The Cortical Activity of Graded Relevance

Zuzana Pinkosova
zuzana.pinkosova@strath.ac.uk
University of Strathclyde, UK

William J. McGeown
william.mcgeown@strath.ac.uk
University of Strathclyde, UK

Yashar Moshfeghi
yashar.moshfeghi@strath.ac.uk
University of Strathclyde, UK

ABSTRACT

Relevance is an essential concept in Information Retrieval (IR). Recent studies using brain imaging have significantly contributed towards the understanding of this concept, but only as a binary notion, i.e. a document being judged as relevant or non-relevant. While such a binary division is prevalent in IR, seminal theories have proposed relevance as a graded variable; i.e. having different degrees. In this paper, we aim to investigate the brain activity associated with relevance when it is treated as a graded concept. Twenty-five participants provided graded relevance judgements in the context of a Question Answering (Q/A) Task, during assessment with an electroencephalogram (EEG). Our findings show that significant differences in event-related potentials (ERPs) were observed in response to information segments processed in the context of high-relevance, low-relevance and no-relevance, supporting the concept of graded relevance. We speculate that differences in attentional engagement, semantic mismatch (between the question and answer) and memory processing underpin the electrophysiological responses to the graded relevance judgements. We believe our conclusions constitute an important step in unravelling the nature of graded relevance and knowledge of the electrophysiological modulation to each grade of relevance will help to improve the design and evaluation of IR systems.

CCS CONCEPTS

• **Information systems → Information retrieval**; • **Users and interactive retrieval**;

KEYWORDS

Information Retrieval, Relevance Judgement, Graded Relevance, Binary Relevance, Brain Signals, EEG, ERPs

ACM Reference Format:
Zuzana Pinkosova, William J. McGeown, and Yashar Moshfeghi. 2020. The Cortical Activity of Graded Relevance. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3397271.3401106

1 INTRODUCTION

Relevance is a key concept in Information Science and Retrieval [45, 57, 59, 66]. While relevance is considered to be multidimensional [12, 45, 57], dynamic and complex [11, 20, 46, 60], there are still debates around the granularity level of relevance judgements that should be collected [44]. To answer this question, it is crucial to understand what each grade of relevance actually means. The value of evaluating information based on graded relevance has begun to receive attention in recent years both from system [40, 55] and user [2, 15, 48] point of views. This is particularly important since the granularity of relevance judgements in previous studies have been based on investigating this phenomenon indirectly, via some sort of mediator [29, 71]. This, therefore, limits the understanding of how searchers perceive different degrees of information relevance [55]. This paper aims to investigate the neural underpinnings of graded relevance directly.

Relevance is also known to be subjective and difficult to quantify [48] since it depends on a searcher's perception of information relating to a specific Information Need (IN) at a certain point in time [8, 57]. However, given the semantic gap between a searcher's IN and their formulated queries [7, 27, 52], IR systems have employed various techniques to capture the subjective aspect of relevance [2] to improve the effectiveness of retrieved results. Examples of such techniques are explicit [38], implicit (e.g.[24, 36]) and physiological [47] feedback. More recently, researchers have shown the possibility of capturing the neural processes associated with relevance, using brain imaging techniques [2, 15, 16, 22, 25, 28, 33, 37, 48, 65]. These studies have either investigated relevance in the context of word associations (i.e. relevance of a word with respect to another) [15, 16, 65] without subjects experiencing any IN; or investigated relevance in the context of Information Retrieval (IR) when IN has been introduced to subjects. In the latter scenario, relevance was investigated only as a binary notion [2, 19, 22, 24, 25, 28, 37, 48, 49] leaving the graded nature of it unexplored. In this paper, we aim to investigate three fundamental research questions:

- **RQ1:** "Is there a clear, detectable, physical manifestation of graded relevance in human brains?";
- **RQ2:** "Do such manifestations differ when a user perceives different degrees of relevance? i.e., when searchers judge a document as highly relevant, low relevant or non-relevant.".
- **RQ3:** "What is the nature of graded relevance from a cognitive neuroscience perspective?";

Answers to these questions will undoubtedly further our understanding of the concept of relevance and will provide the evidence needed to strengthen the theoretical foundations. This study is the first to incorporate relevance theory and a cognitive neuroscience approach to investigate the neural correlates of graded relevance judgements. Here we utilise an experimental design that enables the investigation of graded relevance within the context of IR and in real-time. In this study, participants will perform a Question Answering (Q/A) retrieval task, while their brain activity is monitored using EEG. Our central aims are to identify: (i) the brain activity associated with distinct graded relevance judgement across time from stimulus onset (ii) test whether there are neural manifestations of cognitive activity underlying each grade of relevance judgement

SIGIR '20, July 25–30, 2020, Virtual Event, China

ACM ISBN 978-1-4503-8016-4/20/07.
https://doi.org/10.1145/3397271.3401106

and (iii) test whether processing distinct grades of relevance is associated with significantly different neural signatures. We are focusing on a range of ERP components - the P300, N400 and P600. Being able to capture brain activity associated with graded relevance could provide better inputs to information systems, which in turn could lead to improved retrieval effectiveness and greater searcher's satisfaction.

2 RELATED WORK

Relevance. Relevance is the fundamental concept in IR [45, 57–59]. It plays a crucial role in the user-system interaction since it is a substantial indicator of system retrieval performance [8, 57]. Despite significant attention dedicated to examining this concept, relevance is still not fully understood, and it is a subject of many ongoing scientific debates. Past research has investigated the concept of relevance at different granularity levels from both the user [29, 71] and system [35, 40] perspective. Within the system side, graded relevance (in comparison to the binary one) has been shown to improve ranking functions [34, 55]. Within the user side, recent research supports the idea of categorical thinking [71], suggesting that users divide retrieved results into 3-5 categories based on relevance [41]. However, levels of granularity were decided based on a self-report mechanism, without clear evidence that those levels have different physical manifestations in the brain. In this paper, we aim to provide evidence for different grades of relevance from a neuroscience perspective.

Capturing Relevance Judgement. Given the importance of the user side of relevance, IR systems have been employing mechanisms and techniques to capture this phenomenon, namely explicit and implicit feedback. Explicit feedback is easy to use, however, difficult to obtain due to the cognitive burden associated with it [47], as the user is required to explicitly state whether presented content is subjectively perceived as relevant or not [67]. Implicit feedback is an unobtrusive data collection method. Popular techniques used to measure implicit relevance feedback are, for example, dwell time (i.e.[36]), eye-tracking and pupillometry [24], and/or the measurements of affective [4, 47] and physiological signals [47]. However, implicit feedback is often found to be noisy, which decreases its accuracy [2]. The findings of novel studies employing neuroscience have shown that brain imaging is an effective method to capture relevance judgement in-real time.

Neuroscience & IR. Recent research has begun to apply brain imaging methods to study aspects of the IR process from a neuroscience perspective. One particular area of emphasis for this research has been to examine the IN process [50, 51]. In addition, it has been found that prediction of the IN state experienced by a user is possible using brain signals [50]. Apart from IN, recent studies have employed brain imaging techniques to gain a better understanding of other parts of the information seeking and retrieval process, such as query formulation [28], search [49, 70] and relevance (e.g. [33]).

Neuroscientific Approach to Relevance. Recent research using a neuroscience approach to investigate relevance might be categorised in two ways based on the context within which the relevance was measured. The first line of brain-imaging research has position relevance within the IR task. For instance, Moshfeghi and colleagues [48] employed functional magnetic resonance imaging (fMRI), to localise differences in brain activity in cortical regions during the processing of relevant vs non-relevant images. The research was able to identify regions engaged in the relevance judgement processing and the increased activation of these regions for relevant items was related to visuospatial working memory [49, 51].

