary oscillation [7]. To address this problem, Marshall introduced the so-called Index of Cognitive Activity (ICA) [13].

More specifically, the ICA measures small rapid pupil dilations from a continuous recording of pupil diameter based on a wavelet analysis [1], and was found to be robust to light changes [14]. Regarding the computational details of the ICA approach, we refer the reader to [4]. Various recent works support the claim that the ICA is a reliable measure of

cognitive processing load in a variety of cognitive tasks. For example, in [14], several cognitive tasks were performed indicating that the those tasks induced significantly higher ICA values than the non-cognitive tasks. In addition, ICA is also considered to be an indicator of cognitive processing load during language comprehension. For instance, dual-task studies have been proposed [6, 17] in which subjects were required to perform both a tracking task and a linguistic comprehension task. In both studies, ICA proved to be sensitive to both tasks. Similar to [6, 17], several linguistic comprehension tasks were conducted in [7]. It was found that ICA increased after the onset of the language violations, further supporting the point that ICA can be a reliable measure of cognitive load in language comprehension processing.

Although the ICA has been considered as a measure of cognitive processing load in language comprehension tasks, the feasibility of using the ICA to predict the cognitive processing load induced by language violations in single trials has not been thoroughly investigated – all previous studies evaluated the sensitivity of the ICA to the linguistic violations across a large number of subjects and items. For non-linguistic tasks, there are however initial indications that the ICA might be suitable for detecting such single-trial effects. In [3], classifiers for detecting workload in n-back tasks were trained on several eye-related measures, including pupil diameter, blinks, and the ICA. An accuracy of 70.4% was achieved in a real-time workload classification task and higher accuracy of 76.8% in an offline classification setting. However, in their models, the feature importance of the ICA is lower than that of the other features. Therefore, the feasibility of using ICA as a feature to predict cognitive load in a linguistic task could not be fully validated.

In this work, we address this open research question by investigating the feasibility and reliability of using the ICA value to predict the cognitive processing load caused by language violations. To the best of our knowledge, this is the first study to use the ICA value as a single feature to detect cognitive load in reading tasks. In addition, we systematically compare several machine learning models employed for cognitive load prediction. Our results showed that the ICA value is a robust and reliable predictor of cognitive load induced by language violations in sentence reading tasks. In addition, the models enable real-time classification, which will benefit real-time language violation detection in real life.

2 METHOD

2.1 Participants

32 college students (22 female, 10 male), whose ages range from 18-35, were recruited to participate in our experiment. Due to data incompleteness caused by technical issues of the eye tracker, data of two participants (1 female, 1 male) were discarded. All participants were right-handed and had a normal or corrected-to-normal vision. All participants were asked to sign an informed consent form before the experiment.

2.2 Apparatus

To record the pupillary response, we employed an Eyelink 1000 Plus[1] eye tracker, with a 35mm lens and 250Hz sampling rate. The experimental environment was created and controlled using the SR Research software *Experiment Builder*[2]. Stimuli were presented on a 22" monitor (1920×1080). Both pupil's area data were tracked in the Ellipse model [8].

2.3 Materials and Procedure

In our experiment, data was collected using a similar experimental procedure as in [7]. Forty pairs of German sentence each were generated for the semantic and syntactic violation conditions using the same rules as in [7], with each pair consisting of one correct sentence as the control condition and one incorrect sentence as the violation condition.

[1]https://www.sr-research.com/eyelink-1000-plus/
[2]https://www.sr-research.com/experiment-builder/

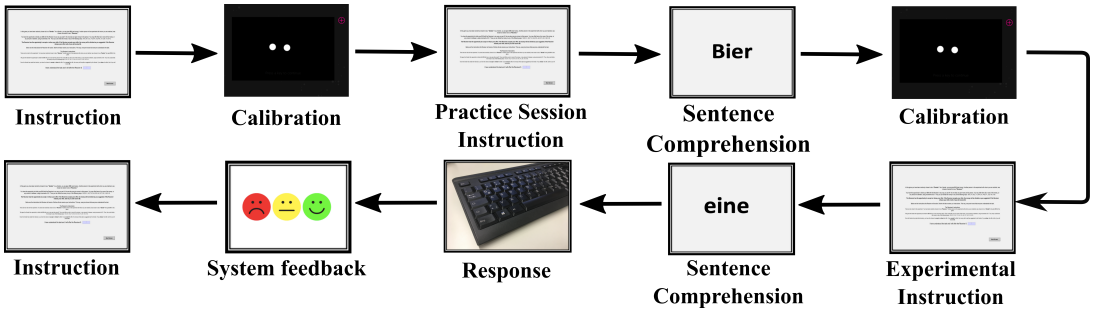

Fig. 1. Experimental protocol.

