Highlights

Toward an educative EEG-based neuroIIR system for adapting contents

Alejandro A. Torres-García, Fernando Martínez-Santiago, Arturo Montejo-Ráez, L. Alfonso Ureña-López

- A novel method for scoring reading fluency based on EEG signals is proposed.
- Students perform worse when the complexity of the text is higher.
- Differences in EEG responses related to texts of 2 complexities were found.
- We found that EEG signals are useful to profile a user's reading ability.
- EEG signals could be suitable for designing neuroTutors or neuroIIRs.

Toward an educative EEG-based neuroIIR system for adapting contents

Alejandro A. Torres-García[a], Fernando Martínez-Santiago[b], Arturo Montejo-Ráez[b], L. Alfonso Ureña-López[b]

[a] *Biosignals Processing and Medical Computing Laboratory. Instituto Nacional de Astrofísica Óptica y Electrónica, Luis Enrique Erro #1, Tonantzintla, 72840, Puebla, Mexico*
[b] *Universidad de Jaén, Las Lagunillas s/n, Jaén, 23071, Spain*

Abstract

The search for information is an integral part of a learner's daily routine, typically facilitated by information retrieval systems. In this study, we examine the relationship between text complexity and reading comprehension from the perspective of neuro information. We aim to determine if it is possible to infer the difficulty of a text for a learner based on their brain activity, as measured by an electroencephalogram (EEG), during the information search process. To achieve this, we applied a subtest of the Woodcock Reading Mastery Test to our participants to assess their average reading fluency. The 18 monitored participants then read paragraphs of varying complexity and answered questions related to the text. With the collected data, we trained a deep learning model, EEGNet, to automatically identify the complexity of the text being read based on the EEG signals. Our findings suggest that it is possible to classify the EEG signal according to the difficulty a student experiences in comprehending the text, getting an average accuracy of 80.53% for all subjects. Some directions to extend our work in real-time scenarios are suggested.

Keywords: interactive information retrieval (IIR), reading comprehension,

*Corresponding author at: Luis Enrique Erro # 1, Puebla, México 72840. Tel +52 222 2472940.

Email addresses: alejandro.torres@ccc.inaoep.mx (Alejandro A. Torres-García), dofer@ujaen.es (Fernando Martínez-Santiago), amontejo@ujaen.es (Arturo Montejo-Ráez), laurena@ujaen.es (L. Alfonso Ureña-López)

neuroscience, Brain-Computer Interfaces (BCI), Electroencephalograms (EEG), text complexity

1. Introduction

Today's students belonging to the so-called Generation Z and Generation Alpha are defined as digital natives, grown up in a world surrounded by technology. They manifest a different attitude towards technology, as it is natural for them and not something to be learned. Thus, the habits of students and people, in general, are changing due to the increasing availability of technology and the constant connectivity offered by the Internet. Now, a student can engage in any activity at any time of day, seven days a week. This has changed the way students study, their habits, and people's expectations regarding how they carry out their school, academic, or social activities.

In this context, the relationship of the student with information search is crucial, and it is here where Information Retrieval (IR) and Interactive Information Retrieval (IIR) are framed. The traditional IR task aims to retrieve documents for a given query, without the participation of the end user beyond the definition of that query. Therefore, IR is focused on the document, inheritance of the traditional Cranfield model [6]. This gap between the IR system and the end user has prevented better search engines generate higher satisfaction and performance of the users [40]. To solve this problem, Interactive Information Retrieval arises [21, 9]. Also known as Human-Computer Information Retrieval (HCIR) [27], IIR refers to the study and evaluation of the interaction of users with information retrieval systems, IR, and their satisfaction with the information retrieved.

Given the centrality of the user in IIR systems, explicit feedback mechanisms such as questionnaires and interviews are required to know about the relevance (utility) that a given document has for each user. However, an alternative to these mechanisms is to acquire information during the search process in a non-invasive way. A group of techniques that have the potential to be used as implicit feedback mechanisms of the user's state is the neurophysiological ones, such as functional magnetic resonance imaging (fMRI), functional near-infrared spectroscopy, or electroencephalography (EEG), magnetoencephalography (MEG), among others. The use of these types of techniques in the field of IIR has come to be referred to as neuroIIR. Some examples of neuroIIR studies based on fMRI can be consulted

in [32, 19, 31, 33], and those based on EEG signals will be approached in Section 2.2.1. Last, the EEG technique has several advantages such as low cost, ease of use, and wearability. Even EEG, compared to other techniques, is the easiest one for directly monitoring the brain, where processes associated with emotions and attention are carried out. This is of interest in understanding the user during their interaction with IIR systems.

Basically, EEG measures brain electrical activity through an array of electrodes placed on the person's scalp, based on the 10-20 international system. It also has an excellent temporal resolution to detect changes in brain activity at the moment. It is also composed of 5 fundamental rhythms or waves: alpha (8-12 Hz), beta (12-32 Hz), theta (4-8 Hz), delta (0-4 Hz), and gamma (32-60 Hz).

Most works merging the fields of information retrieval and EEG-based Brain-Computer Interfaces (BCI) were focused on determining the level of relevance of a document. These have also analyzed the information obtained with eye trackers [11]. However, a neuroIIR should take advantage of neurophysiological signals coming directly from brain activity, such as EEG and the above-mentioned techniques.

This is where the present work fits in, addressing the task within the field of neuroIIR. We propose the use of EEG signals to know the complexity perceived by the student when reading a text. That an IIR system has such knowledge of the user allows the system's behavior to be adjusted to each user. For example, the IIR system may eventually adjust the search process, considering not only the relevance of the document, as is usual in IR, but also the complexity of the text and the student's own reading fluency. However, this work is limited to the study of the relationship between textual complexity and reading comprehension through the study of neurophysiological signals, specifically EEG.

Despite a related issue has been approached using only eye tracker signals in [36], aiming to identify differences in text difficulty and between native and non-native readers; only two works [28, 1] have approached it using only EEG signals. However, their average outcomes were about the chance level. A detailed description and a discussion of both works are presented in Sections 2.2.1 and 2.2.2.

In this work, the data set and the recording protocol designed to characterize the brain responses (measured with EEG) are presented in detail. Also, it has been evaluated if a deep learning technique, such as EEGNet, is capable to discriminate between these two reading complexities (easy or

defficult) from the EEG signals of a given user.

In summary, the following research questions are sought to be answered:

- Does the type of complexity of the texts affect EEG signals from the user's recorded brain activity?
- Is it possible to train a supervised machine learning model from these EEG signals so that the complexity of the texts can be inferred automatically?