Additionally, examining the neural activity underlying post-relevance judgement revealed that there were differences in the processing of non-relevant and relevant words, that persisted for approximately 260 to 320 ms for relevant words and 500 to 530ms for an irrelevant word [19].

Relevance has been inferred using EEG [25] or in combination with pupillometry or/and eye-tracking devices [22] within the context of the IR task, not only for textual stimuli but also for videos [37] and images [2]. For instance, Allegreti et al. (2015) examined the processing of relevant vs non-relevant images, finding the most significant differences to occur between 500 – 800ms [2]. Kim and Kim (2019) explored the ERPs associated with topical relevance of video skims and classified the data based on two specific ERP components (N400 and P600), which have been shown to be indicators of relevant and non-relevant judgements [37]. Moreover, recent findings have shown that relevance can be predicted in real-time from EEG brain signals and eye movements while the user engages with the system and IR task [28].

Another line of research has examined relevance in the context of word associations, employing EEG in isolation, or in combination with eye gaze [15, 16, 65]. In these scenarios, participants did not experience IN, but they engaged in judging word association to the topic. The findings of these studies have shown that brain signals differ when subjects process relevant vs. non-relevant words across time [16]. Later, Eugster and colleagues (2016) introduced a brain-relevance paradigm enabling recommendation of information to users without any explicit user interaction, based on EEG signals alone evoked by users' engagement with the textual content [15].

Despite valuable insight provided through past neuroimaging studies examining the complex relevance process, it is important to note that relevance has been investigated in binary terms. Hence, it is not clear whether graded relevance judgements are associated with significantly different cognitive processes and whether using more than two relevance categories would be associated with significantly different neural activity. This work takes an essential step to understanding the neural processes involved with graded relevance judgements.

3 METHODOLOGY

Participants. Using opportunistic sampling, data were collected from 25 participants. From these, two were excluded due to excessive movement induced artefacts. Within the 23 remaining participants, there were 14 females and 9 males, the mean age was 25.39, the SD[1] 5.00 years, and the range 19 to 39 years. All participants had normal or corrected-to-normal vision. Overall, nine participants reported to be native English speakers, and the rest had high English proficiency. On average, participants had the experience of 17.65 (±3.46) years of formal education. Only one participant reported being left-handed. Most of the participants were either

[1] Standard Deviation

undergraduate or postgraduate students (60.87%), and the rest were either employed in skilled jobs (30.43%) or unemployed (8.67%).

Design. We used a within-subject experimental design, in which participants performed a Q/A task. The independent variable was graded relevance judgement (with three levels: "No-Relevance" (NONR), "Low Relevance" (LOWR) and "High Relevance" (HIGHR). We controlled the number of relevant vs non-relevant sentences (i.e. answers) presented to the user, but we did not control the number of words presented to each participant. This allowed us to simulate an information search and retrieval, as we did not enforce participants to go through the whole answer, once the relevance judgement has been made. The dependent variables were the EEG signals of the brain, gathered from the users during the Q/A task.

ERP Components. To capture the brain response associated with relevance, we measured small voltage fluctuations across the scalp, that are time-locked to a specific event or stimulus. These are known as ERPs. In this study we focused on the main ERP components associated with relevance (i.e. P300, N400, P600) (e.g. [15, 37]). The P300 component is a positive-going voltage deflection that peaks around 300ms post-stimulus [43]. This time-locked ERP component has been included in the analysis as it is thought to reflect the amount of cognitive resource employed for information processing [53]. The N400 component is a negative-going potential, reaching its peak at 400ms from the stimulus presentation [43]. This ERP component is usually elicited by irrelevant/incongruent stimuli, and its amplitude is associated with semantic integration. Words that can be integrated more easily to the context elicit smaller (more positive) amplitude and vice versa [43]. The P600 component is characterised as a positive deflection, peaking around 600ms from stimulus onset [43]. It is elicited during cognitive processing involved in text comprehension [32].

Apparatus. The experiment was conducted in a neurophysiology laboratory at the University of Strathclyde. Stimuli were presented with E-Prime 2.0, installed on a PC with 22" Mitsubishi Diamond Pro 2040u NF CRT monitor with the resolution of 2048 x 1536 and a refresh rate of 75 Hz. The screen was positioned approximately 60cm from the participant. A soft light was used to avoid distractions. Participants were able to interact with the PC using the keyboard. EEG recordings were obtained using a 128-channel geodesic sensor net (Electrical Geodesics Inc) using Net station 4.3.1 software. We aimed to keep electrode-scalp impedance below 50 kΩ for all participants. The sampling rate was set at 1000 Hz. To set the system for recording, we followed Electrical Geodesics Inc guidelines. The EEG sensor net was soaked in the potassium chloride (KCl) solution for conductivity. Each participant's head was measured to determine the correct EEG net size, and the net was positioned using standardised procedures, ensuring that the vertex is halfway between the inion and nasion and halfway between both bilateral preauricular points. The EEG net VREF electrode was positioned at the marked vertex. A NetAmps 200 amplifier was used for synchronisation between the behavioural responses of participants and their brain signals. Finally, we used entry, Post-task and Exit Questionnaires for each participant.

Questionnaires. At the beginning of the experiment, an Entry Questionnaire was introduced to gather background and demographic information about the participant and participants were screened to assess whether they are eligible to take part in the study. Any participants who had existing conditions (i.e. neurological or psychiatric disorder) that might impact the EEG signal recordings were excluded. Once participants completed the task, they were asked to fill in a Post-task Questionnaire asking them to rate how they perceived the encountered task. Finally, all participants completed an Exit Questionnaire, which examined the perception of their overall performance.

3.1 Procedure

The user study was carried out in the following manner. Ethical permission for the study was obtained from the Department of Computer and Information Sciences Ethics Committee at the University of Strathclyde. The formal meeting with the participants took place in the laboratory setting. At the beginning of the session, all the participants were briefed as to the procedure and the purpose of the experiment through the information sheet. Then, they were asked to provide informed consent. All participants were notified about their right to withdraw at any time during the study, without giving a reason and without any consequences. After that, participants were instructed to fill in an Entry Questionnaire. Prior to the main experimental trials, participants underwent training, which resembled the main experimental task. This ensured that all the participants have a good general understanding of the procedure. The training was not limited by time and participants were able to repeat it if required. In total, each participant completed 64 trials. To avoid fatigue, the stimuli were presented to the participants in two blocks of 32 trials each, separated by a break. The average duration of the main experimental task was approximately 53.57 minutes and each participant was presented on average with 799.61 words (±161.11).

3.1.1 *Procedure of the Main Experimental Task.*

A schematic representation of the main experimental task is illustrated in Figure 1. At the beginning of the trial, participants were presented with the task instructions. Next, they viewed a question randomly selected from the data set. Once participants read and fully understood the question, they were presented with the fixation cross, to indicate the location of the answer presentation. The answer was then presented word by word in the middle of the screen to control free-viewing and to minimise saccade-related measurement artefacts. This approach has traditionally been applied in the ERP studies examining neurological correlates of reading [14]. As a result, the ERPs were time-locked to the word presentation. Each word was presented for 950ms. This reading pace was appropriate to model fluent reading as much as possible and to avoid the presence of the overlapping effect of two consecutive words on the ERPs [15]. Participants were instructed to read individual words, that formed sentences representing either a relevant or non-relevant answer. They had an option of terminating the presentation of the words and continuing to the next step. As brain activity was recorded throughout the task, to avoid the possibility of confounding hemispheric effects, the hand used for button responses during the task was counter-balanced across participants (each participant used only the left hand or right hand, as instructed).

