Universal Access in the Information Society (2023) 22:51–67
https://doi.org/10.1007/s10209-021-00826-y

LONG PAPER

The impact of reading fluency level on interactive information retrieval

Fernando Martínez-Santiago[1] · **Alejandro A. Torres-García**[1] · **Arturo Montejo-Ráez**[1] · **Nicolás Gutiérrez-Palma**[1]

Accepted: 23 June 2021 / Published online: 16 August 2021

Abstract

Given an information need and the corresponding set of documents retrieved, it is known that user assessments for such documents differ from one user to another. One frequent reason that is put forward is the discordance between text complexity and user reading fluency. We explore this relationship from three different dimensions: quantitative features, subjective-assessed difficulty, and reader/text factors. In order to evaluate quantitative features, we wondered whether it is possible to find differences between documents that are evaluated by the user and those that are ignored according to the complexity of the document. Secondly, a task related to the evaluation of the relevance of short texts is proposed. For this end, users evaluated the relevance of these short texts by answering 20 queries. Documents complexity and relevance assessments were done previously by some human experts. Then, the relationship between participants assessments, experts assessments and document complexity is studied. Finally, a third experimentation was performed under the prism of neuro-Information Retrieval: while the participants were monitored with an electroencephalogram (EEG) headset, we tried to find a correlation among EEG signal, text difficulty and the level of comprehension of texts being read during the EEG recording. In light of the results obtained, we found some weak evidence showing that users responded to queries according to text complexity and user's reading fluency. For the second and third group of experiments, we administered a sub-test from the Woodcock Reading Mastery Test to ensure that participants had a roughly average reading fluency. Nevertheless, we think that additional variables should be studied in the future in order to achieve a sound explanation of the interaction between text complexity and user profile.

Keywords Interactive information retrieval · Reading comprehension · Reading fluency · User assessments · Neuro-information retrieval · Electroencephalography

1 Introduction

Text complexity refers to the level of challenge a text provides based on a trio of considerations [26]: quantitative features, subjective difficulty, and reader/text factors. Quantitative features of text complexity are the features that can be counted or quantified: sentence length, number of syllables, word length, word frequency [28, 42], perplexity and other features that can be calculated on the computer. The subjective features of a text are the aspects and nuances of it that cannot be measured by a simple formula. They require careful content analysis [49]. The third leg of the text complexity triad shifts the emphasis from the text itself to reflections on readers and their levels of preparation for tackling both the target text and the objective of the reading task. For each one of these three dimensions, in the present paper we study the relationship between text complexity and tasks related to

✉ Fernando Martínez-Santiago
dofer@ujaen.es

Alejandro A. Torres-García
alejandro.torres@ccc.inaoep.mx

Arturo Montejo-Ráez
amontejo@ujaen.es

Nicolás Gutiérrez-Palma
ngpalma@ujaen.es

[1] Institute Intelligent System for Information Access (SINAI) Advanced Studies Center in Information and Communication Technologies (CEATIC), Universidad de Jaén, Campus Las Lagunillas s/n., 23071 Jaén, Spain

seeking information as a consequence of a given information need. In this context, an additional research question naturally arises: how the user's reading comprehension and fluency leverages the comprehension of texts with different complexity levels. In order to shed light on these questions, we propose a number of experiments inspired by the three dimensions of text complexity introduced above. More concisely, we propose a number of experiments where the user is requested to assess relevance of a retrieved documents. We wondered whether it is possible to cluster user assessments according to:

- Quantitative features of text complexity (research question 1, RQ1).
- Subjective features on the basis of text complexity level provided by humans experts (RQ2).
- Features extracted from the user's mental state when reading (RQ3).

In this paper, we apply perplexity as a quantitative feature of text complexity (research question RQ1, see sect. 2.2). Perplexity has a certain mass of evidence that correlates this measure with on the one hand the precision and recall of information retrieval (IR) systems [4], and on the other hand syntactic complexity [7, 46]. As a consequence we wonder whether, given a probabilistic language model, there are significant differences between the perplexity of the set of documents that are evaluated by the user and those documents that are not evaluated. Precision and recall are two well-known scores to measure the quality of IR systems [5]. Precision measures the ratio of relevant documents among those retrieved and recall the ratio of found documents among all those considered relevant in the collection.

A second way to explore text complexity is by means of subjective assessed difficulty (research question RQ2). We make use of the NEWSELA corpus for this end. The NEWSELA corpus [49] allows us to distinguish easy and complex texts by means of subjective assessed difficulty of the text. In this corpus, difficult expressions have been annotated with a level of difficulty between 1 and 4 by human experts. More details about NEWSELA are provided in Sect. 3.3.

The third dimension of text complexity is related to reflections about readers and their levels of preparation for tackling both the target text and the objective of the reading task. As a consequence, several experiments have been carried out where both the reader's internal state when reading and the user's reading skills are the object of study:

Regarding the **reader's internal state** when he or she is reading, we propose an approach from the Neuro-Information Science field, since there is a growing interest in the use of NeuroIS methods in interactive information retrieval (IIR) research [21, 30]. More concisely, we are intrigued about whether it is possible to find differences in the electrical cerebral activity when texts with different levels of complexity are read. For this purpose, we analyze user electroencephalography (EEG) recordings with the aim of distinguishing when a user is reading a hard or easy text, given a sufficient level of comprehension of the document.

Regarding **participants reading fluency**, it has been measured and integrated into the design of the experimentation related to RQ2 and RQ3. Reading fluency can be defined in several ways, but traditionally it has been related to text reading speed and accuracy (e.g., [1]). In terms of the user's experience, it refers to effortless and efficient reading. Therefore, the reader's ability for fluent reading would critically affect their experience with the IR system. For this purpose, we analyzed participants' performance in a very simple reading task under time pressure. Participants who performed better on this task would pattern differently on other user assessments.

The rest of this paper is structured as follows: firstly, those topics that are needed to accomplish our study are introduced, i.e., reading fluency, text complexity and neuro-Information Retrieval. Then we describe the experimentation framework. In relation to RQ1, we make use of the data collection provided by the PIR-CLEF lab, made up of recordings of research sessions of 10 English-speaker users in coping with one or two information needs. For the case of RQ2, we developed our own corpus based on a set of Spanish documents selected from NEWSELA related to a number of topics. Then, this corpus is used in order to accomplish an information seeking task by 42 Spanish participants. Finally, the EEG of 18 of those 42 participants when reading short texts extracted from NEWSELA documents was recorded and analyzed. We finish with some conclusions and suggestions for future work.

2 Related work

Information Retrieval (IR) is the process of obtaining relevant documents to a given user need, usually under the shape of a query. Thus, an Information Retrieval System will return a list of potentially related documents. The list of documents may contain enough information to help the user decide which documents will answer her information needs. By browsing and opening some of them, the user may decide to refine the query, entering an iterative process until a decision to finish this process is taken. This is why we often talk about Interactive Information Retrieval (IIR) [38], as it is usually inherent to the activity of looking for information. A main concept in IR is that of *relevance*, as it defines whether a document is a valid answer to a user need or not. Borlund [9] studied this topic in depth and enumerates the different aspects that such a topic integrates, which are the reason for the lack of consensus among annotators. We will consider

relevance annotations of the NEWSELA corpus as ground truth, without entering into further analysis on this matter.

Intensive research has been done on the evaluation of IIR systems [10, 24] and certain key aspects, like scrolling behavior, repeat visits or reading time, among others, impact the score assigned to the level of satisfaction when dealing with a text-based search engine. Early works in IR showed that readability could benefit the IIR process [6, 32]. But readability is usually considered only on text characteristics, rather on user abilities for reading, like grammatical or lexical skills on the target language, for instance. Our work focuses on that side of the interactive retrieval model, exploring how reading fluency impacts the performance of the search process. This could contribute to a better understanding on how reading fluency could affect "relevance" in IR.

2.1 Reading fluency

There is evidence that text reading fluency is related to reading comprehension (e. g., [8]). There are several theories to explain this relation. On the one hand, according to the automaticity theory [25], fluent reading is closely related to the automaticity of low-level reading processes, such as word decoding. The more automatized these processes are, the more cognitive resources (limited in nature) are available to perform semantic/high-level comprehension processes, and thus better reading comprehension. The cognitive resources released by fluency mainly involve working memory. Accordingly, [3] found evidence of an indirect link between working memory and reading comprehension through decoding, a low-level process necessary for fluent reading. The relation between working memory and reading has been recently analyzed by [35] in a meta-analysis study. In accordance with [3], they found that the connection between working memory and reading was partialed out when decoding and vocabulary were controlled for. Similarly to decoding, vocabulary processing may be considered another low-level factor necessary for fluent reading. [3] also found a direct link between working memory and reading comprehension, as well as direct connections for other factors such as attention and executive processing. If reading is not fluent, it would be expected that readers have more difficulties to sustain their attention and make decisions about what they are reading, negatively affecting reading comprehension. On the other hand, reading with appropriate expressiveness and intonation (i.e., appropriate prosody) has also been included as a key factor within the concept of reading fluency [31], and it has been related to the construction of the sentence meaning or text microstructure. Whatever the approach followed, speed, accuracy and expressiveness seem to be complementary aspects of fluency that are related to reading comprehension.

