Why do Users Issue Good Queries? Neural Correlates of Term Specificity

Lauri Kangassalo
University of Helsinki
lauri.kangassalo@helsinki.fi

Michiel Spapé
University of Helsinki
michiel.spape@helsinki.fi

Giulio Jacucci
University of Helsinki
giulio.jacucci@helsinki.fi

Tuukka Ruotsalo
Aalto University and University of Helsinki
tuukka.ruotsalo@aalto.fi

ABSTRACT

Despite advances in the past few decades in studying what kind of queries users input to search engines and how to suggest queries for the users, the fundamental question of what makes human cognition able to estimate goodness of query terms is largely unanswered. For example, a person searching information about "cats" is able to choose query terms, such as "housecat", "feline", or "animal" and avoid terms like "similar", "variety", and "distinguish". We investigated the association between the specificity of terms occurring in documents and human brain activity measured via electroencephalography (EEG). We analyzed the brain activity data of fifteen participants, recorded in response to reading terms from Wikipedia documents. Term specificity was shown to be associated with the amplitude of evoked brain responses. The results indicate that by being able to determine which terms carry maximal information about, and can best discriminate between, documents, people have the capability to enter good query terms. Moreover, our results suggest that the effective query term selection process, often observed in practical search behavior studies, has a neural basis. We believe our findings constitute an important step in revealing the cognitive processing behind query formulation and evaluating informativeness of language in general.

CCS CONCEPTS

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

KEYWORDS

Term specificity; Neural correlates; Human neurophysiology

ACM Reference Format:
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), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3331184.3331243

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331243

1 INTRODUCTION

Information retrieval systems have two primary components that determine their performance: queries representing information needs of users and the retrieval system computing the relevance and presenting information in response to the queries. Consequently, information retrieval research has mainly focused on improving the retrieval methods and computational support for creating queries. Behavioral evidence from laboratory and in-the-wild studies, on the other hand, have shown that users write and select queries that can be successfully used as input in search engines [19, 45, 46, 48]. Consequently, the user's ability to select query terms has become a commonly accepted assumption. But do we really know if this assumption holds?

While observed search behavior shows that users write effective queries [41, 44], and previous work shows that information needs and relevance can be traced to brain activity [13, 14, 17, 32], the underlying mechanism on why humans are able to select specific terms in their queries remains uncharted. For example, the terms "India", "Asia", or "Gandhi" are better at distinguishing the topic "India" from other topics than the terms "is", "vast", or "hot". But how are we able to make this distinction between important and less important terms? Previous work has shown behavioral evidence that this happens in practice [3, 43, 44, 46], but does not provide us with direct evidence of a neural origin of our ability to distinguish specific from non-specific terms. Are queries simply learned by trial-and-error when interacting with search engines, or are queries something that have a neural origin – are they based on some fundamental, cognitive functions that determine which terms are specific enough for an information need and which terms are less specific?

We set out to study whether term specificity is associated with brain activity and which neural correlates are distinguishable for specific and non-specific terms, as illustrated in Figure 1. More precisely, we define the following research question: Is the specificity of a term associated with the amplitude of its evoked brain activity?

A study was conducted in which EEG was recorded while participants read documents. The event related potential evoked by each word was analysed in order to measure the association between electrophysiological features and the information that the word carries in the document context.

The results show that term specificity is associated with amplified brain activity occurring between 200 to 800 ms following the presentation of the term. This suggests that the human capability to recognize terms that are specific for a particular document has

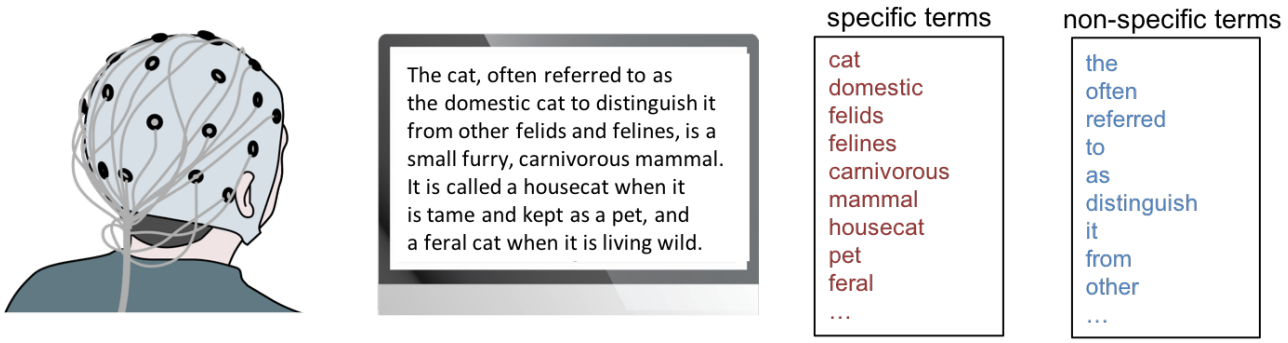

Figure 1: We show an association between term specificity and the amplitude of the evoked brain potentials measured via Electroencephalography (EEG) in response to reading terms from Wikipedia documents. Specific terms, such as 'cat', 'felids', or 'housecat' in the document about cats are associated with significantly different brain activity than non-specific terms, such as 'the', 'often', or 'distinguish'. © Tuukka Ruotsalo

a neural correlate. As participants had never seen the documents before, their reading should not have been influenced by a prior understanding of the document, or a particular search goal. Learning can therefore not account for the observed effects, suggesting a natural origin that generalizes to new terms and documents.