After that, participants were again presented with the same answer appearing on the screen in the same order (up to the point of presentation abandonment). In this stage of the trial, the answers

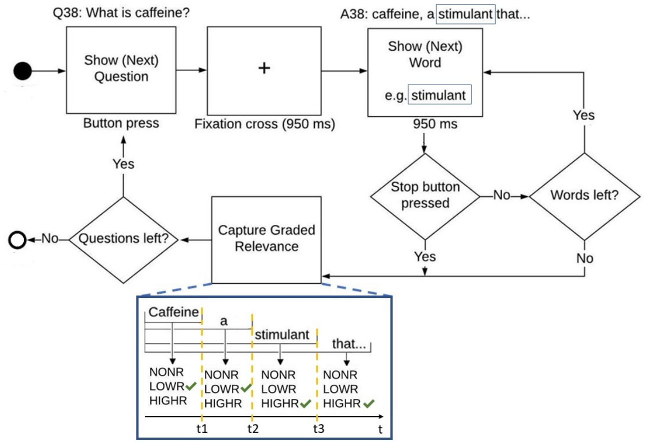


Figure 1: The figure shows the structure of a task. From the left (START), the question is presented in a randomised order. Once ready, the participant presses a button on the keyboard to start. Firstly, a fixation cross was presented for 950 ms. Then, an answer is presented word by word. Each word is shown for 950ms. The participant is able to stop the word presentation once enough information is gathered. Next, participants proceed with the graded relevance judgement (HIGHR, LOWR, NONR), with no time restrictions. Within this step, the participant judges relevance based on the subjectively perceived information accumulation process. Hence, while the answers were presented word-by-word again, participants were asked to submit a subjective graded relevance judgement for the information segment presented to them from the first stimulus (i.e. word) up to and including the current stimulus. The process is repeated for all 64 questions (END).

were presented word by word as continuous text. Participants were instructed to assign a (subjectively perceived) graded relevance judgement (NONR, LOWR, and HIGHR) for each word segment of the answer while taking into account information accumulation, rather than judging words in isolation with relationship to the question. The graded relevance judgements were then assigned retrospectively, enabling this detailed information to be applied to the corresponding EEG segments of interest. The interpretation of each graded relevance category depended on each participant's subjective understanding, which enabled capturing the subjective nature of relevance judgement [57].

All text events were presented in Arial font, size 16. After completion of the main experimental task, participants were instructed to fill out a Post-Task Questionnaire and an Exit Questionnaire. All participants were debriefed after the experiment.

Question/Answering Dataset. For our experiment, we adopted the Question Answering dataset developed and used by Moshfeghi et al. [51]. We chose this dataset since it has been widely in previous studies investigating IR phenomena from a neuroscience perspective [49, 51]. The adopted dataset was further expanded through the selection of additional questions and answers from TREC-8 and TREC-2001. We chose these two Tracks since they (i) provide the correct answer to the question, and (ii) they are independent of one another. We ensured that selected questions were not ambiguous or time-dependent, making sure the answers provided in the Track were still appropriate by manually validating each answer using a search engine.

The created data set was then split into two parts (Data Set A and Data Set B)[2], each containing 64 questions, answers and relevance assessments in total. The decision of splitting the dataset in two was made during the pilot study after observing the length of the experiment and to reduce the fatigue level of the participants. We made sure that Data Set A and B had similar characteristics (as shown in Table 1) to avoid introducing any bias in our results. For example, we ensured that both data sets were balanced to be apriori 50% relevant and 50% no-relevant. In addition, we balanced the answer length (long vs short), and question difficulty (easy vs difficult)[3]. This was done to reduce any potential bias that might occur from the emphasis of one particular question/answer type.

An example of an easy question presented to the participants was "Who was Galileo?", which was followed by the short, relevant answer "In 1642, astronomer Galileo died in Arcetri, Italy.". The order of the questions was randomised for each participant. This randomisation ensured that the recorded signals and effects were related to the users' subjective relevance judgement of the presented stimuli, and not related to the stimulus presentation frequency, potentially causing an oddball effect [15]. We ensured that answers from both data sets were approximately the same lengths, based on the mean number of characters per answer category. Participants were then randomly assigned to one of the two data sets.

Table 1: The Mean length and SD of the answers based on category for Data Set A and Data Set B

	Data Sets			
	A		B	
Answer Length	Mean $_A$	SD$_A$	Mean $_B$	SD $_B$
Total	14.88	6.24	15.02	6.31
Relevant	15.00	6.24	14.97	6.34
No-Relevant	14.75	6.33	15.06	6.38
Difficult	15.19	6.42	14.69	6.02
Easy	14.56	6.13	15.34	6.67
Long	20.84	2.17	21.09	1.99
Short	8.91	0.89	8.94	0.80

Pilot Studies. Prior to running the main study, a pilot study was performed employing 4 participants, whose data were not included in the analysis. Detailed feedback obtained from participants involved in the pilot study was used to adjust the study presentation and design. After the final pilot study, it was determined that the participants were able to complete the user study without problems and that the system was capturing all necessary data.

Pre-processing steps. The brain activity was recorded from participants as the answer was unfolding to them. The obtained signals were then matched with the graded relevance judgement up to the point where the participant stopped the answer presentation. All collected data were first visually inspected. Then we

[2] The Question Answering Data Set A and the Data Set B are available upon request.

[3] To assess the difficulty level, two annotators separately judged the difficulty of the questions (i.e. hard or easy) and then selected a subset of 128 questions where both annotators agreed upon their difficulties, i.e. 64 of them were hard, and 64 were easy questions.

applied a low-pass filter of 30Hz. By attenuating the higher frequencies, the anti-aliasing is reduced, which is an important step before data down-sampling. We have then down-sampled the data from 1000Hz to 250Hz, and then a high pass filter of 0.3Hz was applied. Filtering is a common pre-processing procedure, used to attentuate frequencies commonly associated with noise rather than signal of interest. Downsampling, another commonly applied procedure, is used to reduce file size for easier manipulation (a level was chosen to maintain the signal quality). We then automatically rejected bad channels (EEG data-streams/sensors that were not functioning properly during the data acquisition and that were high in noise throughout the task). On average, we removed 13.08 bad channels (±7.43). The CleanLine EEGLAB plugin was used to filter line noise. The vertex (Cz) sensor was initially used as the reference electrode, but re-referencing to average (across all electrodes) was subsequently performed (to provide an approximation of zero microvolts for the reference at each timepoint). All epochs (the time-windows of interest) were then extracted from 200ms before stimulus presentation to 950ms afterwards. To detect and remove components associated with ocular, cardiac and muscular artefacts based on their power spectrum and time-course, we performed Independent Component Analysis (ICA). ADJUST was then used to remove artefacts. A mean number of 14.91 (±8.26) components were identified and removed. Bad channels were interpolated (reconstructed) using a spherical interpolation method. Next, we removed the outermost belt of electrodes of the sensor net (19 peripheral channels: E43, E48, E49, E56, E63, E68, E73, E81, E88, E94, E99, E107, E113, E119, E120, E125, E126, E127, E128)[9]. Epochs were then extracted again from 100ms before stimulus presentation to 950ms afterwards based on the stimulus label ('NONR', 'LOWR' and 'HIGHR'). All epochs were baseline-corrected and bad epochs were then automatically identified and removed using the epoch rejection plugin. On average, we rejected 35.91 (±12.26) of NONR epochs, 27.83 (±13.86) HIGHR epochs and 23.89 (±12.38) LOWR epochs.