If such a correlation were to be confirmed, this would be a first step towards a neuroIIR/neuroTutor system that allows students to adapt the content to be studied based on the complexity of the text and their own reading fluency. With this system, a personalized learning experience would be sought for each student, helping them to better understand a given content. In that case, the neuroIIR system should present the content to this student.

The rest of the article is organized as follows. Next, areas that are particularly related to the work presented are described, that is, reading comprehension in general and information retrieval in particular, as well as EEG applications in the field of brain-computer interfaces. In Section 3, the materials and methods that make up the experimental framework are presented. In Section 4, the results obtained are presented and discussed. Finally, some conclusions are presented, and we outline what we consider to be the next steps in the framework of the experiment presented, aimed at achieving stronger evidence of the hypotheses posed.

2. Related Work

This work proposes the use of EEG to understand the user's mental state when searching for information by reading related documents. This is influenced by the relationship between reading comprehension and textual complexity, interactive information retrieval, and its relationship with neuropsychology, as introduced below.

2.1. Interactive Information Retrieval

Information Retrieval is the process of obtaining relevant documents for a given user's information need, typically in the form of a query. However, the relevance of a document is not an objective, intrinsic quality of the document,

but is subjective. Even when only expert annotators are considered, there is a lack of consensus [3].

In light of this, the usual way in which each user decides the relevance or usability of the retrieved documents is by browsing and opening some of them and eventually refining the query. This would lead to an iterative process until the information search is completed. This process, which is defined as Interactive Information Retrieval, is inherently interactive and subjective [37], so the task of information search revolves around the individual and not the document and its relevance.

Intensive research has been conducted on the evaluation of IIR systems [4] and the study of the interaction of these systems with the user. Early IR works showed that readability could bias the IIR process [2]. However, readability is usually considered only in the text's features, rather than the user's reading capabilities, such as grammatical or lexical skills in the target language, for example. Our work focuses on that aspect of the interactive retrieval model, exploring how reading fluency influences the performance of the search process.

Reading comprehension is a complex process that relies on the development of various cognitive functions. Various studies have shown that reading efficiency is related to the ability to decode visual stimuli, naming speed, vocabulary breadth, operational memory capacity, and ability to maintain attention and concentration [10, 20]. In the field of neuropsychology, different cognitive models of the reading process have been proposed that rely on processing information through different brain structures [38]. Such information processing requires a series of cognitive skills such as attention, memory, language, and abstraction. Reading comprehension has been found to be a complex mental process, involving many factors such as language knowledge, attention and interest in the subject of reading, or reading fluency. This work focuses particularly on the impact of text complexity. That is, given a set of users who exhibit similar reading skills, how does text complexity affect the individual's brain activity monitored by EEG.

The quantitative characteristics of text complexity are those that can be counted or quantified: sentence length, number of syllables, word length, word frequency, perplexity, and other characteristics that can be calculated on a computer [25, 39]. The subjective characteristics of a text are the aspects and nuances of it that cannot be measured with a simple formula. In this case, the complexity of the text is evaluated by a set of human experts. An example of this is the Newsela corpus [44], which will be used in our exper-

iments and further detailed in Section 3. This work focuses on the reader's comprehension incidence based on the subjective difficulty of the text, as defined in the Newsela collection, with the quantitative characteristics of the text outside the scope of the experiment.

2.2. Brain-Computer Interfaces

Our major goal is to assess if a machine learning model with a supervised approach can identify differences in the EEG signals based on the complexity level of a given text. A neuroIIR system (like ours) focused on studying users' brain activity during information searching is also part of the brain-computer interfaces research area. BCIs are systems that try to provide a person with a new non-muscular communication channel for transmitting messages and commands to the exterior world by recording his/her brain activity [42].

In the next section, we present the most related works to ours, that is, works following a machine-learning-based strategy for approaching the automatic classification of EEG responses recorded during search-related tasks.

2.2.1. EEG-based BCIs for NeuroIIR

In recent years, EEG-based NeuroIIR systems have attracted great interest from researchers. For example, a special issue on Neuro-Information Science [12] was published by the Journal of the Association for Information Science & Technology in 2019. Also, two workshops have been carried out entitled NeuroIIR [13] and NeuroIR [1].

In the special issue, despite three studies analyzed EEG signals [17, 22, 43], only two followed a machine learning approach [17, 22]. The first work, described in [17], studied whether the assessment of implicit relevance judgments can be improved by interpreting users' EEG responses to search results. From the EEG signals and eye movements, their model generated feedback of relevance in real-time. The second work [22] analysed if users consider segments of video relevant to a given topic or not. For making the judgments of relevance, their classification stage was focused on detecting changes in N400 and P600 peaks. The model overcame image-colour-histograms-based algorithms.

For the above-mentioned conferences, only three works [24, 8, 14] approached the use of neuroimaging techniques with different purposes and

[1]The full list of papers can be consulted in `https://sites.google.com/view/neuroiir`

outcomes. In the first work [24], emotion recognition was analyzed. The second one [8] aimed to predict the relevance of texts. Whereas the third one [14] collected and described a dataset called NAILS, which is composed of labeled images based on the neural activity recorded using EEG.

In [15] it was introduced the dataset called ZuCo 2.0, which was recorded using an EEG device and an eye-tracker on 19 subjects (but only 18 were analyzed) read sentences in a normal reading or a task-specific reading. For task-specific reading, the subjects had to find any specific information in the text. This dataset was recorded during the reading of 739 sentences extracted from the Wikipedia corpus generated in [7]. Each participant had to read 349 in normal reading and 390 in task-specific reading. Subjects read the sentences of the two tasks at their own speed. After that, they had to answer some control questions. During normal reading, subjects had to read only for reading comprehension. Whereas for task-specific reading, the subjects had to find any specific information in the text.

Using the ZuCo 2.0 dataset, in [16], a cross-subject classification strategy was evaluated for discriminating between the two reading-related tasks. Specifically, different features were computed from eye-tracker and EEG signals. The best average accuracies were gotten using only eye-tracking features, and eye-tracking along with EEG mean features with 69% and 68%, respectively. An interesting set of features computed with Bidirectional Encoder Representations from Transformers (BERT) was used as a text-based baseline, obtaining an accuracy of 65%.

In [41], an online neuroIIR system based on EEG and eye-tracker signals was developed, aiming to identify subjective relevance. For this, downsampled EEG signals (from 1000 Hz to 20Hz) were used as features and linear discriminant analysis was applied for the classification stage.

In [28] was presented the first work focused on finding differences in EEG responses related to the reading of difficult and easy texts from the NEWSELA corpus. Also, this work measured the reading fluency of the subjects with the Woodcock Reading Mastery Test. As to the EEG signal processing, these were characterized by temporal and frequency features. A random forest with 50 trees was applied for classifying the features. Finally, the average classification outcome obtained was an AUC of 0.45.