2.2 Text complexity

There are different metrics of complexity that have been proposed by various authors since more than fifty years ago. Some of these measures directly provide the recommended age for a reader, such as the *García López* [18] measure, others offer more difficult measures to interpret indexes, such as lexical complexity *Anula* [2], the sentence complexity index or the depth dependency tree *Saggion* [39], among others. Actually, some of them, like the old Flesch score [16], have been used to improve IR systems [6].

In general, few aspects are captured by these features, which essentially rely on basic metrics, mainly lexical, like the number of syllables in a word, the number of rare words, punctuation marks or sentence length. As language models have gained more attention in many language processing tasks like speech recognition or machine translation, measuring text complexity in the model as a useful tool to measure the underlying language. Perplexity is a metric that can be directly related to the complexity of a language model, as is explained in the next section.

2.2.1 Perplexity

The canonical measure of the goodness of a statistical language model is normally reported in terms of perplexity, measurement of how well a probability distribution or probability model predicts a sample. Intuitively, perplexity can be understood as a measure of uncertainty. The perplexity of a language model can be seen as the level of perplexity when predicting the following symbol[11]. In the scope of Information Retrieval[36, 41], we propose statistical language modeling as an alternative to the standard *tf.idf*[37] method of retrieval. In information retrieval, *tf.idf*, short for term frequency–inverse document frequency is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. The tf.idf value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word. Following these studies [4], we found some evidence that the perplexity of the language model has a systematic relationship with the achievable precision recall performance when using traditional Information Retrieval systems. More recently, [46] finds a correlation between perplexity calculated on the basis of part-of-speech (POS) tags and syntactic complexity.

2.3 Electro-encephalography in the field of Neuro-information science

The third dimension of text complexity is related to reflections on readers and their levels of preparation for tackling both the target text and the objective of the reading task. To

the best of our knowledge, our work is the first one aiming to detect differences in the EEG signals depending on the complexity level of a text. We focus on EEG due to its cost, ease of use, wearability, temporal resolution compared to other neuroimaging techniques.

Despite this, nowadays there is a growing interest in the use of methods from neuro-psychology in IIR research such as electro-encephalography (EEG) analysis and eye-tracking. One of the goals is to develop new search models that can account for neurological responses to information stimuli and the influence of cognitive and affective states on users' information behavior. The last is motivated in the work described in [30], which summarizes preliminary evidence for the potential use of analyzing neuroimaging techniques (EEG and fMRI) and eye-tracking during the search process. However, this work also posited that it could be difficult the translation of the knowledge from neuroscience to IR. Hence, they suggested to develop studies focus on neuro-psychology metrics related to search task.

Some preliminary efforts for merging IIR systems and neuro-psychology techniques can be found in the NeuroIIR [21] & NeuroIR international workshops[1]. Nevertheless, only three works explored the used of EEG signals with different purposes such as emotion recognition [27], the creation of a dataset of images neurally labelled [22] using the EEG signals called NAILS, and the prediction of the relevance of a text [14].

Using both eye trackers and EEG signals, two works have focused on determining the level of relevance of a document [19, 20]. Particularly, in [20], a 14-channels EEG device (Emotiv EPOC) was used and a protocol was designed to determine if text document relevance can affect the measurements of EEG signals and eye tracker date differently at early, middle, late stages of reading. They recorded the measurements of both devices from 24 subjects. Also, they applied Proximal Support Vector Machine for the classification stage to the features computed from EEG signals (569 features) and eye tracker data (25 features). This work found that it is possible to distinguish between relevant and irrelevant text documents using the above-mentioned signals. Especially, the biggest differences were found in late stages of reading. Despite this and the outcomes using EEG were slightly above the chance level, the best outcomes were obtained using either EEG and eye tracker together or only eye tracker.

In [19], they analyzed the combination of using an eye tracker and a single channel EEG device (Myndplay Brainband XL). The analysis was carried out on 26 subjects. They mainly assessed if pupil dilation and attention-related measurements taken from the EEG is different between initial visits and re-visits to relevant and irrelevant web pages. At the end of their experimentation, they found significant differences in pupil dilation on visits and revisits to relevant and irrelevant web pages. Nevertheless, these differences were only found in a few conditions using EEG signals when alpha band and attention levels were studied.

In the field of recommendation systems, in [15], 17 participants (analyzing only 15) were asked to read Wikipedia documents about a selection of topics while their EEG was recorded. The subjects explicitly judged as relevant or irrelevant each word of the documents analyzed. The authors designed a protocol for presenting each word of the first six sentences of each document, with which a supervised classification model was able to find the relevant word from the EEG signals. Specifically, they used shrinkage Linear Discriminant Analysis (shrinkage-LDA) due to it is robust to the class imbalance in this experiment. After that, based on the predicted relevant words, the system was able to retrieve documents related to the identified relevant topic. As to the EEG signals classification, the system reached AUC values above the chance level for identifying relevant and irrelevant words for 13 out of 15 subjects.

3 Methods

As stated in the Introduction section, we propose three different research questions to be answered using suitable resources that are available at the moment of carrying out the experimentation. The PIR-CLEF data collection is used to accomplish the first question (RQ1). Research questions RQ2 and RQ3 are interactive experiments where Spanish native speakers have to solve some tasks related with reading fluency. To this end, Woodcock Reading Mastery Test is used to measure the reading fluency and Spanish NEWSELA corpus is the selected resource to evaluate the interactive document retrieval process. At this moment, it is necessary to note that the PIR-CLEF dataset is available for English only so this experimentation on perplexity and user relevance measures (RQ1) is conducted in English only. The following sub-sections provide a brief overview of these resources.

3.1 PIR-CLEF dataset

PIR-CLEF data collection is made up of user profile data and raw search data produced by guided search sessions undertaken by 10 volunteer users. The data provided include the queries submitted, the ranked lists of documents retrieved using a standard search, the items clicked by the user, and document relevance for the user on a 4-grade scale. Three data provided include the queries submitted, the ranked lists

[1] Links to the full list of papers of each workshop are available in https://sites.google.com/view/neuroiir

Table 1 Characterization of the Newsela Corpus (grades 2–6)

Grade level	2	3	4	5	6	Total
Number of texts	59	116	161	146	113	628
Number of texts	59	116	161	146	113	628
Avg. vocabulary size	110	124.46	188.40	210.98	256.51	338.90
Avg. document length	316.56	378.47	577.04	648.98	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 Characterization of the Newsela Corpus (grades 7–10 and 12)

Grade level	7	8	9	10	12	Total
Number of texts	155	115	112	1	245	595
Avg. vocabulary size	278.37	314.70	300.39	425	376.08	177.60
Avg. document length	832.39	930.73	886.84	1249	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	20.93	9.06

of documents retrieved using a standard search, the items clicked by the user, and document relevance for the user on a 4-grade scale. Users were recruited in the researchers working environment. They were between 25 and 40 years old. Their occupations are distributed as follows: four of them are researchers. The people are students and the rest of participants are employees. Six of them are women, with a mean age of 29 years old ($\sigma = 4.7$). On average, men are 33,4 years old ($\sigma = 6.5$). Each session was performed by the users on a topic of their choice, and each search was over a subset of the ClueWeb12 web collection.

Thus, the participants carry out a series of task-based sessions in a controlled way. As a result of these sessions, inter alia, every user assesses at least 19 documents following a stratified sampling method called *2strata strategy* [47]. More details of this strategy and user logs obtained as a result of its application are provided in the overviews of the different editions of the PIR-CLEF campaign [33, 34].

3.2 Woodcock reading mastery test

We used the subtest 2 of the Spanish version of the WM battery [48]. This subtest consists of 105 sentences that could be true or false (e. g., "You can find birds in the countryside" vs. "Dogs are flying animals"). Participants have to read each sentence silently within a time window of three minutes. The difficulty of this task is rather related to speed and accuracy than to the sentence meaning. Sentences are increasingly longer and then progressively more difficult. We scored the number of sentences correctly responded within the those three minutes.