3.2 ERP - Components of Interest

To analyse our three components of interest (P300, N400, and P600), we selected three electrode configurations (that partially overlap) and three time-windows (that do not overlap), based on the findings of previous studies as well as a visual inspection of the grand average waveforms. For the P300 component, we chose a time-window of 220–440 ms from stimulus onset [3, 53] and the representative subset of 22 centro-parietal channels over the midline (E55, E62), left (LH: E30, E36, E37, E42, E52, E53, E54, E60, E61, E67) and right hemispheres (RH: E77, E78, E79, E85, E86, E87, E92, E93, E104, E105) as shown in Figure 2a. The electrode selection was based on the previous literature examining the P300 component through the calculation of mean amplitudes [54]. The mean amplitudes for N400 component were measured between 400 – 600 ms [18]. We identified 14 electrodes, covering parietal-temporal regions bilaterally (LH: E31, E37, E42 E53, E54, E61 and RH: E79 E78, E86, E87, E80, E93), including mid-line channels (E55, E62) [26] as shown in Figure 2b. To analyse the P600 component, we again selected the electrodes and the time-window that the component is typically measured most strongly. We selected the time window between 550 – 750 ms [6] and a bilateral group of electrodes covering central, parietal and temporal regions (E52, E53, E54, E59, E60, E61, E66, E67, E72, E77, E78, E79, E80, E85, E86, E87, E92, E93) [21] as shown in Figure 2c.

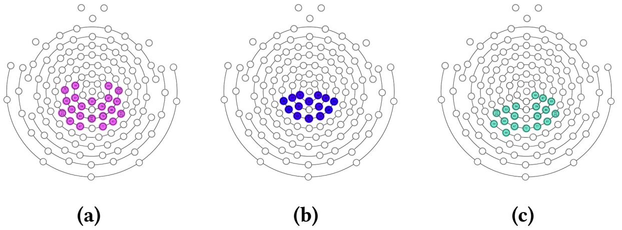


Figure 2: Hydrocel Geodesic Sensor Net-128 channel map. The different colours indicate the selection of electrodes for each component of interest (a) P300, (b) N400, (c) P600

4 RESULTS

Questionnaire Analysis. Before analysing the results we investigated participants' perception of the task. In the Post-task Questionnaire, we asked participants how they found the task, questions presented to them, familiarity with questions and whether they felt comfortable throughout the task (answers: 1: "Strongly Agree", 2: "Agree", 3:"Somewhat Agree", 4:"Neither Agree or Disagree", 5:"Somewhat Disagree", 6: "Disagree", 7: "Strongly Disagree"). The results shown in Figure 3 indicate that participants found the task (M = 1.96, SD = 1.26) and questions (M = 1.91, SD = .75) rather interesting. Perceived difficulty of the task (M = 3.74, SD = 1.84) and questions (M = 4.13, SD = 1.58) was rated as moderate. Presented questions (M = 1.96, SD = 1.33) and task in general (M = 2.00, SD = 1.21) were considered as readable. Additionally, both, questions (M = 2.17, SD = 1.50) and task (M = 2.13, SD = 1.18) were also considered as understandable. On average, participants felt moderate physical comfort (M = 2.74, SD = 1.57) and task was not rated as stressful (M = 4.57, SD = 1.85). Questions selected for the experiment were perceived by participants as moderately familiar (M = 3.26, SD = 1.48) and relevant to them (M = 2.48, SD = 1.53).

Using the Exit Questionnaire, we examined participants' perception of their overall performance during the task (answers: 1: "Strongly Agree", 2: "Agree", 3:"Somewhat Agree", 4:"Neither Agree or Disagree", 5:"Somewhat Disagree", 6: "Disagree", 7: "Strongly Disagree"). Overall, the results indicate that participants felt that they had enough time to press a button to terminate the answer presentation (M = 1.61, SD = .84). Additionally, they found the speed of the word presentation (M = 1.36, SD = .58) to be appropriate for reading. The font size (M = 1.30, SD = .70) and monitor luminance (M = 1.87, SD = 1.14) were also rated to be task appropriate. Most of the participants felt comfortable (M = 2.23, SD = 1.19) during the task and the EEG cap was not causing them any discomfort (M = 1.65, SD = .98). Participants found following the procedure to be easy (M = 1.70, SD = .88), with instructions being clear (M = 1.26, SD = .45) and they were mostly satisfied with their performance (M = 2.04, SD = .98).

Graded Relevance Component Analysis. The ERP analysis relied upon a participant's graded relevance judgement assessment (coded as NONR, LOWR, and HIGHR) to the presented information.

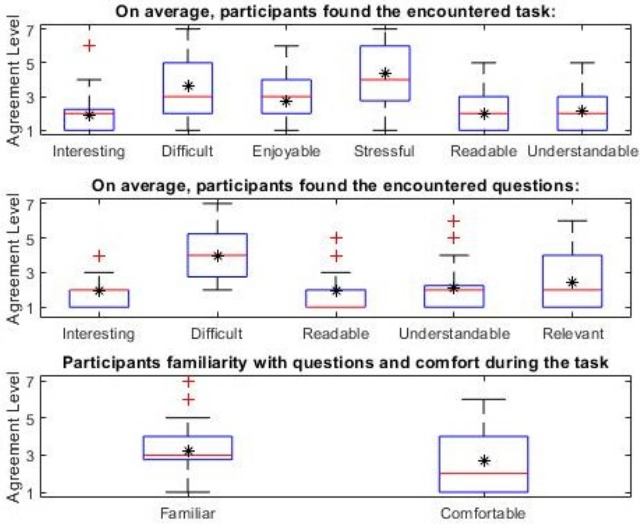


Figure 3: Post-task Questionnaire: Box plot of the participants' perception of the encountered task. The asterisk (*) represents the mean value, while cross (+) represents the outlier value.

Table 2 shows the statistics for the collected graded relevance judgements across participants. Analyses focused on the P300, N400 and P600 components during the graded relevance judgement process. The electrode configurations and time-windows for each component was based on the previous literature and adjusted by visual inspection (see Section 3.2). The ERPs were obtained by averaging across the selected electrode configurations. For this study, we calculated the mean amplitude, which enabled us to factor out latency jitter and to obtain more robust results [10, 42].

Table 2: The summary statistics of the relevance judgements made by participants on the observed stimuli

Condition	Avg	SD	MIN	MAX
NONR	399.39	138.90	182	686
LOWR	162.52	71.34	57	310
HIGHR	228.96	106.70	42	496
Total	793.44	159.88	358	961

To avoid possible misinterpretations when comparing the three conditions for peak amplitude of the components (e.g., when high frequency fluctuations may influence the results), mean amplitudes were instead used. These were determined by averaging the activity within each time window of interest for each of the electrode groupings. To investigate the statistical significance of differences between recorded neurological signal associated with graded relevance judgements (NONR, LOWR, and HIGHR) for P300, N400, and P600 components, we applied repeated measures ANOVAs. To do so, we first investigated whether the assumptions that are required for repeated measures ANOVA are met. The dependent variables were approximately normally distributed. Mauchly's Test was used to investigate whether the assumption of sphericity was met for each condition. Since this was not the case, we used the Greenhouse-Geisser method, which enabled us to obtain F-ratios with greater accuracy. The epsilon (ϵ), estimating the amount of sphericity, is reported. For scenarios in which the ANOVA test indicated statistically significant differences between the recorded neurological signals, we conducted pairwise comparisons of the conditions (i.e. HIGHR vs LOWR, HIGHR vs NONR, etc.) using Bonferroni tests. Results were considered significant at $p < 0.01$.