Another investigation, described in [1], assessed whether both different readability difficulties and texts read at different presentation speeds can be distinguished by a model trained for mental workload detection (low and high workload) based on EEG signals. They found the average predictive

values were higher for difficult texts compared to easy texts, with 0.618 and 0.553 respectively.

2.2.2. Summary on EEG-based BCIs for NeuroIIR

In short, most of the works only approached the identification of relevant/irrelevant topics. Another issue that has also been approached is the identification of differences between normal and task-specific reading. Nevertheless, the task of classifying the EEG responses to different levels of text complexity has been approached only in [28] and [1]. In [1], a relevant factor as the reading fluency of the subjects was not analyzed. Besides, the global performances were not presented but the predictive values for each class of texts' difficulty. Whereas in [28], the machine-learning-based classification of the EEG signals was not the main aim of the work. Finally, both works classified the EEG signals using traditional machine learning algorithms such as LDA [1] and random forest [28]. Accordingly, the question of if the outcomes obtained are more related to the complexity of the task compared to the machine learning strategy arises. For this reason, in this work, we will assess if the outcomes can be improved with a novel machine-learning strategy (based on a deep-learning method). Furthermore, since the work described in [28] measured the reading fluency of the subjects, we will compare the present work with this.

3. Materials and methods

The main elements of the experimentation framework are described in this section, including those taken from the area of reading comprehension, neurosciences, and, particularly, EEG-based BCIs.

3.1. Woodcock-Muñoz reading fluency test

A group of volunteer participants in this experiment was given the Woodcock-Muñoz reading fluency test [34]. The aim is to start with a set of individuals who have average reading comprehension. It should be remembered that we are interested in knowing the impact of text complexity on their comprehension and how it affects the EEG signal recording. A text being "easy" or "difficult" depends to a large extent on the individual's reading comprehension, and that is why it is important to ensure that the population being tested has similar reading comprehension, mitigating the bias that extreme cases could introduce.

For this purpose, subtest 2 of the Spanish version of the Woodcock-Muñoz battery was selected. This subtest consists of 105 sentences that can be true or false (for example, "You can find birds in the field" versus "Dogs are flying animals"). Participants have to read each sentence silently for three minutes. The difficulty of this task is related to both the speed and accuracy and the meaning of the sentence. The sentences are getting longer and progressively more difficult. The number of sentences answered correctly in those three minutes is scored. It is a particularly suitable test for the defined experimental framework, in which participants have a limited time window to read each text.

This test was completed by 42 participants (31 men, 11 women, average age = 22.4, std = 3.7), all Spanish and, therefore, with Spanish as their native language. They are university students recruited from three different degrees (psychology, computer engineering, and electrical engineering). Five of these participants were not considered because they showed abnormally low values in the Woodcock-Muñoz reading fluency test (see 1). Of the remaining 37, 18 subjects gave their consent to have their brain activity recorded, on whom the complete experiment was carried out, as described in section 4. On average, 44.76 out of 65 questions are answered correctly (68.86%, s = 17.15), 1.02 questions are answered incorrectly (2%, s = 0.93), and 19.22 questions are not answered (29.1%, s = 17.08). Note that the goal is to study the impact of text complexity. Therefore, it is advisable for users to be at a similar level of reading fluency. In other words, the difference in reading comprehension based on text complexity is studied in a population that has comparable reading fluency.

3.2. Corpus Newsela

This corpus distinguishes between easy and complex texts through subjective difficulty, evaluated from the text so that difficult expressions have been annotated with a difficulty level between 1 and 4 by human experts. Newsela is available for research on textual difficulty, among other disciplines. This corpus includes thousands of articles, both in English and Spanish, of news professionally adapted for different reading complexities. It consists of a total of 1,130 news articles distributed across 11 grades or courses. Each article has four different versions, according to the different grade levels, and produced by Newsela editors, a company specializing in reading materials for use in pre-university classrooms. Thus, the corpus consists of five different subsets: Original, Simp-1, Simp-2, Simp-3, and Simp-4. Some statistics

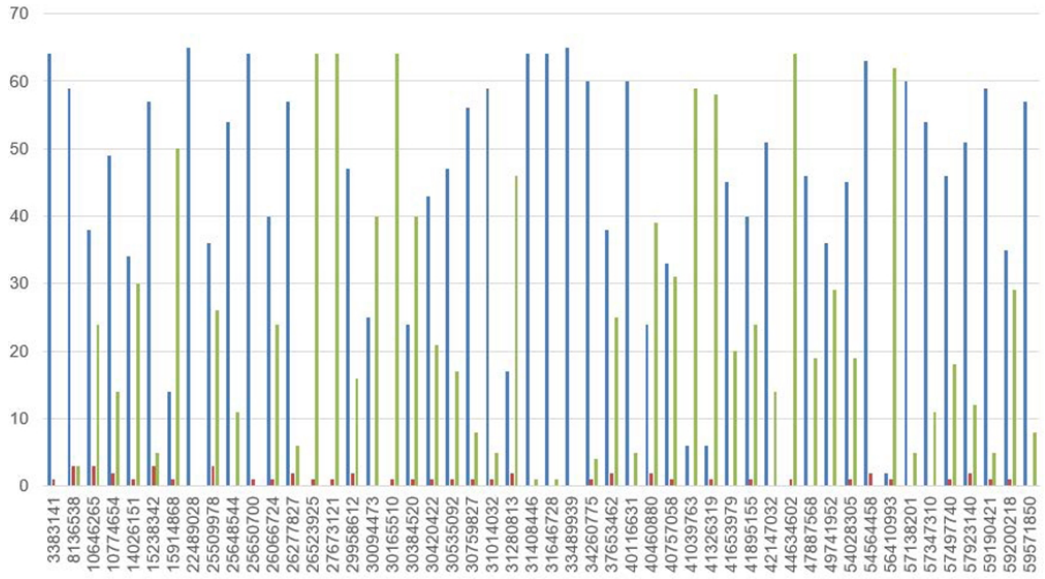

Figure 1: Results of the Woodcock-Muñoz Reading Fluency Test. For each participant, the blue, red, and green columns correspond to the number of correct, incorrect, and unanswered questions, respectively.

about the corpus are shown in Figures 1 and 2 and in Tables 1 and 2. Note that the "Total" column refers to the total of all grades in the corpus, not just those in the table.

A set of paragraphs from the Newsela corpus [44] were presented to the subjects in order to create the EEG signals corpus recorded during reading texts with different complexity. Although the Newsela corpus has texts with several levels of reading complexity (levels 1 to 4 labeled by experts), only texts with easy and difficult levels (levels 1 and 4) were used for the creation of the EEG corpus.