3.3 NEWSELA corpus

Newsela[2] corpus is available for research on text difficulty, among other disciplines [49]. This corpus includes thousands of articles, in both English and Spanish, of professionally adapted news items for different complexities of reading. It consists of a total of 1,130 news articles. Each article has four different versions, according to different grade levels, and produced by editors at Newsela, a specialized company on reading materials for pre-college classroom use. Thus, the corpus is composed of five different subsets: *original, Simp-1, Simp-2, Simp-3* and *Simp-4*. The number of grade levels in the Newsela corpus and some statistics about them are shown in Tables 1 and 2. Note that the "Total" column refers to the total for all grades in the corpus, not only for those in the table.

4 Results

Following the proposed methods, this section presents a description and discussion of the experimental results according to RQ1, RQ2 and RQ3 research questions previously introduced.

[2] https://newsela.com/data

4.1 RQ1: Relationship between language model perplexity and user relevance measures

Following previous works [36, 41], we hypothesize that for a given probabilistic language model there are significant differences between the set of documents that are evaluated by the user and those documents that are not evaluated in terms of perplexity. To this end, an experimentation was carried out using the test collection provided by the PIR-CLEF laboratory inspired by the work of [40].

4.1.1 Text complexity calculus on the basis of perplexity

We used trigram language models with interpolated Kneser–Kney discounting trained using the SRI language modeling toolkit [43]. We generated different models by varying the training corpus. More concisely, we used the Simple-wiki, Sphinx-70k and ClueWeb12 corpora.

Simple-wiki [12] contains 137K sentence Simple English Wikipedia articles. Sphinx-70k uses CMUSphinx US English generic acoustic model[3], is the most general language model that we have considered and the best suited to represent the English language. Finally, a list of documents was retrieved from ClueWeb12 by using every set of queries related to each topic. To this end, an online ClueWeb12 search service[4] was applied in order to retrieve the 100 first ranked documents. As a consequence, we obtained a different language model for each topic proposed in the PIR-CLEF dataset.

Once statistical language models are calculated, the ranked list of documents for each user and query are clustered by following a criteria on the basis of user assessments on these ranks:

- Relevant documents (user relevance judgment is 3 or 4);
- Non-relevant documents (user relevance judgment is 1 or 2);
- Documents without user assessments (there is no user relevance judgment in spite of the fact that they are part of the ranked list of documents retrieved. As a consequence, those documents are unread by users).

Finally, the perplexity of these three different sets of documents per each user and query pair was measured to test if there were statistically significant differences between these measures.

[3] https://sourceforge.net/projects/cmusphinx/files/AcousticandLanguageModels/US

[4] ClueWeb12 search service available at http://clueweb.adaptcentre.ie/WebSearcher/search (18-02-2019)

4.1.2 Results

When the dataset is small, the P-Value from t-Student is likely to be the most usual test but it requires a normal distribution of the dataset. For this reason, we applied the Shapiro-Wilk test that is suited for small datasets and we found that it is not always possible to assert that the datasets considered follow a normal distribution. As a consequence, we applied a non-parametric test, the Mann-Whitney-Wilcoxon U test.

When language models based on Simple-wiki and ClueWeb12 search datasets are applied, we found no significant differences between the perplexity of the three sets of documents considered (relevant, non-relevant or unjudged).

When Sphinkx-70k is used to train the language model, we find some evidence that the perplexity of judged documents (relevant or not relevant) are greater than those that are unjudged (U-value=59, critical U-value at p<0,05=51). This is quite surprising since it could be interpreted as a tendency of the user to evaluate the most complex texts. Once we revise some of the non-judged documents we find that it is quite frequent that these documents do not have any textual content at all, only lists of sections, menus and stylesheets, but none or very little meaningful text.

4.2 RQ2: Evaluation of subjective assessed difficulty of the text

We now try to answer RQ2: according to difficulty of the text, how are documents evaluated by users in contrast with relevance assessment provided by human experts (subjective assessed difficulty)?

The research question differs in some ways from the previous one. Firstly, we focused on a subjective feature of the text so the difficulty level of the text is not calculated but judged by human experts. We consider that a complexity (difficulty) score determined by humans is closer to the real readability level of the text. For this reason, we use the NEWSELA corpus as depicted in Sect. 3.3. Details of the experimentation framework are depicted below. Secondly, we evaluate user performance when comparing relevance assessments provided by experts and those that come from each participant. Finally, participants are asked to judge documents by using a binary relevant/non-relevant scale of values for their assessments. The reason for following this approach comes from the study in [45], where it was found that users are more precise when an easier scale of assessment is used.

In summary, the goal is to find out whether there are significant differences in the degree of concordance between experts and participants according to text complexity.

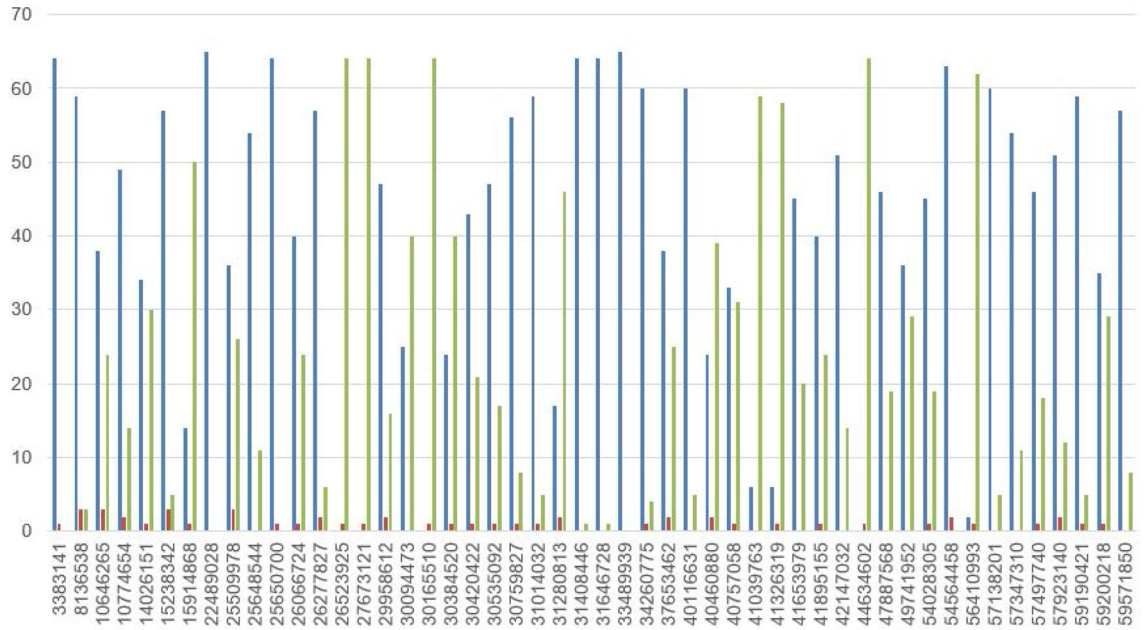

Fig. 1 Woodcock-Muñoz reading fluency test results. For each participant, blue, red and green columns correspond with the number of right, wrong and unanswered questions

4.2.1 Participants reading fluency level

From a methodological point of view, it is relevant that all the participants were administered the Woodcock-Muñoz reading fluency test (see Sect. 3.2 for more details) as a way to ensure that all of them achieved a roughly average reading fluency skill. This test was completed by 42 participants (31 men, 11 women, mean age=22.4, s=3.7), all of them are Spanish and consequently, Spanish is their first language. They are University students recruited from three different grades (psychology, computer engineering and electrical engineering). Five of these participants were not considered because they showed abnormally low values on the Woodcock-Muñoz reading fluency test (see Fig. 1). On average, 44.76 of 65 questions are answered correctly(68.86%, s= 17.15), 1.02 questions wrongly(2%, s=0.93), and 19.22 questions are not answered (29.1%, s=17.08). Note that the end is to study the impact of the complexity of the text. Therefore, it is appropriate that users are in a similar reading fluency level. That is, it is studied the difference in reading comprehension based on the complexity of the text in a population that has a comparable reading fluency.

4.2.2 Gathering of user assessments

Only once participants accomplished the reading fluency test, are they in condition to start with the second part of the experiment where data gathering takes place (see Fig. 2). The data gathering process takes place over three main phases: query[5] development, final query description, and relevance assessment. The IR system provides a total of 20 pre-stored queries so every user has to (i) execute the given query, (ii) open and, eventually, read some documents from the list of documents that is obtained as a result of the query execution, (iii) submit a summary of her/his findings with regard to the accomplished search task and (iv) judge the relevance of a set of sampled results for each topic that s/he has developed during the search session. Figure 3 shows an example for the query with title *"Intercultural communities"*. Figures 4 and 5 are an example, a fragment of the Web interface that is shown to the user: as a result of the execution of a given query, the IR system lists the title of 20 documents. Eventually, the user selects one document from the list. Then, the whole text of the document is shown. Finally the user assess the relevance of the document for the given query.