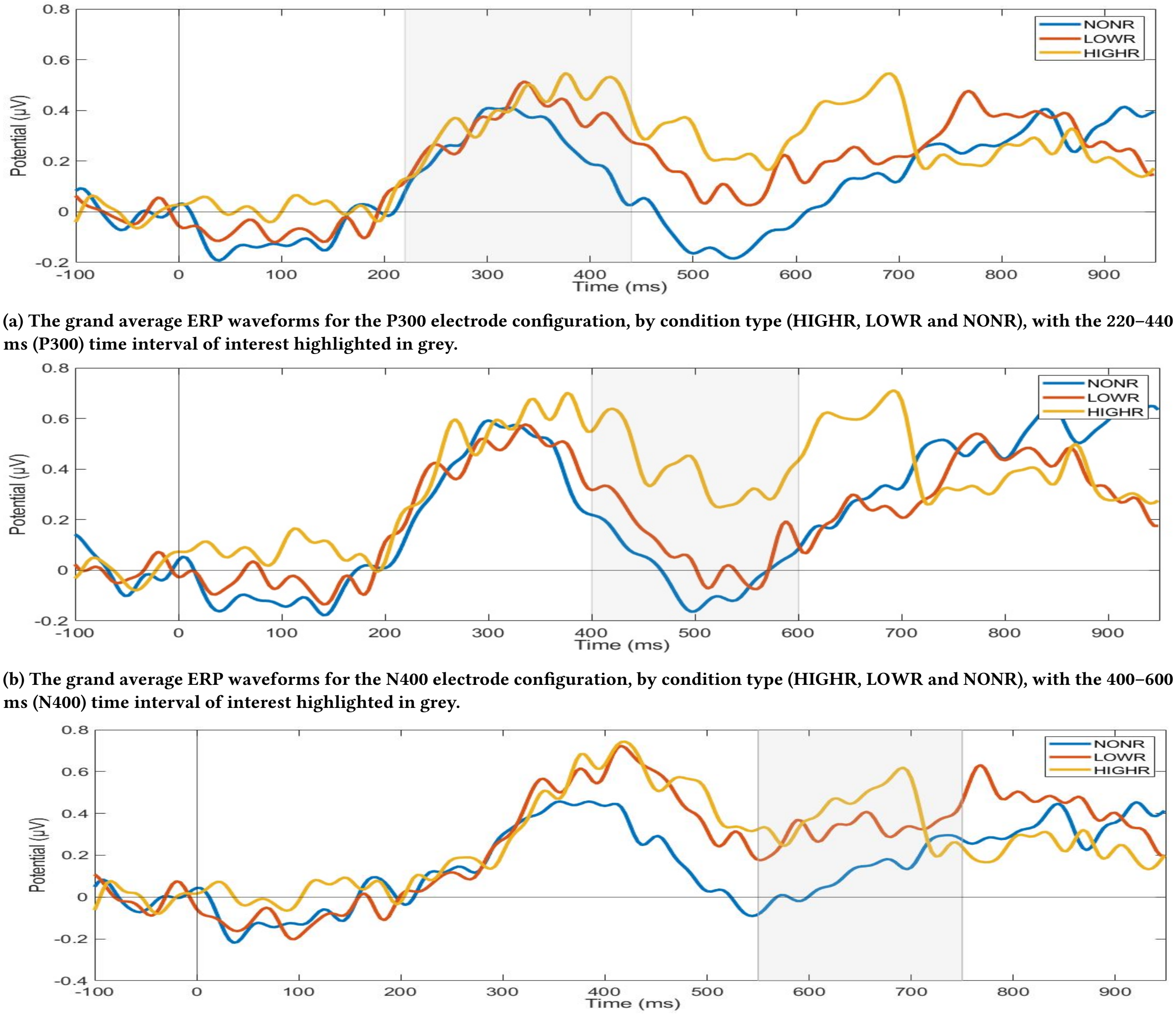
4.0.1 Main Findings. Our experimental results show that significant differences exist in brain activity when judging information as having high-relevance, low-relevance, or no-relevance. Differences were present in all components of interest (the P300, N400, and P600 components), which suggests that a variety of distinct cognitive processes are underpinning the graded relevance evaluations.

4.0.2 P300. The P300 component waveforms grand-averaged across participants for the NONR, LOWR and HIGHR conditions are shown in Figure 4a. Component latency considered for the analysis is highlighted in grey. The P300 amplitude was highest for HIGHR condition (M = .39 μV, SD = .11), followed by the LOWR condition (M = .35 μV, SD = .10) and NONR condition (M = .26 μV, SD = .11). The results of repeated measures ANOVA with Greenouse-Geisser correction showed that the brain activation associated with the P300 component differed significantly across NONR, LOWR and HIGHR conditions [F(1.16, 63.98) = 42.42, $p < 0.001$, $\epsilon = .44$]. Post-Hoc tests using the Bonferroni correction revealed that the highest mean difference was between HIGHR and NONR conditions (M_{diff} = .13 μV , $p < .001$), following the LOWR and NONR (M_{diff} = .0.83 μV, $p < .001$) and LOWR and HIGHR conditions (M_{diff} = .05 μV, $p < .001$).

Information judged as HIGHR was associated with significantly greater P300 amplitudes across central and centro-parietal electrode sites when compared to LOWR and NONR. Hence, during this early stage of implicit relevance judgement, users' selective attention is allocated towards highly relevant stimuli, which are also easier to process in terms of cognitive load [1, 53]. These findings are consistent with the previous literature, showing that the degree of subjectively perceived information relevance is proportional to the P300 component amplitude [2, 23].

4.0.3 N400. Figure 4b presents the N400 component waveforms for the NONR, LOWR and HIGHR conditions. The mean negativity of the N400 component was the most prominent in NONR condition (M = .00 μV, SD = .11), following the LOWR (M = .10 μV, SD = .13) and HIGHR condition (M = .40 μV, SD = .11). A repeated measures ANOVA carried out with Greenouse-Geisser correction revealed significant differences in the mean amplitude across the graded relevance judgement conditions [F(1.64, 82.09) = 784.16, p < 0.001, $\epsilon = .94$]. Post-Hoc comparisons using the Bonferroni correction indicated that the highest mean amplitude differences for the N400 component were measured between the HIGHR and NONR conditions (M_{diff} = .40 μV , $p < .001$), followed by the LOWR and HIGHR (M_{diff} = .0.30 μV, $p < .001$) and LOWR and NONR conditions (M_{diff} = .10 μV, p < .001).

The processing of NONR, LOWR and HIGHR information was associated with negative-going deflection, which was the most prominent for the NONR condition, which is consistent with previous literature [15, 37, 65]. Although highly relevant information reduces the N400 amplitude [63], our results indicate that the N400



(a) The grand average ERP waveforms for the P300 electrode configuration, by condition type (HIGHR, LOWR and NONR), with the 220–440 ms (P300) time interval of interest highlighted in grey.

(b) The grand average ERP waveforms for the N400 electrode configuration, by condition type (HIGHR, LOWR and NONR), with the 400–600 ms (N400) time interval of interest highlighted in grey.

(c) The grand average ERP waveforms for the P600 electrode configuration, by condition type (HIGHR, LOWR and NONR), with the 460–680 ms (P600) time interval of interest highlighted in grey.

Figure 4: The grand average ERP waveforms from the selected electrode configurations. Time intervals of interest (in milliseconds) are highlighted in grey for each component. Change in the amplitude of the ERP relative to the prestimulus baseline (-100ms to 0ms) is represented on the y-axis (in microvolts), and time following stimulus onset (from 0ms onwards) is represented on the x-axis.

deflection was also salient for the LOWR condition. The observed difference between the HIGHR and LOWR category for the N400 component might indicate that LOWR (and NONR) stimuli require significantly greater cognitive effort to process and integrate within the given context [13]. The context within this experiment has been provided through the question, and hence it is possible to assume that higher N400 amplitudes elicited during NONR condition signalise contextual violation [5] and a degree of uncertainty during LOWR condition [64].

4.0.4 P600. The grand-averaged P600 waveforms are displayed in Figure 4c for each condition (NONR, LOWR and HIGHR). The mean amplitude of the P600 component was the highest in HIGHR condition, (M = .40 μV, SD = .13), following the LOWR (M = .33 μV, SD = .06) and NONR condition (M = .12 μV, SD = .11).