The results imply that query-term selection has a neural basis and that the human information processing system adjusts itself online in response to terms that carry information in a document context.

2 BACKGROUND

Our work is related to various, conventionally distinct, research areas: early information science work on term specificity and query formulation, cognitive neuroscience and psycholinguistics, and the utilization of neuroscience methods in information retrieval research.

2.1 Term specificity and query formulation

The early research on information retrieval has already revealed the importance of term specificity [21] and its relation to effectiveness of weighting schemes [39, 43]. Researchers have proposed a variety of measures for term specificity starting from the early work on tf-idf [21] to generative query likelihood of language models [39].

In addition to specificity of terms in queries, researchers have also studied the effect of the type of queries [46], query length [6], and query composition process [3, 42, 45] with respect to retrieval effectiveness. Methods and tools have been proposed to predict query performance [8] and assist the user in the query construction process [7, 24, 45]. However, in the end, the queries are only as good as the searchers' capability to formulate or recognize them. Consequently, the human mental process of recalling or recognizing the query terms that are specific for a particular document is crucial in the search process.

Previous research on selecting query terms, issuing query phrases, and studying their effectiveness is, however, based largely on behavioral evidence [3, 11, 41, 50]. That is, the studies have concentrated on analysing query logs. Consequently, we have little to no evidence on how humans are able to come up with the query terms in the first place. What are the cognitive processes that support query construction?

2.2 Cognitive neuroscience and psycholinguistics

Electroencephalography, the measurement of differences in electric potentials recorded from the scalp, has proven remarkably useful in the investigation of the temporal processing of external stimuli. Only if a large cluster of neurons fire synchronously in a similar direction, will their summed postsynaptic potential be strong enough to be measured from the scalp. The study of event-related potentials offers an elegant way around this conundrum: Knowledge of the exact timing of stimuli allows one to time-lock EEG data, so that by repeatedly presenting a similar stimulus, averaging the activity will gradually cancel out all random, unrelated activity. The event-related potential technique thus involves the analysis of electrophysiological activity that is evoked by events that have an onset defined to the millisecond. The temporal precision, in other words, allows us to attribute scalp electric potentials to brain activity.

The onset of differences between evoked responses are often seen as diagnostic to the depth of processing: early (<100 ms) differences tend to result from lower level differences in stimuli, such as in spatial location [20], while later (>250 ms) differences commonly involve more cognitive operations, such as context updating [10], and semantic integration [25].

While the specificity of search queries itself has received little attention in cognitive science or psycholinguistics, the study of event-related potential shows the remarkable degree to which evoked brain activity involves preferential processing of distinct, meaningful information. Indeed, two of the most studied event-related potentials in psychophysiology are the N2 and P3 families [27]. The N2 is found ca. 200-400 ms after presentation of a stimulus that deviates in some way from a repetitively occurring series of stimuli, such as a high tone after low tones, or a red circle after green circles [33]. If, additionally, such a stimulus is meaningful, whether that is defined by it being novel, improbable, informative, or task-relevant [10, 38, 47], it will evoke a successive positive

potential after ca. 300-700 ms, commonly called the P3. Together, these potentials that are intimately associated with specificity and informativeness therefore not only account for a significant part of the literature on psychophysiology, but also for much of the evoked brain response in general.

The study of psycholinguistics furthermore shows that semantically and syntactically distinct words evoke brain responses that are indirectly related to term specificity. In particular, Kutas and Hillyard [26] showed that unexpected deviants in a linguistic context will evoke a strong negativity 400 ms after the semantically ill-fitting word. Numerous studies (see [25], for a review) have since improved our understanding of the N400 component as a response to a semantic integration process. In contrast to the semantic negativity of the N400, syntactic oddities were found to be associated with a successive positivity at ca 600 ms, commonly called the P600 or syntactic positive shift [18]. Rather than involving syntactic analysis, however, it has been suggested that the grammatic oddity prompts an attempt to re-evaluate the preceding sentence in order to comprehend it [22]. This suggests that P600 can be related to deeper cognitive processing involved in comprehending text in general.

The majority of previous research treat stimuli (words, sentences, discourses etc.) as a manipulated variable, hand-crafted to produce a certain effect in the brain via purposely caused anomalies. While such an approach has provided valuable insights on the cognitive processes associated with language processing, the results may not generalise towards naturally occurring human language processing. In contrast to studies based on experimentally manipulated syntax and semantics, we investigate, to our knowledge for the first time, the effect of term specificity for human cognitive processing in a real-world document context.