The experimental procedure is depicted in Fig 1. After a brief introduction to the experimental procedure, the participants were seated in front of the monitor with a viewing distance of about 70 cm and asked to stabilize their head on the head-rest rack in a comfortable position. Then, a built-in 13-point calibration process was performed to ensure high-quality eye-tracking data. In the following practice session, participants were presented with 4 sentence pairs (2 semantic and 2 syntactic pairs). The sentence stimulus was presented word by word on the screen. For each word, the display time was 450ms, followed by a 150ms interval before the next word appeared. Participants were asked to press the button immediately upon finding a violation in the sentence that they were reading. Feedback was provided after each trial to inform the participants whether or not they correctly identified the violation. Data of the test trials were not recorded. In the following, actual experimental session, each participant completed a total of 80 trials (20 semantic and 20 syntactic sentence pairs). The sentences were randomly assigned to each trial. Those trials were divided into 4 blocks, with each block containing 20 sentences. After each block, participants took a break. The camera set-up adjustment was repeated before the new block. Therefore, a total of 30×40 trials were conducted for each violation. The duration of the experiment was about 20-30 minutes.

3 MACHINE LEARNING CLASSIFICATION MODELS

ICA Feature Extraction: Following the previous studies [6, 7], we calculated ICA values by counting the ICA events (rapid pupil dilation) for a time window of 100-millisecond. These studies provide strong evidence that ICA is a reliable measure of cognitive load invoked by linguistic stimuli such as semantic and syntactic violations in sentences read by participants. Interestingly, in [7], the authors found that both semantic ($\chi^2 = 8.1104$, $p = 0.0025$) and syntactic ($\chi^2 = 4.66$, $p = 0.031$) violation were significant positive predictors of ICA values in the post critical region onset.

There were 1108 and 1134 frames for the semantic and syntactic violation detection groups, respectively. Due to a delay of about half a second of ICA effect on sentence violation [7], time points in the time region of 400ms to 1100ms after each violation stimulus were chosen as the input in our models. For the control group (correct sentence trials), time points from the same time region were extracted as the input. Therefore, 8 ICA values corresponding to the time points were used as features in our prediction models. The data were normalized using a min-max normalization method.

Prediction model: In order to detect participants' cognitive processing load induced by language violations (semantic and syntactic violations are detected separately) in sentence comprehension tasks, we build four classifiers, namely a Support Vector Machine (SVM), a k-Nearest Neighbors (k-NN) model, a classifier based on Random Forest, and Gradient Boosting. To make the model subject-independent, nested 10-fold cross-validation was performed to average the models

Table 1. Classifiers evaluation results for Semantic and Syntactic violation conditions

Classifiers	Semantic				Syntactic			
	Accuracy	F1-Score	Precision	Recall	Accuracy	F1-Score	Precision	Recall
SVM	0.6718	0.6664	0.6696	0.6601	0.6471	0.6455	0.6469	0.6451
kNN	0.6484	0.6487	0.6394	0.6383	0.6285	0.6129	0.6204	0.6344
RandomForest	0.6267	0.6248	0.6351	0.6228	0.6259	0.6067	0.6165	0.6253
GradientBoosting	**0.7448**	0.7395	0.9518	0.7375	**0.7161**	0.7178	0.7055	0.6917

against each of the folds and then finalize our models. More specifically, in each iteration, we randomly selected three participants as the test set and the rest of the twenty-seven participants as the training set. For each iteration, the three participants that were not selected in the previous cycles should be selected as the test set. Therefore, all participants are used as training and test set. In this context, our models generalize to unseen data, which mitigates the risk of overfitting. The final performance scores of the prediction models reported in this paper are the mean values of the scores on the test set. According to the training results, we set the best hyper-parameters for each classifier: for SVM, the RBF kernel function was used; for the k-NN classifier, 6-NN was evaluated; for Random Forest, 500 trees were used to train the model; for the Gradient Boosting classifier, the number of boosting stages to be performed was set to 600. We evaluate the classification performance of the models in terms of accuracy, f1-score, precision, and recall.

4 RESULTS AND DISCUSSION

In our study, several classification models were trained to predict the cognitive processing load induced by the semantic and syntactic violations in sentence reading tasks.Models were trained using Python 3.6[3] on a host computer equipped with a 6×3GHz Intel i5-9500 CPU and 16GB RAM. The chance level for both violation conditions was 50%. Table 1 presents the overall performance measurement scores for all four classifiers for the semantic and syntactic violation detection, respectively. The results show that training performance was similar for the two violation conditions. Of those classifiers, Gradient Boosting performed best, with accuracy of 74.48% and 71.61% for semantic and syntactic violations detection, respectively, outperforming all other classifiers by 6.9% or more. The results show that all the classifiers trained in our study performed above chance level. Particularly, our models enable feature extraction and classification in real-time, as the computational effort is very low.