4.2.3 Document collection

The document set is made up of 368 documents written in Spanish distributed among 20 queries. Every query has a title and a description, a field similar to the one shown in Fig. 3.

[5] Note that the term "query" in the context of an Information Retrieval system must be assimilated to a textual title and/or description, expression of a given information need, such is exemplified in Fig. 3.

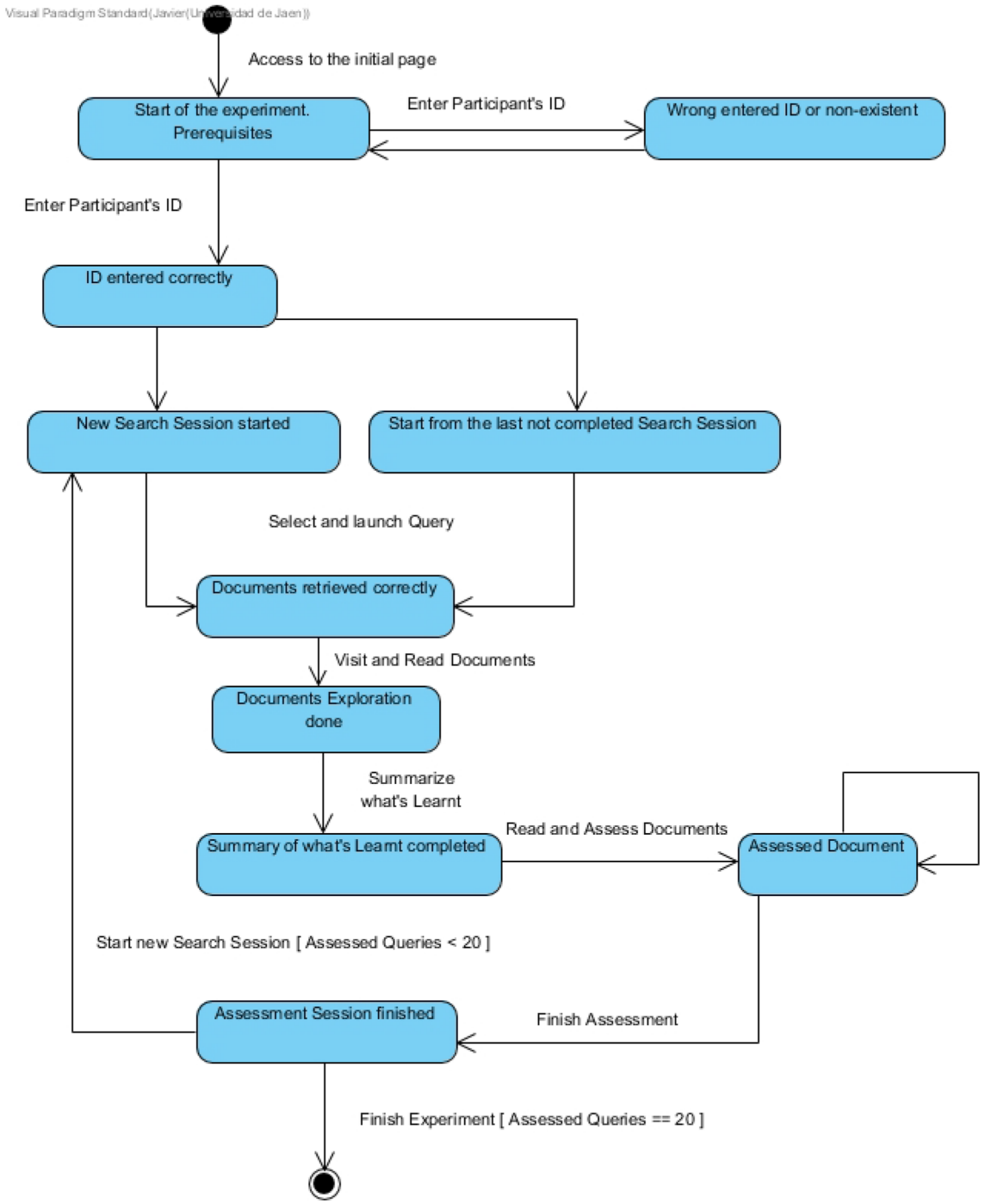

Fig. 2 Data gathering scheme

As a result, every query is related to 20 documents, some of them shared among different queries. The sources of the document set are NEWSELA and the Web. From NEWSELA, we have chosen 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 is the lowest or the highest. Consequently, documents from NEWSELA whose text complexity is in the middle of the scale (2 and 3 categories) are not considered. Because the number of documents obtained by this method is low (4.8 on average per query, s = 3.17), we completed the collection and achieved 20 documents per query by searching for documents related in some way to every query. In order to accomplish this task, our first option was the ClueWeb12 dataset, but it is difficult to find useful Spanish

```
<top>
<num> 001 </num>
<title> Comunidades interculturales</title>
<desc> Encuentra ejemplos de sociedades que estén conformadas
por diversas etnias y culturas, y que
coexistan pacíficamente, en armonía
 e integradas como parte de una misma comunidad</desc>
<auth> Fernando Martínez</auth>
<lang>ES</lang>
</top>
```

Fig. 3 Example of query: Intercultural communities. Find examples of societies that are made up of different ethnicities and cultures, and that coexist peacefully, in harmony and integrated as part of the same community

documents according to the definition of the task. Finally, we opted to look for documents related to every query on the Web.

Once the document set is defined, the following step is the creation of relevance assessments. Thus, every query and document pair is judged by three human experts, achieving an inter-rate agreement (kappa value) of $K = 0.83$. The distribution of assessments is depicted in Table 3.

4.2.4 Results

With the aim of obtaining significant differences in participants' performance, the analysis of the user assessments was accomplished by partitioning both documents and participants according to document complexity and reading fluency, respectively. User performance is measured in terms of precision and recall values. This is interpreted as a measure of participants-experts agreement in the task of judging documents in relation to a given query.

As expected, the best results were obtained by considering NEWSELA-easy and the most proficient readers (p80 group) and the worst results when NEWSELA-hard and p20 participants were considered, but the differences are very modest and hardly statistically significant (Table 4). For this reason, in order to look into the relationship between participant reading fluency and sensitivity (true positive) and specificity (true negatives) measures of the assessments, the Pearson correlation coefficient was calculated, obtaining $R = 0.245$. Although technically a positive correlation, the relationship between both variables is weak. The value of the coefficient of determination R^2 is 0.06. In a similar way, by considering false positives and false negatives participants assessments, $R = -0.3024$ which is a moderate inverse correlation.

4.3 RQ3: Relation between text complexity and brain activity

According to text complexity(RQ3), can we find differences in the EEG analysis of the brain activity of the users?

We designed a recording protocol for the EEG signals, which we explain below. A fixation cross appeared at the start of the timeline. At second 2, the subject heard a beep, while the fixation cross was activated, to keep their attention before the main stimulus of our experiment. Later, a paragraph to be internally read appeared at the 3rd second. A set of paragraphs was taken from the NEWSELA corpus. Each paragraph had an associated reading complexity, easy or difficult. These were defined in the original corpus along with other intermediate complexities. For sending a trigger to the EEG signals to delimit the end of the paragraph reading, the subject had to press a key after reading each paragraph. In addition, in order to avoid the subjects being distracted from the experiment the system had a maximum duration defined for the number of words times 0.5 s, which was never activated. After the paragraphs were shown, a set of 3 true/false questions were displayed on the screen. The subjects were instructed to respond only to those questions of which they were more sure of a correct answer (empty answers were allowed). Also, the subjects did not have any time limitation. Finally, a black screen was displayed to indicate to the subject a pause whose duration was 1 s. The timeline of the protocol can be seen in Fig. 6.

Figures 7 and 8 show examples of the texts' reading complexity. In this case, Fig. 7 is an example of a difficult text. Figure 8 is an example of an easy text for reading. Translation into English is given in parenthesis.