3.3. EEG dataset

For recording the EEG dataset [28], a 14-channel wireless Emotiv EPOC was used, which has a sampling rate of 128 Hz. The channel names of the 10-20 international system are AF3, AF4, F3, F4, F7, F8, FC5, FC6, P7, P8, T7, T8, O1, and O2. EPOC also includes two references (P3/CMS, P4/DRL) and a gyroscope to detect movements along the axes x and y. Nonetheless, in this work, we only studied brain activity.

Table 1: Newsela corpus characterization (academic grades 1-6)

Grade	2	3	4	5	6	Total
Number of texts	59	116	161	146	113	628
Average vocabulary size	110.00	124.46	188.40	210.98	256.51	338.90
Average document size	316.56	378.47	577.04	646.98	768.33	1007.96
Shortest document	203	235	337	240	487	288
Largest document	645	923	1296	1296	1669	1249
Avg. sentence length	9.19	11.04	12.78	15.66	18.38	13.41
Avg. complex sentences	5.54	6.75	8.51	10.89	13.59	20.56

Table 2: Newsela corpus characterization (academic grades 7-10 and 12)

Grade	7	8	9	10	12	Total
Number of texts	155	115	112	1	245	595
Average vocabulary size	278.37	314.70	300.39	425.00	376.08	177.60
Average document size	832.39	930.73	886.84	1249.00	1140.83	537.88
Shortest document	288	466	315	1249	296	203
Largest document	1969	2043	1208	1249	2923	1669
Avg. sentence length	21.32	23.74	27.55	29.05	26.24	24.68
Avg. complex sentences	16.13	18.84	22.39	25.00	20.93	9.06

The dataset[2] is composed of the EEG signals of 18 subjects. For each subject, a set of 40 epochs (instances) was recorded during the inner reading of texts with two levels of complexity (easy and hard) based on their labels in NEWSELA. For each level of complexity, a set of 20 epochs was recorded, that is, 20 texts labeled as hard (NEWSELA's level 4) and 20 texts labeled as easy (NEWSELA's level 1). In Section 3.3.1 are described more details about the protocol designed to record subjects' brain activity. Meantime, a preprocessing scheme, explained in Section 3.3.2, was applied to the signals looking to reduce the impact of noise and artifacts.

Last, in Table 3 are shown four statistical values (mean, standard deviation, maximum and minimum) to describe the durations of the epochs of each subject. It is important to highlight that 128 samples are equal to 1 second. In addition, in Table 3, a variation in duration can be observed, which is explained mainly by the reading speed of the subjects and the protocol designed in [28]. Since the relevance of the experimental protocol for the present work, we briefly present it in Section 3.3.1.

3.3.1. Protocol for recording the EEG signals

A protocol for recording EEG signals during the interaction with an IIR system has been designed and implemented, which aims to stimulate the subjects with documents whose complexity of reading is simple or difficult, to be able to identify changes in their brain state depending on the type of reading complexity presented. At the end of this stage, there will be one dataset of EEG signals recorded during user interaction with an IIR system.

The recording protocol of the EEG signals can be seen in Figure 2. In the beginning, a fixation cross was presented to indicate the beginning of the record of an epoch (example or instance) of the task in question. Subsequently, 2 seconds after the start, the subjects heard a beep to indicate that they paid attention since the main stimulus of the experiment was about to appear. 3 seconds after the start, a paragraph with variable complexity of reading (easy or difficult) that had to be read internally by the subjects was shown. Subsequently, the subjects had to press a key to indicate that they had finished reading the paragraph. From this, it is observed that the dura-

[2]Another scheme of classification can be analyzed, which is based on relabeling the epochs according to the questionnaires answered by the subjects in each epoch and the mean of right answers of the subjects in each epoch (see Appendix Appendix A for more details).

Table 3: Mean, standard deviation, maximum and minimum of the number of samples for each subject's epochs

subj	mean	std	max	min
S1	5792.23	4.55	5805	5780
S2	4082.22	825.38	5724	2494
S3	4064.90	842.39	5724	2496
S4	1775.13	311.32	2297	1192
S5	1769.70	314.13	2302	1136
S6	2186.88	535.81	3494	1123
S7	1929.35	413.97	2884	1179
S8	2292.68	480.43	3410	1220
S9	2303.50	570.22	3742	1456
S10	1891.35	439.34	2965	1027
S11	1819.40	375.22	2501	1096
S12	2140.05	501.85	3057	1193
S13	1727.67	390.60	2397	750
S14	1801.83	388.29	2717	1179
S15	2310.15	510.09	3494	1290
S16	2587.47	576.71	3438	1581
S17	1065.20	201.28	1581	680
S18	2173.65	494.84	3423	1399

tion of the reading will be variable for each subject. Likewise, the stimulation program had a predefined duration for changing this window, which was not activated, defined as the number of words per 0.5 s. This aimed to prevent the subjects from being distracted from the task at hand. After reading the paragraphs, the subjects were presented with a questionnaire with 3 true/-false questions with no time limit. In addition, to select only those questions where the subjects' answers were clear, there was the possibility of leaving any of them unanswered. Finally, a black window was shown to indicate to the subjects a pause lasting one second and the end of the recording of an epoch.

Some examples of the types of paragraphs that were shown to the subjects can be seen in Figures 3 and 4. An example of a text with a level of reading complexity difficult is shown in Figure 3. While in Figure 4 we can see a text with an easy level of complexity.

3.3.2. EEG signal preprocessing

As above mentioned, the EEG signals of the dataset were pre-processed with an automatic artifact removal method. An algorithm based on inde-

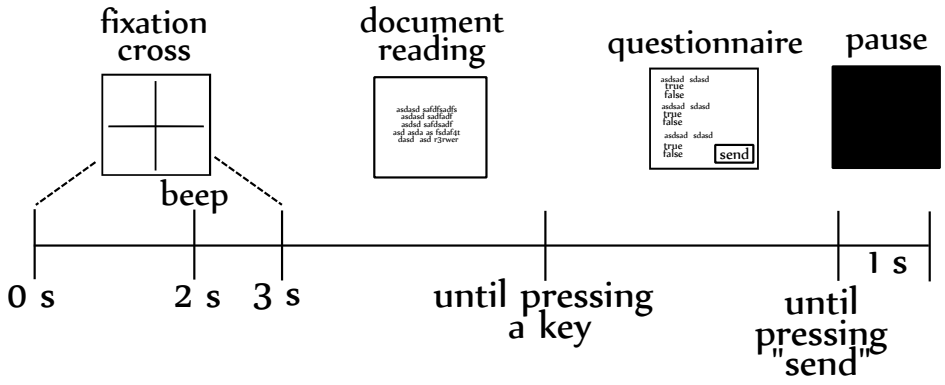

Figure 2: Timeline of the recording protocol

pendent component analysis (ICA) called ADJUST [29] was applied. This algorithm was chosen because it usually performs well in removing signals associated with flickering, eye movement, and generic discontinuities. Furthermore, the EEG signals were filtered using a 5th-order Butterworth bandpass filter (1-50 Hz). Finally, common average reference (CAR) was also used to eliminate that activity that is common to all channels at a given instant of time.