The results of repeated measures ANOVA carried out using Greenhouse-Geisser correction method revealed significant differences in the mean amplitude across the graded relevance judgement conditions [F(1.34, 67.07) = 108.31, p < 0.001, ϵ = .68]. Post-Hoc tests using Bonferroni correction revealed that the highest mean difference was between HIGHR and NONR conditions (M_{diff} = .27 , p < .001), following the LOWR and NONR (M_{diff} = .21, p < .001) and HIGHR and LOWR conditions (M_{diff} = .07, p < .01) displayed in Figure 4c over the centro-parietal areas. Our findings are in alignment with previous studies, suggesting that processing of NONR information is associated with low P600 amplitudes [15, 37]. As the P600 amplitude is highest for the HIGHR condition, this may reflect that the amount of information carried by the processed term is higher in comparison to the LOWR condition [32].

5 DISCUSSION

The paradigm developed for this study enabled judgements of relevance to be assessed in a graded fashion, and for the corresponding neural activity to be recorded. Specifically, participants reflected on sentences that they had seen in response to a question and reported after each word of the sentences what their perception of relevance was at that time (HIGHR, LOWR or NONR). Participants were, therefore, processing each word of the sentence within the context of whether they subjectively perceived the information segment at that time to be relevant to the question.

The key findings which emerged from the study are that levels of neural activity across time are dependent on whether the person perceived the sentence as of high relevance, low relevance or no-relevance to the question (this finding addresses **RQ1)**. Significant differences were detected in the three late-stage ERPs of interest (P300, N400 and P600), that followed words that were processed in the context of whether the information segment was deemed to be of high relevance, low relevance or no-relevance to the question (this finding addresses **RQ2**). The differences in neural activity suggest that during assessment of relevance, a variety of cognitive processes are relied upon to different degrees: For example the higher the relevance, it may be, the greater the attentional engagement, the higher the perception of semantic relatedness (the lower the semantic incongruency between context of the question and the answer), and the greater the requirement for engagement of memory (relevant information might be deemed more important to encode and recall than irrelevant information) (this discussion provides an early step towards answering **RQ3**).

High Relevance: The results suggest that greater attentional resources are allocated to highly relevant stimuli, as indicated through the differences observed within the P300 component. P300 amplitude has been shown to be proportional to attentional engagement [30]. Greater P300 amplitude has also been suggested to reflect the quantity of information transmitted [56], the quantity of useful information [31], relevance to the self [23], or relevance to the task [17, 62], or to judgements of relevance specifically [2]. Eugster et al. [15], when assessing term relevance, did not find a difference in the P300 between relevant and irrelevant words. It is possible that our P300 measure is influenced by the N400 deflections, but the P300 and N400 may both be modulated during the assessment of relevance. The underlying processes also might not be wholly independent [5]. Future research will provide clarification.

There was a clear reduction of the N400 within the time-interval of 400 – 600 ms. Typically, in studies of language, the N400 provides an index of semantic relatedness; it is larger when there is a semantic mismatch than semantic congruency (see e.g., [61]). Given the task requirements of the current study and the differences we observed in the N400 in response to graded relevance, it seems that in this case the component is modulated not necessarily in response to the meaning of the word, but to the relatedness of the sentence to the question. Words processed in the context of high relevance are semantically aligned to the question, which likely explains the attenuated N400 response. The N400 results fit closely with the study by Eugster and colleagues [15], who found a reduced N400 for relevant words, compared to irrelevant words.

In our study, words linked to sentences of high relevance were associated with the highest P600 amplitudes. Similarly, Eugster et al. [15] found that their relevant words elicited larger P600 components than irrelevant words. The link between higher relevance and P600 amplitude is not completely clear. The P600 has often been associated with syntactic processing, and later research has flagged a semantic-thematic role (however, larger P600s are found due to violations – see e.g., [39]). A more likely reason, or at least a partial explanation, may be that the late stage positivity is instead linked to memory processing (e.g., through a process such as recognising that the answer is relevant and is linked to the question). A late positive complex has been observed during memory recognition, it is higher for old versus new stimuli, occurs at around 600ms after a stimulus and also has a central posterior topography [69].

It should be noted that the task requirements of our study differed from that of Eugster and colleagues, due to our assessment of graded relevance (high, low, no), as opposed to a binary judgement. A further key difference was in the way the relevance judgements were made - participants in our study did not evaluate individual words for relevance, but rather, in response to pre-established IN, they repeatedly assessed the relevance of the entire information segment each time a new word was presented.

No-Relevance: Words processed without any perceived relevance to the question had the lowest mean P300 values, the greatest N400, and the lowest P600. Conversely to the words processed in the high relevance context, the words processed that are not relevant to the question may have a low P300 due to cognitive factors such as low attentional engagement, a large N400 due to a mismatch between the semantic material offered in the answer given the context of the question (larger semantic incongruity), and a lower P600 due to reduced memory processing given that the answer is not relevant to the question (e.g., information to be retained results in larger P600 amplitudes than information to be forgotten - see e.g.[68].

Low Relevance: A crucial question relates to the manner in which words are processed in the low relevance context. Specifically, is the processing of these words more similar to words viewed in the high relevant context or the non-relevant context? The ERP component amplitudes for the words processed in the context of low relevance fell somewhere in between those processed in the context of high relevance and those processed in the context of low relevance (significant differences were seen for all three components across the three conditions). A key difference appears to be in the N400. Although the N400 to the word segments deemed of low relevance differed significantly from those deemed of high and of no-relevance, the waveform pattern more closely resembled the waveform for no-relevance, suggesting that low relevance segments may be perceived more like non-relevant segments, until a critical threshold is reached. On the other hand, although low relevant segments differed significantly from the high relevance and no-relevance segments for the P600 component, the low relevance segment waveform more closely resembled the high relevance waveform. This may reflect a dual nature of the stimuli evaluated as having low relevance. When words are processed in the context of low relevance, the sentences semantically have not crossed any critical threshold for relevance (there is still semantic incongruency), whereas there is still the potential for relevance to increase in these stimuli, so they must be processed adequately in memory and matched to the question (in a similar fashion as words in the context of high relevance).

6 CONCLUSION AND FUTURE DIRECTIONS

In conclusion, our findings provide support for the concept of graded relevance, given the clear differences in neural activity when information segments are perceived as having high relevance, low relevance or no-relevance. The P300, N400 and P600 all differed due to the perceived relevance of the answer. Despite a number of ERP components being identified that relate in different ways to the level of relevance perceived, it will be important to understand how robust/reliable these differences are and how alterations in the questions (e.g., the difficulty level) or in the answer (e.g., the length of the response) may interact with these features. Future competing hypotheses may be whether there may be a dual nature to relevance processing, wherein cases of ambiguity (e.g., low relevance - where the person reserves judgement while seeking more information) some neural features may be more like non-relevant features (e.g., as seen in pattern of the N400 waveforms in the current study), and others are more like high relevance features (e.g., the pattern of P600 waveforms), or alternatively perhaps whether more fine-grained distinctions may be reflected in the neural signal. More fine-grained distinctions could be tested with relevance scales that offer more options for evaluation than we did in the current study (e.g., no, low, moderate and high relevance; or a judgement scale with even more options could be used). These results further our understanding of the concept of relevance and provide evidence needed to strengthen its theoretical foundations. Finally, we believe our conclusions constitute an important step in unravelling the nature of graded relevance and knowledge of the electrophysiological modulation to each grade of relevance.

Acknowledgement: This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/R513349/1].

REFERENCES

[1] Lubna Ahmed and Jan W de Fockert. 2012. Working memory load can both improve and impair selective attention: evidence from the Navon paradigm. *Attention, Perception, & Psychophysics* 74, 7 (2012), 1397–1405.