Therefore, all classifiers trained in our study are suitable to discriminate the ICA value between violation (incorrect sentence) and control (correct sentence) conditions in the sentence reading task. Our findings further reveal that the ICA is a robust and reliable predictor of language-related cognitive processing load in a single-trial setting. These results help to position the ICA as an effective feature for identifying language violations even in a single-trial instance. In addition, since the ICA is robust to illumination changes, it can be used in real-world settings where lighting condition cannot be fully controlled. Therefore, such a language violation detection method could be used, for instance, as a language teaching assistant that helps teachers to better assist their students in a real-time scenario. Future work should investigate whether this detection method works with a consumer-level eye tracker at a lower sampling rate. Furthermore, in addition to detecting language violations from the correct sentences, it is interesting to investigate whether it is possible to distinguish between two types of language violations.

[3]https://www.python.org/

REFERENCES

[1] Felix Abramovich, Trevor C Bailey, and Theofanis Sapatinas. 2000. Wavelet analysis and its statistical applications. *Journal of the Royal Statistical Society: Series D (The Statistician)* 49, 1 (2000), 1–29.

[2] Dag Alnæs, Markus Handal Sneve, Thomas Espeseth, Tor Endestad, Steven Harry Pieter van de Pavert, and Bruno Laeng. 2014. Pupil size signals mental effort deployed during multiple object tracking and predicts brain activity in the dorsal attention network and the locus coeruleus. *Journal of vision* 14, 4 (2014), 1–1.

[3] Tobias Appel, Christian Scharinger, Peter Gerjets, and Enkelejda Kasneci. 2018. Cross-Subject Workload Classification Using Pupil-Related Measures. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & amp; Applications* (Warsaw, Poland) *(ETRA '18)*. Association for Computing Machinery, New York, NY, USA, Article 4, 8 pages. https://doi.org/10.1145/3204493.3204531

[4] Deborah A Boehm-Davis, Wayne D Gray, Leonard Adelman, Sandra Marshall, and Robert Pozos. 2003. *Understanding and measuring cognitive workload: A coordinated multidisciplinary approach.* Technical Report. GEORGE MASON UNIV FAIRFAX VA DEPT OF PSYCHOLOGY.

[5] Efe Bozkir, David Geisler, and Enkelejda Kasneci. 2019. Assessment of Driver Attention during a Safety Critical Situation in VR to Generate VR-Based Training. In *ACM Symposium on Applied Perception 2019* (Barcelona, Spain). ACM, New York, NY, USA, Article 23, 5 pages. https://doi.org/10.1145/3343036.3343138

[6] Vera Demberg. 2013. Pupillometry: the index of cognitive activity in a dual-task study. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 35. eScholarship, USA, California, 2154–2159.

[7] Vera Demberg and Asad Sayeed. 2016. The frequency of rapid pupil dilations as a measure of linguistic processing difficulty. *PloS one* 11, 1 (01 2016), e0146194. https://doi.org/10.1371/journal.pone.0146194

[8] EyeLink. 2017. *The Eyelink 1000 Plus Eye Tracker.* https://www.sr-research.com/wp-content/uploads/2017/07/EyeLink-1000-Plus-Brochure-2016-November.pdf

[9] Stefan Frank and Robin Thompson. 2012. Early effects of word surprisal on pupil size during reading. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 34. eScholarship, USA, California, 1554–1559.

[10] Jukka Hyönä, Jorma Tommola, and Anna-Mari Alaja. 1995. Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *The Quarterly Journal of Experimental Psychology* 48, 3 (1995), 598–612. https://doi.org/10.1080/14640749508401407 arXiv:https://doi.org/10.1080/14640749508401407

[11] Marcel A Just and Patricia A Carpenter. 1993. The intensity dimension of thought: pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 47, 2 (1993), 310.

[12] Krzysztof Krejtz, Andrew T Duchowski, Anna Niedzielska, Cezary Biele, and Izabela Krejtz. 2018. Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PloS one* 13, 9 (2018), 1–23. https://doi.org/10.1371/journal.pone.0203629

[13] Sandra P Marshall. 2000. Method and apparatus for eye tracking and monitoring pupil dilation to evaluate cognitive activity. US Patent 6,090,051.

[14] Sandra P Marshall. 2002. The index of cognitive activity: Measuring cognitive workload. In *Proceedings of the IEEE 7th conference on Human Factors and Power Plants*. IEEE, IEEE, New York, NY, USA, 7–7.

[15] Oskar Palinko and Andrew L Kun. 2012. Exploring the effects of visual cognitive load and illumination on pupil diameter in driving simulators.. In *ETRA*. Association for Computing Machinery, New York, NY, USA, 413–416.

[16] Pauline van der Wel and Henk van Steenbergen. 2018. Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic bulletin & review* 25, 6 (2018), 2005–2015.

[17] Jorrig Vogels, Vera Demberg, and Jutta Kray. 2018. The Index of Cognitive Activity as a measure of cognitive processing load in dual task settings. *Frontiers in Psychology* 9 (2018), 2276.