3.4. Supervised learning

The ultimate goal of the neuroIIR/neuroTutor system outlined in this work is to automatically provide a student with the content that fits him/her better. For this reason, our work will focus in this first stage on the automatic identification of differences in EEG signals based on the inner reading of either difficult or easy texts. Hence, the approach will be a supervised learning type, in which a learning algorithm will beforehand know the relevant segments of EEG signals and what complexity of the text was read by the subject.

A supervised learning algorithm attempts to generate a model that can infer a function from the training data. This function will allow the mapping of a set of training instances, described with a set of values or characteristics, to their corresponding class. Once the model is trained, this is tested with instances not used during the training process to evaluate if the model can be generalized to these new cases [18].

Even though there is a wide range of algorithms to perform the supervised classification task (such as Random Forest or Support Vector Machines

Muchos inmigrantes a los EEUU pagan cientos de dólares por un servicio que les ayude a validar sus credenciales educativas. Algunos de esos servicios son tan solo una "estafa", mientras que otros son auténticos, pero aun en el segundo de los casos, cada estado acepta documentación de un servicio de validación específico.
(Many immigrants to the US pay hundreds of dollars for a service to help validate their educational credentials. Some of these services are just a scam, while others are legitimate, but even in the latter case, each state accepts documentation from a specific validation service.)

Q1: Algunos servicios de validación de estudios son un fraude.
Q2: Es frecuente que los inmigrantes contraten servicios que le facilitan los papeles necesarios para residir en el país.
Q3: Cada estado tiene sus propios mecanismos de validación educativa.

(Q1: Some educational validation services are a scam.
Q2: It's common for immigrants to hire services that provide them with the necessary documents to reside in the country.
Q3: Each state has its own educational validation mechanisms)

Figure 3: Example of a document with high text complexity and its corresponding questionnaire.

(SVM), among others), deep learning technologies have become the most used and promising ones in recent years for approaching almost all areas of Artificial Intelligence: Natural Language Processing [35], Robotics [30], Computer Vision [5] among others. These networks can learn latent encodings from unstructured data. In the case of Convolutional Neural Networks (CNNs), several transformations are applied to data matrices. The model itself learns these transformations (convolutions) through a training process.

In our research, a convolutional deep neural network type, EEGNet, was selected. This network has got good results for the most used neuro paradigms in BCI, such as motor imagery and P300 signals [23].

3.4.1. EEGNet

For analyzing whether a deep learning-based computational learning algorithm can discriminate between the EEG signals recorded while reading the two levels of complexity of texts, the algorithm called EEGNet was evalu-

Como Nueva York es una ciudad de apartamentos pequeños, la gente suele tirar las cosas para tener más espacio. Algunas personas se llevan estas cosas a su casa. Otros buscan comida, como bagels, arroz y pasta. Robin Nagle es profesora de la Universidad de Nueva York que estudia las cosas que tira la gente. Nagle ha hecho todo lo posible para que los fabricantes usen más materiales reciclables.
(As New York is a city of small apartments, people tend to throw things away to create more space. Some people take these things to their homes. Others search for food, like bagels, rice, and pasta. Robin Nagle is a professor at New York University who studies the things people throw away. Nagle has done everything possible to encourage manufacturers to use more recyclable materials.)

Q1: Algunas personas aprovechan lo que otros tiran.
Q2: En Nueva York la gente tira a la basura cosas porque están viejas.
Q3: Robin Nagle lucha por concienciar a la ciudadanía para que recicle.

(Q1: Some people take advantage of what others throw away.
Q2: In New York, people throw things away because they are old.
Q3: Robin Nagle fights to raise awareness to recycle.)

Figure 4: Example of a document with low text complexity and its corresponding questionnaire.

ated. EEGNet, proposed in [23], is a compact convolutional network specially designed for tasks related to EEG-based BCIs. EEGNet has the advantage of not having to define in advance a feature extraction technique, which is usually dependent on the selected neuro paradigm. In fact, the EEGNet input is the EEG signals without any transformation, which is why these algorithms are also referred to as *end-to-end* approaches. Furthermore, EEGNet is robust to a small size of the training datasets, which is often a common factor in BCI-related data. The latter also highlights the selection of EEGNet with respect to other deep learning models, since the performances of the latter usually depend on a large amount of data, which is not common in BCI applications (as in the current work) [26].

As mentioned above, the input to EEGNet is the EEG signals. From them and the application of a temporal convolution, a set of frequency filters is learned. Next, and taking as input the feature maps obtained in the previ-

ous stage, a depth convolution is applied to learn a set of frequency-specific spatial filters. Subsequently, EEGNet learns a temporal summary of the feature maps individually and how to mix the feature maps optimally, using a separable convolution that groups both a depth and a point convolution. Figure 5 outlines the architecture of EEGNet graphically.

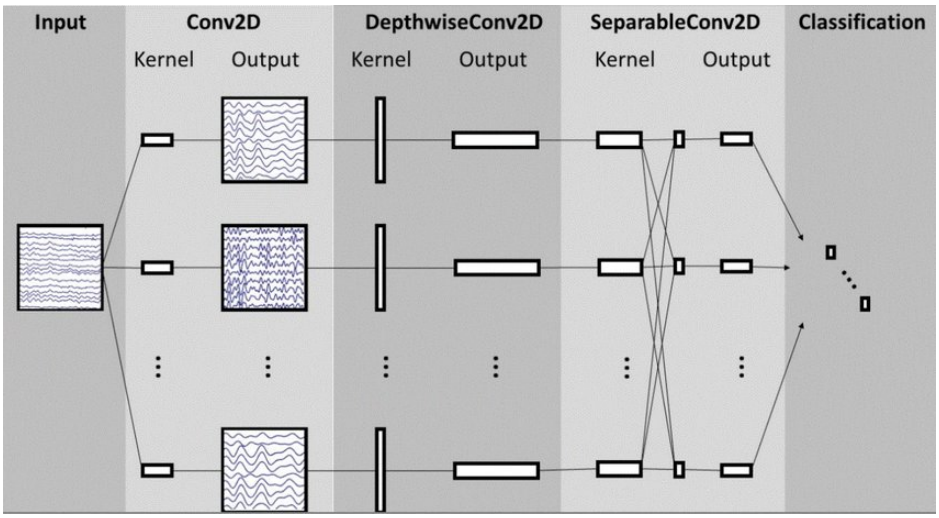

Figure 5: EEGNet architecture (taken from [23])