[2] Marco Allegretti, Yashar Moshfeghi, Maria Hadjigeorgieva, Frank E. Pollick, Joemon M. Jose, and Gabriella Pasi. 2015. When Relevance Judgement is Happening? An EEG-Based Study. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 719–722. https://doi.org/10.1145/2766462.2767811

[3] Hafeez Ullah Amin, Aamir Saeed Malik, Nidal Kamel, Weng-Tink Chooi, and Muhammad Hussain. 2015. P300 correlates with learning & memory abilities and fluid intelligence. *Journal of neuroengineering and rehabilitation* 12, 1 (2015), 87.

[4] Ioannis Arapakis, Joemon M. Jose, and Philip D. Gray. 2008. Affective feedback: an investigation into the role of emotions in the information seeking process. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)*. Association for Computing Machinery, New York, NY, USA, 395–402. https://doi.org/10.1145/1390334.1390403

[5] Yael Arbel, Kevin M Spencer, and Emanuel Donchin. 2011. The N400 and the P300 are not all that independent. *Psychophysiology* 48, 6 (2011), 861–875.

[6] Yossi Arzouan, Abraham Goldstein, and Miriam Faust. 2007. Brainwaves are stethoscopes: ERP correlates of novel metaphor comprehension. *Brain research* 1160 (2007), 69–81.

[7] N. J. Belkin and H. M. Brooks. 1982. Ask for information retrieval: Part ii. results of a design study. *Journal of Documentation* 38, 3 (3 1982), 145–164. https://doi.org/10.1108/eb026726

[8] Pia Borlund. 2003. The concept of relevance in IR. *Journal of the American Society for information Science and Technology* 54, 10 (2003), 913–925.

[9] Marta Calbi, Francesca Siri, Katrin Heimann, Daniel Barratt, Vittorio Gallese, Anna Kolesnikov, and Maria Alessandra Umiltà. 2019. How context influences the interpretation of facial expressions: a source localization high-density EEG study on the "Kuleshov effect". *Scientific reports* 9, 1 (2019), 1–16.

[10] Peter E Clayson, Scott A Baldwin, and Michael J Larson. 2013. How does noise affect amplitude and latency measurement of event-related potentials (ERPs)? A methodological critique and simulation study. *Psychophysiology* 50, 2 (2013), 174–186.

[11] Colleen Cool, Nick Belkin, Ophir Frieder, and Paul Kantor. 1993. Characteristics of text affecting relevance judgments. In *National online meeting*, Vol. 14. LEARNED INFORMATION (EUROPE) LTD, 77–77.

[12] Erica Cosijn and Peter Ingwersen. 2000. Dimensions of relevance. *Information Processing & Management* 36, 4 (2000), 533–550.

[13] J Bruno Debruille. 2007. The N400 potential could index a semantic inhibition. *Brain research reviews* 56, 2 (2007), 472–477.

[14] Olaf Dimigen, Werner Sommer, Annette Hohlfeld, Arthur M Jacobs, and Reinhold Kliegl. 2011. Coregistration of eye movements and EEG in natural reading: analyses and review. *Journal of Experimental Psychology: General* 140, 4 (2011), 552.

[15] Manuel JA Eugster, Tuukka Ruotsalo, Michiel M Spapé, Oswald Barral, Niklas Ravaja, Giulio Jacucci, and Samuel Kaski. 2016. Natural brain-information interfaces: Recommending information by relevance inferred from human brain signals. *Scientific reports* 6 (2016), 38580.

[16] Manuel J.A. Eugster, Tuukka Ruotsalo, Michiel M. Spapé, Ilkka Kosunen, Oswald Barral, Niklas Ravaja, Giulio Jacucci, and Samuel Kaski. 2014. Predicting Term-Relevance from Brain Signals. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '14)*. Association for Computing Machinery, New York, NY, USA, 425–434. https://doi.org/10.1145/2600428.2609594

[17] Lawrence A Farwell and Emanuel Donchin. 1991. The truth will out: Interrogative polygraphy ("lie detection") with event-related brain potentials. *Psychophysiology* 28, 5 (1991), 531–547.

[18] Bálint Forgács, Eugenio Parise, Gergely Csibra, György Gergely, Lisa Jacquey, and Judit Gervain. 2019. Fourteen-month-old infants track the language comprehension of communicative partners. *Developmental science* 22, 2 (2019), e12751.

[19] Aline Frey, Gelu Ionescu, Benoit Lemaire, Francisco López-Orozco, Thierry Baccino, and Anne Guérin-Dugué. 2013. Decision-making in information seeking on texts: an eye-fixation-related potentials investigation. *Frontiers in systems neuroscience* 7 (2013), 39.

[20] Thomas J Froehlich. 1994. Relevance reconsidered—Towards an agenda for the 21st century: Introduction to special topic issue on relevance research. *Journal of the American Society for Information Science* 45, 3 (1994), 124–134.

[21] Felicidad Marcia Garcia. 2017. *Brain responses to contrastive and noncontrastive morphosyntactic structures in African American English and Mainstream American English: ERP evidence for the neural indices of dialect.* Ph.D. Dissertation. Columbia University.

[22] Jan-Eike Golenia, Markus A Wenzel, Mihail Bogojeski, and Benjamin Blankertz. 2018. Implicit relevance feedback from electroencephalography and eye tracking in image search. *Journal of neural engineering* 15, 2 (2018), 026002.

[23] Heather M Gray, Nalini Ambady, William T Lowenthal, and Patricia Deldin. 2004. P300 as an index of attention to self-relevant stimuli. *Journal of experimental social psychology* 40, 2 (2004), 216–224.

[24] Jacek Gwizdka. 2014. Characterizing Relevance with Eye-Tracking Measures. In *Proceedings of the 5th Information Interaction in Context Symposium (IIiX '14)*. Association for Computing Machinery, New York, NY, USA, 58–67. https://doi.org/10.1145/2637002.2637011

[25] Jacek Gwizdka, Rahilsadat Hosseini, Michael Cole, and Shouyi Wang. 2017. Temporal dynamics of eye-tracking and EEG during reading and relevance decisions. *Journal of the Association for Information Science and Technology* 68, 10 (2017), 2299–2312.

[26] Sandra Hasko, Katarina Groth, Jennifer Bruder, Jürgen Bartling, and Gerd Schulte-Körne. 2014. What does the brain of children with developmental dyslexia tell us

about reading improvement? ERP evidence from an intervention study. *Frontiers in human neuroscience* 8 (2014), 441.

[27] Birger Hjørland. 2010. The foundation of the concept of relevance. *Journal of the american society for information science and technology* 61, 2 (2010), 217–237.

[28] Giulio Jacucci, Oswald Barral, Pedram Daee, Markus Wenzel, Baris Serim, Tuukka Ruotsalo, Patrik Pluchino, Jonathan Freeman, Luciano Gamberini, Samuel Kaski, et al. 2019. Integrating neurophysiologic relevance feedback in intent modeling for information retrieval. *Journal of the Association for Information Science and Technology* (2019).

[29] Jiepu Jiang, Daqing He, Diane Kelly, and James Allan. 2017. Understanding Ephemeral State of Relevance. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR '17)*. Association for Computing Machinery, New York, NY, USA, 137–146. https://doi.org/10.1145/3020165.3020176

[30] R Johnson Jr. 1988. The amplitude of the P300 component of the event-related potential: Review and synthesis. *Advances in psychophysiology* 3 (1988), 69–137.

[31] Ray Johnson Jr and Emanuel Donchin. 1978. On how P300 amplitude varies with the utility of the eliciting stimuli. *Electroencephalography and Clinical Neurophysiology* 44, 4 (1978), 424–437.