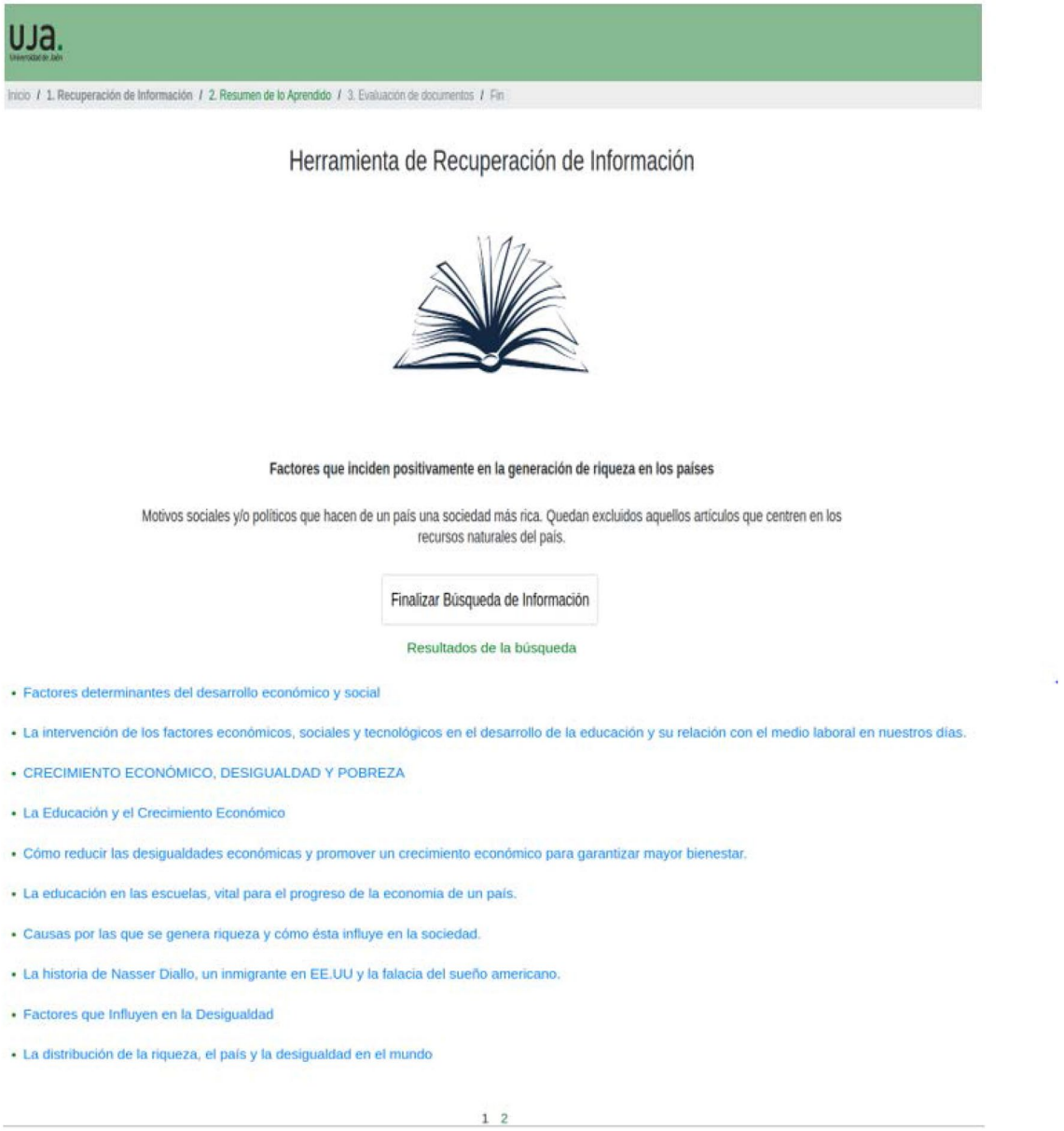

Fig. 4 List of relevant documents for a given query (factors that have a positive impact on the generation of wealth in the countries of the world) as they are shown to the participants

4.3.1 EEG Material

The EEG signals were recorded with an EPOC kit [13] from the company Emotiv, consisting of 14 channels (AF3, AF4, F3, F4, F7, F8, FC5, FC6, P7, P8, T7, T8, O1, O2. Ordering data: P3/CMS, P4/DRL) and it includes a gyroscope in order to record movements along the x and y axes. The EPOC is wireless and has a sampling frequency of 128 Hz (Fig. 9).

4.3.2 Dataset filtering

Since our focus is to determine whether it is possible to distinguish when a text is difficult or easy to read by analyzing the recording of the participant EEG signal, a filter was applied so that participants whose results in the reading fluency test were above percentile 80 or under percentile 20 were not considered. Thus, a total of 18 participants took part in this study. The reason is that we are interested in readers whose reading proficiency is roughly average.

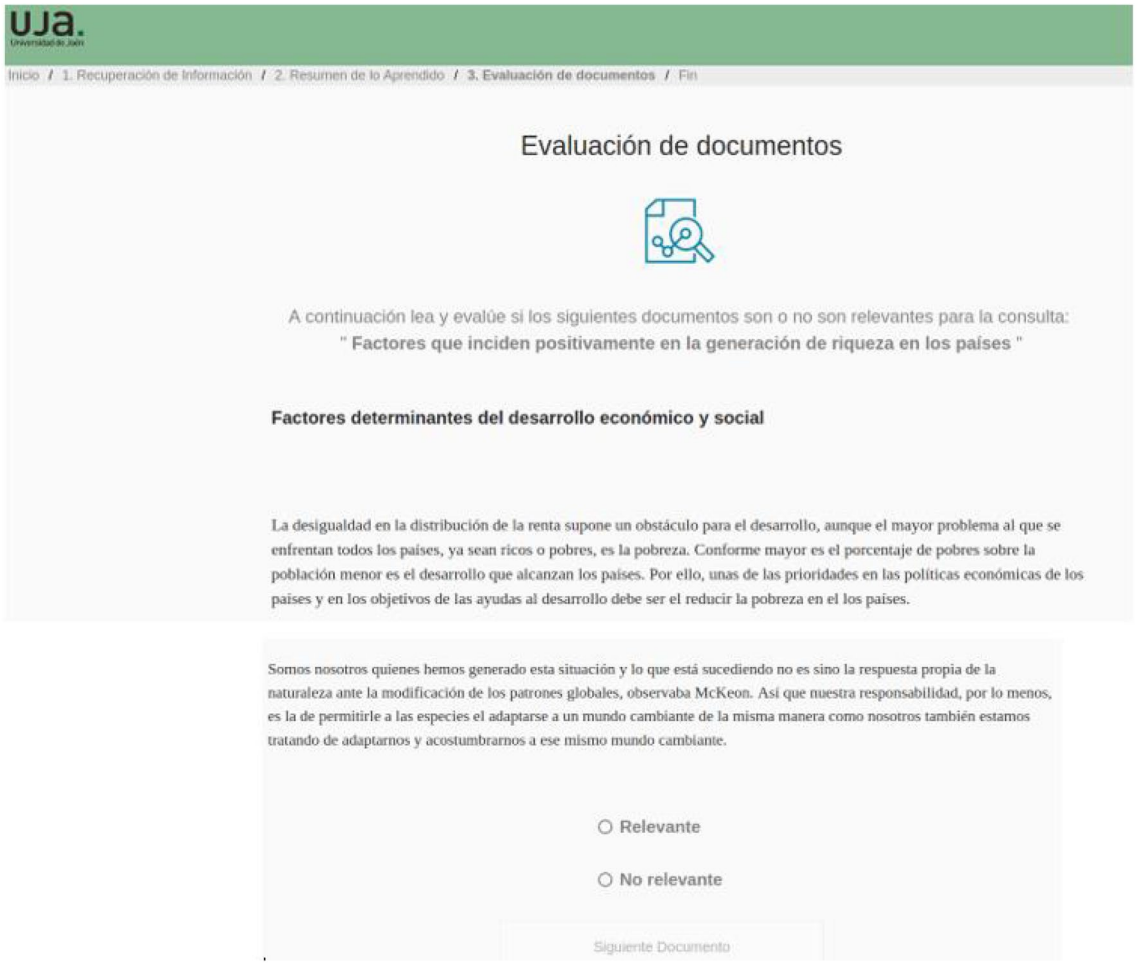

Fig. 5 Once a participant selects a given document of the proposed list of relevant document, it is shown together radio button below the text document so the user assess the relevance, or not, of the given document

Table 3 Relevant(+)/Non-relevant(-) document distribution. NEWSELA-easy are those documents whose complexity is the lowest. Similarly, NEWSELA-hard are the hardest ones

	Web		NEWSELA-easy		NEWSELA-hard	
Relevance	+	-	+	-	+	-
Mean per query	9.15	6.05	2.1 (s=0.62)	2.7 (s=1.47)	2.6 (s=0.68)	2.3 (s=1.21)
Total	183	121	18	30	24	24

Table 4 Precision and recall measures obtained by participants considering, on the one hand, the Web, NEWSELA-easy, NEWSELA-hard subsets and on the other hand participants under percentile 20 (p20) and over percentile 80 (p80) of reading fluency performance

	Web		NEWSELA-easy		NEWSELA-hard	
Participants	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
All	0.81	0.85	0.82	0.90	0.78	0.83
p20	0.77	0.82	0.76	0.82	0.71	0.81
p80	0.85	0.87	0.84	0.92	0.81	0.91