4. Experiments and results

In this section, we describe the experiments carried out to assess if a machine learning algorithm can detect differences in the EEG signals of students when they read easy and difficult texts. This would suggest that students' signals can reflect how difficult the students perceive a text. Before this, both the reading comprehension of the group of students in the study and changes in reading comprehension depending on the two complexity levels of the texts (taken from NEWSELA) are validated. For the first validation, a group of students with a similar average reading comprehension ability was selected, as described in Section 3.1. Whereas for the second validation, students had to carry out an information search task through a simulated IIR system made up of a small pre-defined set of "easy" and "difficult" documents (most of them belonging to NEWSELA). Later, 18 of the 42 subjects of the first validation were recorded with EEG headsets while they read easy or

difficult documents, different from those used in the IIR task, but belonging to the same document collection. After these validations, evidence related to worst reading performances was found when text complexity is high. This suggests that these changes could be detected from EEG signals.

4.1. Information Retrieval task

The IIR system provides a total of 368 documents that relate to one or more of 20 predefined queries. That is, it is not a real IIR system, since the learner is not free to make up the queries, nor is the document base on which to search for information extensive. The aim is to narrow down the system as much as possible so that all participants have to judge exactly the same documents using the same queries, and also to make them used to the IIR environment chosen. An example of these 20 queries is shown in Figure 6. For each query, the student has to (i) execute the given query, (ii) open and eventually read some documents from the list of documents obtained as a result of the query execution, (iii) submit a summary of his findings with respect to the search task performed, and (iv) judge the relevance of a set of sampled results for each topic he has developed during the search session.

The sources of the document set are the Newsela corpus and the Web. From Newsela, we selected those documents whose topic and/or content is related in some way to at least one query. In addition, the difficulty of the selected documents corresponds to the lowest (1) or highest (4) level of complexity. Since the number of documents obtained with this method is low (4.8 documents per query on average, s = 3.17), we searched the Web for documents related to each query until we obtained 20 documents related, relevant or not, to each query. Once the set of documents has been defined, the next step is the creation of relevance assessments. Thus, each query-document pair is judged by three human experts, reaching an inter-index agreement (kappa value) of K = 0.83. This results in 225 positive document relevance judgments for all queries, compared to 185 negative ones. On average, for each query, there are 13.85 documents marked as relevant to that query (s=0.9), compared to 11.6 documents on average marked as not relevant (s=1.2).

Finally, user performance is measured in terms of precision and coverage, that is, respectively, the proportion of documents of the documents judged as relevant by the learner how many actually are, and what proportion of the total relevant documents have been found. These measures are interpreted as a measure of the agreement between participants and experts in

```
{
      "lang": "ES",
     "num": " 001 ",
     "title": " Factores que inciden positivamente en la generación de riqueza
               en los países ",
     "desc": " Motivos sociales y/o políticos que hacen de un pais una
               sociedad más rica. Quedan excluidos aquellos artículos que se
               centren en los recursos naturales del pais",
     "auth": " Fernando Martínez"
}
```


(Factors that obtain a positive impact on the generation of wealth in countries. Social and/or political reasons that make a country a richer society. Articles that focus on the country's natural resources are excluded.)

Figure 6: Example of query

the task of judging documents in relation to a given query. The best results were obtained when considering Newsela-easy, achieving a precision of 0.82 and coverage of 0.90 (F1=0.43). In contrast, the results filtered by Newsela-difficult yielded a precision of 0.78 and a coverage of 0.83 (F1=0.40). Thus, we find some evidence that indeed students perform worse when the complexity of the text is higher. The question we try to answer in the next section is whether this greater difficulty is detectable by using deep neural networks from the EEG recordings of the student when he/she performs reading comprehension tasks.

4.2. Classification of EEG responses to the level of text complexity

For the experimentation, the EEGNet deep learning algorithm was studied to classify the EEG signals of the subjects obtained during text reading with 2 different complexities: 20 with an easy level and 20 with a difficult level. Furthermore, only the segments during which people read the paragraphs were used as epochs for the analysis. Also, since both the length of the paragraphs and the reading speed of people were different, this resulted in epochs with different lengths (see Table 3). Therefore, based on this distribution and looking for analyzing the same epoch size, we decided to analyze the intermediate part of them, particularly the 3 intermediate seconds of each one.

EEGNet was evaluated using accuracy, since this data set is balanced. This measure was obtained for each subject by applying cross-validation with 10 partitions. Repeatedly, a partition is reserved to evaluate the model and the rest to train it, and so on until 10 accuracy evaluations are obtained, which are averaged to calculate the final performance. This results in a more robust way of evaluating and is less dependent on the luck related to the creation of evaluation and training sets. Summing up, the process is done for each subject separately to obtain the final performance of the model for each one. This subject-dependent evaluation was chosen because the neuroIIR/NeuroTutor system will be for personal use.

The EEGNet algorithm was trained for 30 epochs (iterations over the full set in this case, not to confuse with the EEG recorded epochs). No early-stopping is applied, so the final trained model is obtained after epoch 30. Also, the model was fitted using the Adam optimizer. Likewise, the 14 EEG channels were analyzed simultaneously. In addition, since the signal frequency sampling is 128 Hz, the EEGNet kernel length parameter was set to 64, that is, half of the frequency sampling, as suggested by [23]. The model was trained on a high-performance computing system based on Intel Xeon Silver 4208 processors along with NVIDIA Tesla V100, NVIDIA Tesla A100, and NVIDIA GeForce RTX 2080Ti graphics cards.

Table 4 shows the results obtained for the recognition of the two text complexities analyzed from the EEG signals. It is interesting to note that, for all subjects, EEGNet obtained randomly higher accuracies for two classes, and its average performance was 80.83%. Particularly, the model achieved performances of accuracy between 67.5% to 90% for subjects S14 and S9, respectively. Likewise, for comparison, the results obtained using random forest with 50 trees and a characterization based on temporal and frequency features obtained after applying discrete wavelet transform [28] are presented. The average results for both classification models show that EEGNet outperforms random forest.

Also, a sign test was applied to verify the statistical significance of this outcome. This test was chosen due to the data's box plot showed a non-parametric and non-symmetric distribution. After its application, the difference between both methods was statistically significant ($p \approx 7.63e^{-6}$, $\alpha = 0.05$).