[32] Lauri Kangassalo, Michiel Spapé, Giulio Jacucci, and Tuukka Ruotsalo. 2019. Why Do Users Issue Good Queries? Neural Correlates of Term Specificity. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 375–384. https://doi.org/10.1145/3331184.3331243

[33] Jukka-Pekka Kauppi, Melih Kandemir, Veli-Matti Saarinen, Lotta Hirvenkari, Lauri Parkkonen, Arto Klami, Riitta Hari, and Samuel Kaski. 2015. Towards brain-activity-controlled information retrieval: Decoding image relevance from MEG signals. *NeuroImage* 112 (2015), 288–298.

[34] Jaana Kekäläinen. 2005. Binary and graded relevance in IR evaluations—comparison of the effects on ranking of IR systems. *Information processing & management* 41, 5 (2005), 1019–1033.

[35] Jaana Kekäläinen and Kalervo Järvelin. 2002. Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology* 53, 13 (2002), 1120–1129.

[36] Diane Kelly and Nicholas J Belkin. 2004. Display time as implicit feedback: understanding task effects. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 377–384.

[37] Hyun Hee Kim and Yong Ho Kim. 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 (2019), 931–941.

[38] Jürgen Koenemann and Nicholas J Belkin. 1996. A case for interaction: A study of interactive information retrieval behavior and effectiveness. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 205–212.

[39] Gina R Kuperberg. 2007. Neural mechanisms of language comprehension: Challenges to syntax. *Brain research* 1146 (2007), 23–49.

[40] Lukas Lerche and Dietmar Jannach. 2014. Using Graded Implicit Feedback for Bayesian Personalized Ranking. In *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys '14)*. Association for Computing Machinery, New York, NY, USA, 353–356. https://doi.org/10.1145/2645710.2645759

[41] Mark Levene, J Bar-Ilan, and M Zhitomirsky-Geffet. 2018. Categorical relevance judgment. *Journal of the Association for Information Science and Technology* (2018).

[42] Steven J Luck. 2014. *An introduction to the event-related potential technique*. MIT press.

[43] Steven J Luck and Emily S Kappenman. 2011. *The Oxford handbook of event-related potential components*. Oxford university press.

[44] Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. 2016. Why is that relevant? Collecting annotator rationales for relevance judgments. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*.

[45] Stefano Mizzaro. 1997. Relevance: The whole history. *Journal of the American society for information science* 48, 9 (1997), 810–832.

[46] Stefano Mizzaro. 1998. How many relevances in information retrieval? *Interacting with computers* 10, 3 (1998), 303–320.

[47] Yashar Moshfeghi and Joemon M Jose. 2013. An effective implicit relevance feedback technique using affective, physiological and behavioural features. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 133–142.

[48] Yashar Moshfeghi, Luisa R Pinto, Frank E Pollick, and Joemon M Jose. 2013. Understanding relevance: an fMRI study. In *European conference on information retrieval*. Springer, 14–25.

[49] Yashar Moshfeghi and Frank E Pollick. 2018. Search process as transitions between neural states. In *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 1683–1692.

[50] Yashar Moshfeghi, Peter Triantafillou, and Frank Pollick. 2019. Towards predicting a realisation of an information need based on brain signals. In *The World Wide Web Conference*. 1300–1309.

[51] Yashar Moshfeghi, Peter Triantafillou, and Frank E Pollick. 2016. Understanding information need: An fMRI study. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 335–344.

[52] Robert N Oddy, NJ Belkin, and Helen M Brooks. 1982. ASK for information retrieval: Part I. Background and theory. *Emerald: Journal of Documentation*. (1982), 61.

[53] John Polich. 2007. Updating P300: an integrative theory of P3a and P3b. *Clinical neurophysiology* 118, 10 (2007), 2128–2148.

[54] Gabriel Gaudencio Rêgo, Camila Campanhã, Julia Horta Tabosa do Egito, and Paulo Sérgio Boggio. 2017. Taking it easy when playing ultimatum game with a Down syndrome proposer: Effects on behavior and medial frontal negativity. *Social neuroscience* 12, 5 (2017), 530–540.

[55] Stephen E Robertson, Evangelos Kanoulas, and Emine Yilmaz. 2010. Extending average precision to graded relevance judgments. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 603–610.

[56] DS Ruchkin and S Sutton. 1978. Emitted P300 potentials and temporal unvertainty. *Electroencephalography and Clinical Neurophysiology* 45, 2 (1978), 268–277.

[57] Tefko Saracevic. 2007. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *Journal of the American Society for information Science and Technology* 58, 13 (2007), 2126–2144.

[58] Tefko Saracevic. 2016. The Notion of Relevance in Information Science: Everybody knows what relevance is. But, what is it really? *Synthesis Lectures on Information Concepts, Retrieval, and Services* 8, 3 (2016), i–109.

[59] Linda Schamber and Michael Eisenberg. 1988. Relevance: The Search for a Definition. (1988).

[60] Linda Schamber, Michael B Eisenberg, and Michael S Nilan. 1990. A re-examination of relevance: toward a dynamic, situational definition. *Information processing & management* 26, 6 (1990), 755–776.

[61] Tatiana Sitnikova, Dean F Salisbury, Gina Kuperberg, and Phillip J Holcomb. 2002. Electrophysiological insights into language processing in schizophrenia. *Psychophysiology* 39, 6 (2002), 851–860.

[62] Kenneth C Squires, Emanuel Donchin, Ronald I Herning, and Gregory McCarthy. 1977. On the influence of task relevance and stimulus probability on event-related-potential components. *Electroencephalography and clinical neurophysiology* 42, 1 (1977), 1–14.

[63] Scott C Steffensen, Allison J Ohran, Daniel N Shipp, Kimberly Hales, Sarah H Stobbs, and Donovan E Fleming. 2008. Gender-selective effects of the P300 and N400 components of the visual evoked potential. *Vision research* 48, 7 (2008), 917–925.

[64] Don T Stuss, TW Picton, and AM Cerri. 1986. Searching for the names of pictures: An event-related potential study. *Psychophysiology* 23, 2 (1986), 215–223.

[65] Markus Andreas Wenzel, Mihail Bogojeski, and Benjamin Blankertz. 2017. Real-time inference of word relevance from electroencephalogram and eye gaze. *Journal of neural engineering* 14, 5 (2017), 056007.

[66] Howard D White. 2017. Relevance theory and distributions of judgments in document retrieval. *Information Processing & Management* 53, 5 (2017), 1080–1102.

[67] Ryen W. White, Ian Ruthven, and Joemon M. Jose. 2002. The Use of Implicit Evidence for Relevance Feedback in Web Retrieval. In *Advances in Information Retrieval*, Fabio Crestani, Mark Girolami, and Cornelis Joost van Rijsbergen (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 93–109.

[68] Xin Xiao, Heather D Lucas, Ken A Paller, Jin-hong Ding, and Chun-yan Guo. 2014. Retrieval intention modulates the effects of directed forgetting instructions on recollection. *PloS one* 9, 8 (2014), e104701.

[69] Haopei Yang, Geoffrey Laforge, Bobby Stojanoski, Emily S Nichols, Ken McRae, and Stefan Köhler. 2019. Late positive complex in event-related potentials tracks memory signals when they are decision relevant. *Scientific reports* 9, 1 (2019), 1–15.

[70] Tengxiao Zhang, Chuting Bao, and Chunqu Xiao. 2019. Promoting effects of color-text congruence in banner advertising. *Color Research & Application* 44, 1 (2019), 125–131.

[71] Maayan Zhitomirsky-Geffet, Judit Bar-Ilan, and Mark Levene. 2015. How and why do users change their assessment of search results over time? *Proceedings of the Association for Information Science and Technology* 52, 1 (2015), 1–4.