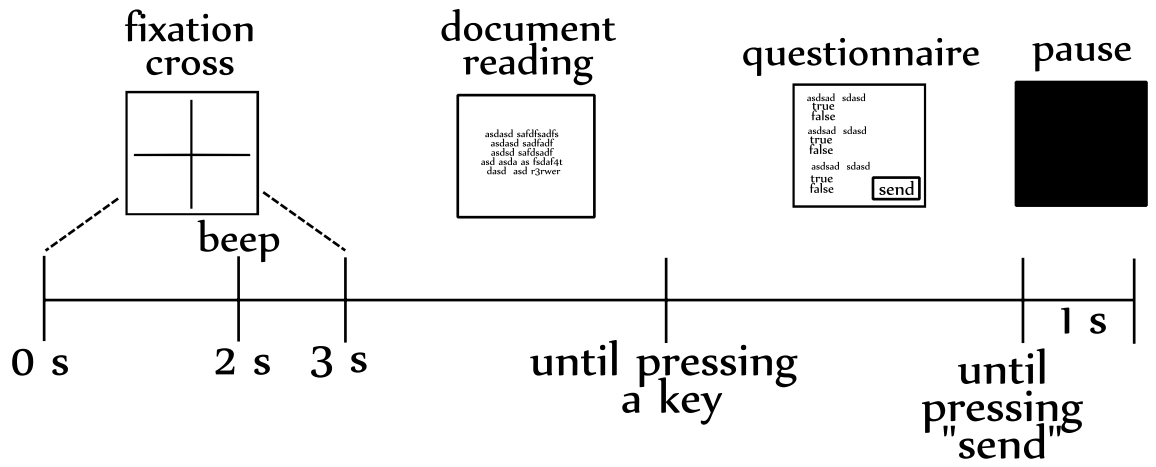

Fig. 6 Timeline of the recording protocol

Fig. 7 An example of a document with high textual complexity, and the corresponding questioner

El cartel de la muestra es obra de un artista menos conocido, Huygh Pietersz Voskuyl. Su notable autorretrato de 1638 muestra una pose clásica de "selfie": mira sobre su hombro derecho hacia el exterior del marco. No se necesita demasiada imaginación para verlo mirando la cámara de un móvil en lugar de un espejo, el recurso utilizado por estos artistas para realizar sus autorretratos.

(*The poster for the exhibition is the work of a lesser-known artist, Huygh Pietersz Voskuyl. His remarkable self-portrait from 1638 shows a classic "selfie" pose: he looks over his right shoulder to the outside of the frame. It does not take much imagination to see him looking into a cell phone camera instead of a mirror, the resource used by these artists to make their self-portraits.*)

Q1: Huygh Pietersz Voskuyl hacía uso de espejos para pintar autorretratos.

(*Huygh Pietersz Voskuyl made use of mirrors to paint self-portraits.*)

Q2: El cartel de la muestra pictórica es un selfie hecho con un móvil.

(*The poster of the painting is a selfie made with a mobile phone.*)

Q3: El cartel de la muestra es obra de Huygh Pietersz Voskuyl.

(*The poster shows a work by Huygh Pietersz Voskuyl.*)

4.3.3 Experiments and results

Using this protocol, we recorded the EEG signals from 18 subjects as depicted above. Each subject read 40 paragraphs balanced between the two complexities (easy and difficult). Here, it is important to highlight that we analyzed only the EEG segments in which the subjects read the paragraphs. Since the paragraphs' length was variable, the duration of these segments, too. For analyzing the same epoch size, we focused on the 3 intermediate seconds of all epochs of interest (during paragraph reading).

For the analysis and processing of the data, we grouped the data following two strategies. The first one was to use a *priori* labels for each paragraph. For this case, each paragraph is labeled as easy or difficult depending on the level of complexity according to the difficulty level of the NEWSELA document that is source of the paragraph. The question here is whether it is possible to distinguish when a user reads an easy or difficult text. The second strategy was to consider a group of the recorded EEG signals defined according to the number of correct answers of each subject from the average score obtained by the subjects in the questionnaire of each

Mucha gente está sembrando en las áreas donde viven los leones. También usan esas tierras para criar otros animales. Además, los cazadores están matando muchos animales salvajes. Los leones casi no tienen qué comer. Los cazadores también matan a los leones. Este año, acabaron con la vida de Cecil, un león muy famoso. Muchas personas protestaron por lo ocurrido.

(*Many people are planting in the areas where the lions live. They also use those lands to raise other animals. In addition, hunters are killing many wild animals. The lions have almost nothing to eat. The hunters are killing the lions too. This year, they killed Cecil, a very famous lion. Many people protested about what happened.*)

Q1: El cambio climático es uno de los motivos por los que hay menos leones.

(*Climate change is one of the reasons why there are fewer lions.*)

Q2: Cecil es el nombre de un león víctima de unos cazadores.

(*Cecil is the name of the lion* victim of the hunters.)

Q3: Los leones cada vez tiene menos espacio y comida.

(*Lions have less and less space and food.*)

Fig. 8 An example of a document with low textual complexity, and the corresponding questioner

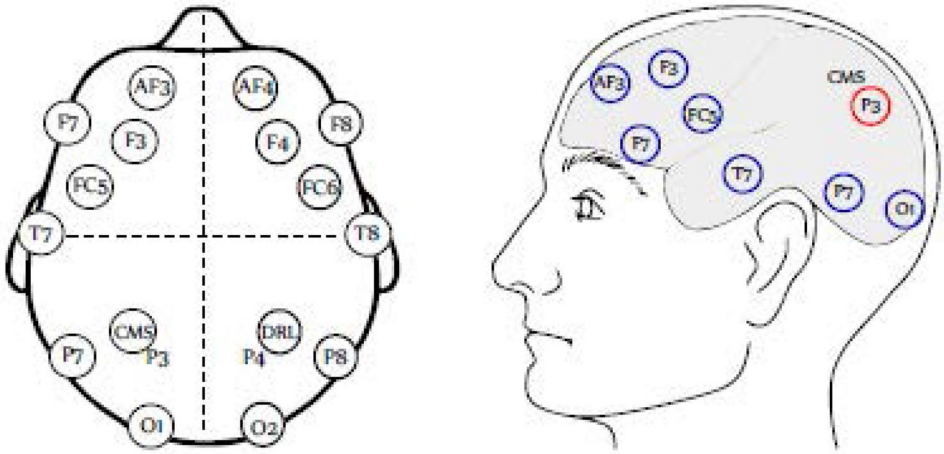

Fig. 9 Location of the electrodes in the EMOTIV headset

paragraph. In other words, we use an a *posteriori* analysis. Every pair $< participant, paragraph >$ is labelled as easy, normal or difficult, according to the performance (correct answers) obtained by the participant on each questionnaire.

Our main objective was to assess whether a classifier could classify between these levels of difficulty for each subject and for the two strategies separately.

Note that the *a priori* group of the EEG records only depends on the textual complexity tag in NEWSELA, while the *a posteriori* partition of those same records varies from one user to another, depending on their performance when answering the test that follows the reading of each paragraph.

For both strategies of analysis, an automatic artifact removal method was applied in order to remove undesired signals. This was the ADJUST algorithm [29], which is an ICA-based algorithm. This algorithm was chosen as the fact that it has good performance rejecting blinks, eye movements and generic discontinuities. At the end of the ADJUST processing the artifacted components were selected. Later we removed the artifacted components in order to create a set of clean EEG signals by only using the non-artifacted components. Then we applied a 5th order pass-band Butterworth filter (1–50 Hz). After that, common average reference was applied to remove the average voltage from all the EEG channels at a same time instant.