Table 4: Percentages of accuracy (± std.) for classification of the levels of complexity of the texts from EEG signals of each subject obtained by EEGNet and random forest

	EEGNet		random Forest	
subj	acc	std	acc	std
S1	85.00	20.00	70.00	10.00
S2	77.50	20.80	40.00	25.49
S3	80.00	18.70	52.50	17.50
S4	82.50	22.50	47.50	13.46
S5	77.50	23.60	50.00	15.81
S6	77.50	23.60	55.00	15.00
S7	80.00	15.00	57.50	11.45
S8	80.00	26.90	55.00	10.00
S9	90.00	16.60	57.50	16.00
S10	85.00	12.20	40.00	22.91
S11	75.00	19.40	45.00	15.00
S12	87.50	16.80	52.50	20.76
S13	82.50	19.50	40.00	12.24
S14	67.50	25.10	50.00	15.81
S15	82.50	19.50	52.50	13.46
S16	85.00	16.60	42.50	16.00
S17	75.00	22.40	45.00	31.22
S18	85.00	20.00	52.50	13.46
mean	80.83	19.96	50.28	16.42
std	5.36	3.71	7.66	5.53

5. Discussion and conclusions

In this work, the design of a neuroTutor/neuroIIR system to support students during the search processes is proposed. In particular, we tested whether a computational learning model can find differences between EEG signals during the reading of easy and difficult texts (labeled a priori in the NEWSELA corpus). To distill as much as possible the causal relationship between textual complexity and its incidence in the recording of EEG signals, other variables that may have an incidence in the process were fixed, particularly the reading ability of the participants and the difficulty of the texts. For the first, the Woodcock-Muñoz reading fluency test was applied to the participants, selecting only those who got a value close to the average.

Both the difficulty of the texts and their incidence during the search for information were validated through an experiment that simulates an information search task, finding some evidence of better results when the texts judged are easy.

With all this, the results obtained for all the subjects were above the level of chance for the two classes, which is promising as a first experiment and allows conjecturing that the automatic adaptation of a tutor and/or IR system is feasible from the use of EEG signals. Furthermore, a comparison with previous work was presented, which used both a different characterization and random forest for classification. The observed result was that the current method (EEGNet) outperformed the previous with an accuracy of 80.83% compared to 50.28% for random forest.

It would be interesting to increase the size of the dataset, both in the number of subjects and in the number of instances. This could help to reduce the standard deviation obtained in the model. For the same purpose, the generation of artificial instances could be useful. Specifically, some methods that could be evaluated are generative adversarial networks (GANs) or combinations of the instances in the time and time-frequency domains.

It could also be evaluated if there is any improvement between analyzing the initial and final segments with respect to the intermediate one analyzed in this work. In addition, a detailed analysis of what would be the most optimal window size to identify the complexity of the texts from the EEG signals would be desirable.

Finally, in the search for implementing an EEG-based neuroIIR/neuroTutor system, it would be appropriate to evaluate if the user's mood affects the system and how it impacts the search process, and in particular the identification of easy and difficult texts. It would also be interesting to effectively determine the level of attention paid to a text document.

Declaration of Competing Interest

We, the authors, declare that we have no known competing financial interests or personal relationships that could have appeared to influence the present work.

Funding

This work was partially financed by the CEATIC of the Universidad de Jaén, Spain through the "Premios de Invitación de Movilidad CEATIC.

Jóvenes Doctores" in 2018 and 2022. This work is also partially funded by the WeLee project (grant 1380939, FEDER Andalucía 2014-2020) of the Junta de Andalucía.

References

[1] Andreessen, L. M., Gerjets, P., Meurers, D., and Zander, T. O. (2021). Toward neuroadaptive support technologies for improving digital reading: a passive bci-based assessment of mental workload imposed by text difficulty and presentation speed during reading. *User Modeling and User-Adapted Interaction*, 31:75–104.

[2] Belkin, N., Chaleva, I., Cole, M., Li, Y. L., Liu, L., Liu, Y., Muresan, G., Smith, C. L., Sun, Y., Yuan, X., and Zhang, X. (2004). Rutgers' hard track experiences at trec 2004. In *Proceedings of the Text REtrieval Conference 2004 (TREC)*. NIST.

[3] Borlund, P. (2003a). The concept of relevance in ir. *Journal of the American Society for information Science and Technology*, 54(10):913–925.

[4] Borlund, P. (2003b). The iir evaluation model: a framework for evaluation of interactive information retrieval systems. *Information research*, 8(3):8–3.

[5] Chai, J., Zeng, H., Li, A., and Ngai, E. W. (2021). Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, 6:100134.

[6] Cleverdon, C., Mills, J., and Keen, M. (1966). Factors determining the performance of indexing systems volume 1. design.

[7] 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 NAACL, Main Conference*, pages 296–303.

[8] Eugster, M., Ruotsalo, T., Spape, M., Kosunen, I. J., de Bellegarde, O. B. M., Ravaja, J. N., Jacucci, G., and Kaski, S. J. I. (2015). Predicting relevance of text from neuro-physiology. In *SIGIR 2015 Workshop on Neuro-Physiological Methods in IR Research (NeuroIR 2015)*.

[9] Fidel, R. (2012). *Human information interaction: An ecological approach to information behavior*. Mit Press.

[10] Gough, P. B. (1984). Word recognition. *Handbook of reading research*, 1:225–253.

[11] Gwizdka, J. and Dillon, A. (2020). Eye-tracking as a method for enhancing research on information search. In *Understanding and Improving Information Search*, pages 161–181. Springer.

[12] Gwizdka, J., Moshfeghi, Y., and Wilson, M. L. (2019). Introduction to the special issue on neuro-information science. *Journal of the Association for Information Science and Technology*, 70(9).

[13] Gwizdka, J. and Mostafa, J. (2017). NeuroIIR: Challenges in Bringing Neuroscience to Research in Human-Information Interaction. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, pages 437–438. ACM.

[14] Healy, G., Wang, Z., Gurrin, C., Ward, T. E., and Smeaton, A. F. (2017). An EEG image-search dataset: A first-of-its-kind in IR/IIR. NAILS: neurally augmented image labelling strategies. In *NeuroIIR 2017*.

[15] Hollenstein, N., Troendle, M., Zhang, C., and Langer, N. (2020). Zuco 2.0: A dataset of physiological recordings during natural reading and annotation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 138–146.

[16] Hollenstein, N., Tröndle, M., Plomecka, M., Kiegeland, S., Özyurt, Y., Jäger, L. A., and Langer, N. (2023). The zuco benchmark on cross-subject reading task classification with eeg and eye-tracking data. *Frontiers in Psychology*.

[17] Jacucci, G., Barral, O., Daee, P., Wenzel, M., Serim, B., Ruotsalo, T., Pluchino, P., Freeman, J., Gamberini, L., Kaski, S., and Blankertz, B. (2019). Integrating neurophysiologic relevance feedback in intent modeling for information retrieval. *Journal of the Association for Information Science and Technology*, 70(9):917–930.

[18] Jensen, R. and Shen, Q. (2008). *Computational intelligence and feature selection: rough and fuzzy approaches*. John Wiley & Sons.

[19] Jones, L. M., Wright, K. D., Jack, A. I., Friedman, J. P., Fresco, D. M., Veinot, T., Lu, W., and Moore, S. M. (2019). The relationships between health information behavior and neural processing in african americans with prehypertension. *Journal of the Association for Information Science and Technology*, 70(9):968–980.

[20] Just, M. A., Carpenter, P. A., and Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of experimental psychology: General*, 111(2):228.

[21] Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and trends in Information Retrieval*, 3(1—2):1–224.

[22] Kim, H. H. and Kim, Y. H. (2019). ERP/MMR algorithm for classifying topic-relevant and topic-irrelevant visual shots of documentary videos. *Journal of the Association for Information Science and Technology*, 70(9):931–941.

[23] Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., and Lance, B. J. (2018). Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013.

[24] Li, X., Zhang, P., Song, D., Yu, G., Hou, Y., and Hu, B. (2015). EEG based emotion identification using unsupervised deep feature learning. In *SIGIR 2015 Workshop on Neuro-Physiological Methods in IR Research (NeuroIR 2015)*.

[25] Lopez-Anguita, R., Montejo-Ráez, A., Martínez-Santiago, F. J., and Carlos Díaz-Galiano, M. (2018). Text readability, complexity metrics and the importance of words. *Procesamiento del Lenguaje Natural*, 1(61):101–108.

[26] Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A., and Yger, F. (2018). A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update. *Journal of neural engineering*, 15(3):031005.

[27] Marchionini, G. (2006). Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4):41–46.

[28] Martínez-Santiago, F., Torres-García, A. A., Montejo-Ráez, A., and Gutiérrez-Palma, N. (2021). The impact of reading fluency level on interactive information retrieval. *Universal Access in the Information Society*, pages 1–17.

[29] Mognon, A., Jovicich, J., Bruzzone, L., and Buiatti, M. (2011). Adjust: An automatic eeg artifact detector based on the joint use of spatial and temporal features. *Psychophysiology*, 48(2):229–240.

[30] Morales, E. F., Murrieta-Cid, R., Becerra, I., and Esquivel-Basaldua, M. A. (2021). A survey on deep learning and deep reinforcement learning in robotics with a tutorial on deep reinforcement learning. *Intelligent Service Robotics*, 14(5):773–805.

[31] Moshfeghi, Y. and Pollick, F. E. (2018). Search process as transitions between neural states. In *Proceedings of the 2018 World Wide Web Conference*, pages 1683–1692.

[32] Moshfeghi, Y. and Pollick, F. E. (2019). Neuropsychological model of the realization of information need. *Journal of the Association for Information Science and Technology*, 70(9):954–967.

[33] Moshfeghi, Y., Triantafillou, P., and Pollick, F. (2019). Towards predicting a realisation of an information need based on brain signals. In *The world wide web conference*, pages 1300–1309.

[34] Muñoz-Sandoval, A. F., Woodcock, R. W., McGrew, K. S., Mather, N., and Ardoino, G. (2009). Batería iii woodcock-muñoz. *Ciencias psicológicas*, 3(2):245–246.

[35] Otter, D. W., Medina, J. R., and Kalita, J. K. (2020). A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems*, 32(2):604–624.

[36] Reich, D. R., Prasse, P., Tschirner, C., Haller, P., Goldhammer, F., and Jäger, L. A. (2022). Inferring native and non-native human reading comprehension and subjective text difficulty from scanpaths in reading. In *2022 Symposium on Eye Tracking Research and Applications*, pages 1–8.

[37] Ruthven, I. (2008). Interactive information retrieval. *Annual review of information science and technology*, 42(1):43–91.

[38] Samuels, S. J. and Kamil, M. L. (1984). Models of the reading process. *Handbook of reading research*, 1:185–224.

[39] Štajner, S., Evans, R., Orasan, C., and Mitkov, R. (2012). What can readability measures really tell us about text complexity. In *Proceedings of workshop on natural language processing for improving textual accessibility*, pages 14–22. Citeseer.

[40] Turpin, A. and Scholer, F. (2006). User performance versus precision measures for simple search tasks. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 11–18, New York, NY, USA. ACM.

[41] Wenzel, M. A., Bogojeski, M., and Blankertz, B. (2017). Real-time inference of word relevance from electroencephalogram and eye gaze. *Journal of neural engineering*, 14(5):056007.

[42] Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., and Vaughan, T. M. (2002). Brain–computer interfaces for communication and control. *Clinical Neurophysiology*, 113(6):767–791.

[43] Xu, C. and Zhang, Q. (2019). The dominant factor of social tags for users' decision behavior on e-commerce websites: Color or text. *Journal of the Association for Information Science and Technology*, 70(9):942–953.

[44] Xu, W., Callison-Burch, C., and Napoles, C. (2015). Problems in current text simplification research: New data can help. *Transactions of the Association of Computational Linguistics*, 3(1):283–297.

Appendix A. Approach of the a-posteriori analysis of the dataset

With this data set, an a-posteriori analysis of the EEG signals obtained from the subjects could also be made. Specifically, the recorded data of each subject can be grouped based on the user's number of right answers and the average of right answers obtained for all subjects for the questionnaire in each paragraph. Each questionnaire was composed of three questions related to the paragraph read during the regarding epoch. Based on this, each instance of the original data set is relabeled with one of the following three complexities: easy, normal, and hard. For example, if a subject correctly answered as many questions as the average of all subjects for that paragraph,

Table A.5: Instances distribution for each subject after the relabeling based on the right answers in the questionnaires for each paragraph (instance)

subj	hard	easy	normal
S1	28	3	9
S2	10	13	17
S3	5	11	24
S4	13	6	21
S5	11	8	21
S6	7	10	23
S7	7	9	24
S8	8	9	23
S9	4	12	24
S10	3	19	18
S11	8	12	20
S12	8	16	16
S13	8	13	19
S14	7	13	20
S15	4	14	22
S16	12	10	18
S17	5	16	19
S18	10	8	22

then the instance associated with it is relabeled as normal. Otherwise, if the subject answered less or more than the average, then the instance is relabeled as difficult or easy, respectively.

Table A.5 shows the global distribution of the instances for all subjects. It is interesting to note that, for most subjects except S1, most instances were relabeled as easy or normal.