Table 5 AUC values obtained with and without feature selection (FS) and the *a priori* strategy (with 2 difficulty levels: easy and difficult)

	Without FS		With FS	
subject	AUC	(± std)	AUC	(± std)
S1	0.79	0.22	0.70	0.33
S2	0.23	0.17	0.34	0.29
S3	0.30	0.27	0.34	0.28
S4	0.29	0.30	0.64	0.18
S5	0.18	0.19	0.45	0.33
S6	0.68	0.26	0.59	0.29
S7	0.26	0.20	0.45	0.27
S8	0.43	0.37	0.44	0.17
S9	0.46	0.33	0.63	0.34
S10	0.48	0.22	0.66	0.37
S11	0.43	0.30	0.35	0.29
S12	0.69	0.30	0.58	0.25
S13	0.64	0.27	0.60	0.2
S14	0.53	0.34	0.74	0.19
S15	0.54	0.20	0.54	0.33
S16	0.40	0.33	0.48	0.3
S17	0.23	0.26	0.59	0.4
S18	0.78	0.17	0.64	0.21

Table 6 AUC obtained with and without feature selection (FS), and the *a posteriori* strategy based on the questionnaires of each paragraph

	Without FS		With FS	
subject	AUC	(± std)	AUC	(± std)
S1	0.54	0.28	0.55	0.30
S2	0.40	0.26	0.58	0.24
S3	0.38	0.25	0.42	0.32
S4	0.35	0.24	0.39	0.25
S5	0.76	0.18	0.62	0.21
S6	0.38	0.23	0.44	0.25
S7	0.38	0.24	0.55	0.24
S8	0.63	0.27	0.35	0.25
S9	0.43	0.27	0.48	0.33
S10	0.58	0.25	0.60	0.32
S11	0.51	0.19	0.33	0.23
S12	0.66	0.18	0.77	0.21
S13	0.41	0.24	0.55	0.23
S14	0.35	0.21	0.59	0.18
S15	0.58	0.28	0.39	0.21
S16	0.60	0.23	0.50	0.12
S17	0.39	0.24	0.45	0.28
S18	0.42	0.24	0.37	0.33

Later temporal and frequency features were extracted for the epochs during the paragraph. Since the variability in the epochs' length, we only focused on the 3 intermediate seconds of these epochs as above-mentioned. As to the extracted temporal features they were the mean, median, standard deviation, variance, maximum, minimum, sum, difference and sum between maximum and minimum, kurtosis, skewness, entropy and zero-crossing rate. These features were chosen as they can capture global shapes and changes in the temporal domain of the EEG signals. Also, we based on the work described in [50], which was applied to a similar problem to ours in imagined speech recognition.

The frequency features were computed after the application of discrete wavelet transform (sDWT) with 4 decomposition levels (D1-D4 and A4) and using a biorthogonal 2.2 wavelet as the mother wavelet function. They also allows the analysis of changes in each one of the brain rhythms. These decomposition levels allow an easy mapping between the levels and the brain rhythms so that D1 captures frequencies between 32-64 Hz (gamma and beta), D2 captures frequencies between 16-32 Hz (beta), D3 captures frequencies between 8-16 Hz (alpha), D4 captures frequencies between 4-8 Hz (theta) and A4 the frequencies up to 4 HZ (delta). In addition, these parameters were selected due to their performance in a similar task called imagined speech [44]. For the coefficients at each decomposition level, the following set of features was computed: instantaneous wavelet energy (IWE), relative wavelet energy (RWE), teager wavelet energy (TWE), mean, median, standard deviation, variance, ratio of the mean in adjacent sub-bands, maximum, minimum, sum, difference and sum between maximum and minimum, kurtosis and skewness.

After the feature vectors for each subjects were computed, we applied correlation-based feature subset selection looking to determine whether feature selection could improve or, at least, keep our results using all the features. Then we classified them using random forests with 50 trees in both strategies (with and without feature selection). We assessed the use of all features (belonging to all channels) because of the following two reasons. The first one is that optimal locations for recording are unknown for this task (as opposed to motor imagery). The second reason is that the performance got using all features is used as a baseline for measuring if an improvement could be gotten applying feature selection.

The results obtained for the a *priori* levels of complexity (easy and difficult) are obtained using 10-folds cross-validation applied to each subject' data separately (see Table 5). Despite the performances for S1 and S18 are at a range of acceptable discrimination (according to [23]), the area under the ROC (receiver operating characteristic) Curve (AUC) for most the subjects suggest no discrimination between the classes, which could suggest that there is no difference in brain activity when a person reads documents with different complexity levels. Furthermore, a sign test showed that there is no difference between applying feature selection or

not applying it ($Z = -0.97$, $p = 0.332$). After analyzing the box-plots of the outcomes, the sign test was chosen due to the data distribution is not normal and asymmetrical.

The outcomes obtained for the *a posteriori* strategy (3 classes for each paragraph: easy, normal and difficult) are shown in Table 6, which is based on the number of correct answers in the questionnaires for each paragraph. Despite Feature Selection got better AUC values than using all the features, a sign test indicated that there is no difference between both schemes ($Z = -1.179$, $p = 0.238$). This is in agreement with the a *priori* strategy's results. Both results suggest that the identification of the complexity level of a paragraph is a difficult task, and maybe no difference could be found in EEG signals.

5 Discussion

Information Retrieval has traditionally been studied in terms of precision and coverage obtained by the various search algorithms, being an algorithm-centric evaluation. In contrast, Interactive Information Retrieval adopts a user-centered perspective, focused on actors in the information seeking process, as a particular case of Human Information Interaction[17]. The present work is framed in this approach, more particularly in studying a specific user's trait, his reading fluency. With respect to other works carried out in this field, and briefly reviewed in Sect. 2, the present work is an effort to make an approximation as sound as possible in order to study user seeking performance text complexity, attending to the three aspects identified in [26] and introduced in Sect. 2.2. On the other hand, users information seeking will necessarily be biased by their reading fluency, but such trait is unseen in related works. Even more, considering that the number of participants is frequently very small (it is not easy to recruit people interested in this type of experiments, whose administration requires considerable time and effort from the participant), we believe that it is especially relevant to ensure that the degree of reading fluency is on-average in order to avoid to some extent results biased by the profile of, perhaps, one or more participants. At this point, we propose to administer the Woodcock reading test so that such cases can be identified and ruled out. In short, the present work is an attempt to reduce the gap between the different perspectives to study text complexity and Interactive Information Retrieval when an on-average reader faces an information seeking task.

6 Conclusions and future work

In this work, we approach the impact of text complexity on the task of Interactive Information Retrieval from different dimensions. Firstly, we focus on quantitative features of text complexity (perplexity) in order to distinguish those documents that are evaluated by the users from those that are not evaluated. A second framework is defined by considering subjective document features trying to evaluate the users' performance when they come to the task of judging documents related to a given query. In addition, in this case it is guaranteed that participants have a good enough reading fluency level. Weak evidence is found that correlates reading fluency and user performance. In the same way, in general, easier documents are slightly more accurately judged by the participants. EEG recordings and posterior analysis evidences how subtle is the distinction when reading texts with different reading complexity and, in general, it is not possible to find significant differences extracted from EEG signals.

As a conclusion, we find a certain mass of evidence that correlates text difficulty and user performance when interacting with an Information Retrieval system as part of an information seeking task but we think that is necessary to explore this relationship more deeply. An obvious first step in this direction would be to gather more data, that is, more participants. Following this line, we think that the integration of fluency reading levels as part of the user profile is a powerful tool that must be applied more in-depth, but it requires the recruitment of more participants and more varied profiles. Finally, we think that reading comprehension is a factor to consider when IR user behavior is studied but additional variables regarding the user profile must be considered when assessing user performance with the aim of enabling us to explain and carry out a more fine-grained analysis of the results. In addition, in case user reading fluency is known or it can be approximated as part of the user profile, in line with the results correlating text difficulty and user performance when searching(RQ2), we believe that it is likely that user experience may be improved when document rankings provided by information systems are the result of integrating document relevance, document complexity and user reading fluency.

Acknowledgements This work has been partially supported by the LIVING-LANG project (RTI2018-094653-B-C21) from the Spanish Government and Fondo Europeo de Desarrollo Regional (FEDER). This work was partially supported by CEATIC from Universidad de Jaen through the "Premios de Invitacion de Movilidad" for young doctors.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Informed consent and participants Research participants are physically and mentally healthy adults who are neither providing personal information nor are required to make any task that is dangerous under any point of view. Thus, according to the ethical standards applicable to us, all of them signed an informed consent where detailed information was provided: study purpose, steps and procedures. They were all informed that the information that they provide will be kept confidential, and that participation is voluntary.

References

1. Adams, M.J.: Beginning to read: A critique by literacy professionals and a response by Marilyn Jager Adams. Read Teacher **44**(6), 370–395 (1990)
2. Anula, A.: Lecturas adaptadas a la enseñanza del español como l2: variables lingüísticas para la determinación del nivel de legibilidad. La evaluación en el aprendizaje y la enseñanza del español como LE L **2**, 162–170 (2008)
3. Arrington, C.N., Kulesz, P.A., Francis, D.J., Fletcher, J.M., Barnes, M.A.: The contribution of attentional control and working memory to reading comprehension and decoding. Sci. Stud. Read. **18**(5), 325–346 (2014)
4. Azzopardi, L., Girolami, M., van Risjbergen, K.: Investigating the relationship between language model perplexity and ir precision-recall measures. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 369–370. ACM (2003)
5. Baeza-Yates, R., Ribeiro-Neto, B., et al.: Modern information retrieval, vol. 463. ACM press, New York (1999)
6. Belkin, N. J., Chaleva, I., Cole, M. J., Li, Y., Liu, L., Liu, Y.-H., Muresan, G., Smith, C. L., Sun, Y., Yuan, X., et al.: Rutgers' hard track experiences at trec 2004. In *Proceedings of the Text REtrieval Conference 2004 (TREC)*. NIST (2004)
7. Berdicevskis, A., Çöltekin, Ç., Ehret, K., von Prince, K., Ross, D., Thompson, B., Yan, C., Demberg, V., Lupyan, G., Rama, T., et al.: Using universal dependencies in cross-linguistic complexity research. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 8–17, (2018)
8. Berninger, V.W., Abbott, R.D., Trivedi, P., Olson, E., Gould, L., Hiramatsu, S., Holsinger, M., McShane, M., Murphy, H., Norton, J., et al.: Applying the multiple dimensions of reading fluency to assessment and instruction. J. Psychoedu. Assessment **28**(1), 3–18 (2010)
9. Borlund, P.: The concept of relevance in ir. J. Am. Soc. Inform. Sci. Technol. **54**(10), 913–925 (2003a)
10. Borlund, P.: The iir evaluation model: a framework for evaluation of interactive information retrieval systems. Inf. Res. **8**(3), 8 (2003b)
11. Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Lai, J.C., Mercer, R.L.: An estimate of an upper bound for the entropy of English. Comput. Linguist. **18**(1), 31–40 (1992)
12. Coster, W., Kauchak, D.: Simple english wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 665–669. Association for Computational Linguistics (2011)
13. Duvinage, M., Castermans, T., Petieau, M., Hoellinger, T., Cheron, G., Dutoit, T.: Performance of the emotiv epoc headset for p300-based applications. Biomed. Eng. Online **12**(1), 56 (2013)
14. Eugster, M., Ruotsalo, T., Spape, M., Kosunen, I. J., de Bellegarde, O. B. M., Ravaja, J. N., Jacucci, G., Kaski, S. J. I.: Predicting relevance of text from neuro-physiology. In *SIGIR 2015 Workshop on Neuro-Physiological Methods in IR Research (NeuroIR 2015)* (2015)
15. Eugster, M.J., Ruotsalo, T., Spapé, M.M., Barral, O., Ravaja, N., Jacucci, G., Kaski, S.: Natural brain-information interfaces: Recommending information by relevance inferred from human brain signals. Sci. Rep. **6**, 38580 (2016)
16. Farr, J.N., Jenkins, J.J., Paterson, D.G.: Simplification of flesch reading ease formula. J. Appl. Psychol. **35**(5), 333 (1951)
17. Fidel, R.: Human information interaction: An ecological approach to information behavior. MIT Press, Cambridge (2012)
18. García López, J.: Legibilidad de los folletos informativos. Pharmaceutical Care España **3**(1), 49–56 (2001)
19. Gwizdka, J.: Inferring web page relevance using pupillometry and single channel eeg. In *Information Systems and Neuroscience*, (pp. 175–183). Springer (2018)
20. Gwizdka, J., Hosseini, R., Cole, M., Wang, S.: Temporal dynamics of eye-tracking and eeg during reading and relevance decisions. J. Assoc. Inf. Sci. Technol. **68**(10), 2299–2312 (2017)
21. Gwizdka, J., Mostafa, J.: 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 (2017)
22. Healy, G., Wang, Z., Gurrin, C., Ward, T. E., Smeaton, A. F.: An EEG image-search dataset: A first-of-its-kind in IR/IIR. NAILS: neurally augmented image labelling strategies. In *NeuroIIR 2017* (2017)
23. Hosmer, D.W., Jr., Lemeshow, S., Sturdivant, R.X.: Applied logistic regression, vol. 398. Wiley, New Jersy (2013)
24. Kelly, D.: Methods for evaluating interactive information retrieval systems with users. Found. Trends Inf. Retrieval **3**(1–2), 1–224 (2009)
25. LaBerge, D., Samuels, S.J.: Toward a theory of automatic information processing in reading. Cogn. Psychol. **6**(2), 293–323 (1974)
26. Lapp, D., Moss, B., Grant, M.: *A close look at close reading: Teaching students to analyze complex texts, Grades K–5*. ASCD (2015)
27. Li, X., Zhang, P., Song, D., Yu, G., Hou, Y., Hu, B.: EEG based emotion identification using unsupervised deep feature learning. In *SIGIR 2015 Workshop on Neuro-Physiological Methods in IR Research (NeuroIR 2015)* (2015)
28. Lopez-Anguita, R., Montejo-Ráez, A., Martínez-Santiago, F.J., Carlos Díaz-Galiano, M.: Text readability, complexity metrics and the importance of words. Procesamiento del Lenguaje Natural **61**, 101–108 (2018)

29. Mognon, A., Jovicich, J., Bruzzone, L., Buiatti, M.: Adjust: An automatic eeg artifact detector based on the joint use of spatial and temporal features. Psychophysiology **48**(2), 229–240 (2011)
30. Mostafa, J., Gwizdka, J.: Deepening the role of the user: Neurophysiological evidence as a basis for studying and improving search. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, pages 63–70. ACM (2016)
31. National Reading Panel (US). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. National Institute of Child Health and Human Development (2000)
32. Newbold, N., McLaughlin, H., Gillam, L.: Rank by readability: Document weighting for information retrieval. In Cunningham, H., Hanbury, A., and Rüger, S., editors, *Advances in Multidisciplinary Retrieval*, pages 20–30, Berlin, Heidelberg. Springer Berlin Heidelberg (2010)
33. Pasi, G., Jones, G. J., Curtis, K., Marrara, S., Sanvitto, C., Ganguly, D., Sen, P.: Evaluation of personalised information retrieval at clef 2018 (pir-clef). In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 335–342. Springer (2018)
34. Pasi, G., Jones, G. J., Marrara, S., Sanvitto, C., Ganguly, D., Sen, P.: Evaluation of personalised information retrieval at clef 2017 (pir-clef): towards a reproducible evaluation framework for pir. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum* (2017)
35. Peng, P., Barnes, M., Wang, C., Wang, W., Li, S., Swanson, H.L., Dardick, W., Tao, S.: A meta-analysis on the relation between reading and working memory. Psychol. Bull. **144**(1), 48 (2018)
36. Ponte, J. M., Croft, W. B.: A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 275–281, New York, NY, USA. ACM (1998)
37. Ramos, J. et al.: Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 133–142. New Jersey, USA (2003)
38. Ruthven, I.: Interactive information retrieval. Annu. Rev. Inf. Sci. Technol. **42**(1), 43–91 (2008)
39. Saggion, H., Štajner, S., Bott, S., Mille, S., Rello, L., Drndarevic, B.: Making it simplext: Implementation and evaluation of a text simplification system for spanish. ACM Trans. Access. Comput. (TACCESS) **6**(4), 14 (2015)
40. Sanvitto, C., Ganguly, D., Jones, G. J., Pasi, G.: A laboratory-based method for the evaluation of personalised search. In *EVIA@ NTCIR*(2016)
41. Song, F., Croft, W. B.: A general language model for information retrieval. In *Proceedings of the eighth international conference on Information and knowledge management*, pages 316–321. ACM (1999)
42. Štajner, S., Evans, R., Orasan, C., Mitkov, R.: 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 (2012)
43. Stolcke, A.: Srilm-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing* (2002)
44. Torres-García, A.A., Reyes-García, C.A., Villaseñor-Pineda, L., García-Aguilar, G.: Implementing a fuzzy inference system in a multi-objective eeg channel selection model for imagined speech classification. Expert Syst. Appl. **59**, 1–12 (2016)
45. Vakkari, P., Sormunen, E.: The influence of relevance levels on the effectiveness of interactive information retrieval. J. Am. Soc. Inform. Sci. Technol. **55**(11), 963–969 (2004)
46. von Prince, K., Demberg, V.: Pos tag perplexity as a measure of syntactic complexity. *Shared Task on Measuring Language Complexity*, page 20 (2018)
47. Voorhees, E. M.: The effect of sampling strategy on inferred measures. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 1119–1122. ACM (2014)
48. Woodcock, R. W., Munoz-Sandoval, A. F., Ruef, M. L., Alvaado, C. G.: *Bateria III Woodcock-Munoz: pruebas de habilidades cognitivas*. Riverside Publishing Company (2005)
49. Xu, W., Callison-Burch, C., Napoles, C.: Problems in current text simplification research: New data can help. Trans. Assoc. Comput. Linguis. **3**(1), 283–297 (2015)
50. Zhao, S., Rudzicz, F.: Classifying phonological categories in imagined and articulated speech. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 992–996. IEEE (2015)