2.3 Neuroscience methods in IR

Recent research has begun to examine information retrieval relevant research problems using methods of neuroscience. Researchers have utilized techniques of brain imaging to reveal the brain activity related to the underlying neural activity of participants engaged in tasks relevant to information retrieval context, such as detecting relevance or information need.

Neurophysiological correlates of relevance have been studied by Moshfeghi et al. [30] using functional Magnetic Resonance Imaging (fMRI) revealing three brain regions in the frontal, parietal and temporal cortex where brain activity differed between processing relevant and non-relevant documents. Moshfeghi and colleagues [31, 32] also reported an experiment revealing a distributed network of brain regions commonly associated with activities related to information need and retrieval, and differing brain activity in processing scenarios when participants knew the answer to a given question and when they did not and needed to search. This study showed that brain imaging techniques can provide us information about human cognitive processing even before it is manifested in information search activity. The temporal pattern of brain activity related to relevance assessment phenomena has also been studied [1, 13, 17]. In line with our results, studies consistently showed a variation within the first 800 ms of a relevance assessment process from the presentation of stimuli within the EEG signals.

Researchers have also aimed beyond studying when and where brain activity can be detected and engineered methods and systems that utilize neurophysiological measurements as input to search engines. Physiological signals have the advantage of implicitly eliciting relevance signals by exposing more items and collect relevance feedback without disrupting the user's search process. For example, Kauppi et al. [23] studied magnetoencephalographic signals alone and in conjunction with gaze signals in order to provide relevance feedback in an image retrieval task. Similarly, Eugster et al. [13] decoded the EEG with the objective of providing relevance feedback in a text retrieval task. Another study [15] demonstrated how the brain's relevance response can be harnessed to improve image searches in ambiguous search tasks.

Moreover, Eugster et al. [14] gave relevant feedback on words according to information extracted from EEG signals. The loop between brain and computer was closed by presenting new recommendations to the users according to the EEG-based feedback, which resulted in a significant information gain for about 70% of the participants of the study. That work constitutes, as far as we know, the first proof-of-concept IR systems that have performed automatic information filtering on the basis of brain activity alone. Similar work was recently reported where closed-loop system included a combination of gaze and EEG signals [49].

Related to information retrieval, mental processes and psychophysiological states measured from a combination of neural and peripheral sources have also been used to affectively annotate information [2, 5, 29]. In such scenarios, affective states can provide important context information for when or how to present information to user, considering awareness, relevance, and other mental states.

These studies have been important in providing evidence on the feasibility of including brain signal based relevance detection in real systems as well as information need and relevance correlates in the human neural system. However, the nuance of how neural correlates are associated with more detailed attributes of queries or documents, such as term specificity, have not, to our knowledge, been previously reported.

2.4 Contributions

Our contributions can be summarized as follows:

(1) To our knowledge, this is the first study showing that detecting term specificity has a neural basis.
(2) Our results suggests that the linguistic processing in the brain reflects the statistical measure of term specificity.
(3) We show that learning cannot sufficiently account for the observed effects. The presented text was new to the participants and effects were evoked immediately upon reading words.
(4) The results imply that term selection in query formulation has a neural basis, which underlies our natural ability to generally select discriminative query terms across documents and information needs.

3 TERM SPECIFICITY

Term specificity in information theory refers to the degree that terms accurately and precisely describe the content they refer to. We defined term specificity using Shannon entropy [52], which

constitutes an analytical measure that can be used to derive the algebraic as well as probabilistic interpretations of term specificity as commonly used in IR literature.

3.1 Defining term specificity

Let $D = (d_1, d_2, ..., d_N)$ be a set of mutually exclusive documents (events). Assume that for term (evidence) t, we can define a probability distribution $P_i = (p_1, p_2, ..., p_N)$ as follows $p, = P(d|t)$ for $i = 1, 2, ..., N$, such that $p_i \geq 0$ and $\sum_{i=0}^{N} p_i = 1$.

Now, we can define the term specificity for an individual term t_i in a document context d using Shannon entropy as follows:

$$H(d|t_i) = -\sum_{d} P(d|t_i) \cdot \log_2 P(d|t_i)$$

In this context, the Shannon entropy $H(d|T)$ can be interpreted as a measure of the degree of information uncertainty based on what we know about the document d evidence given as a set of terms T. Similarly, we can also compare the informativeness of terms. Consider a finite set of terms forming the vocabulary of the document collection, $T = t_1, t_2, ..., t_h$ and suppose $H(P_{t_j}) > H(P_{t_k}))$ for a pair of terms t_j, t_k. One may conclude that the term t_j is better than the term t_k, because t_j produces a state of lower uncertainty measured via entropy. Thus, the evaluation of a term based on the entropy function essentially amounts to the measurement of specificity of the term. Intuitively, the entropy also signifies the term's ability to distinguish documents from one another, as if $H(d|t_i)$ approaches the maximum, the distribution approaches uniformity.

Wong and Yao [52] have shown that commonly used term weighting schemes, such as tf-idf, can be derived from such definition. However, as entropy is directly interpretable in terms of conditional probabilities of terms and documents, we directly utilize entropy as our term-specificity measure.

3.2 Data and term-specificity estimation

In order to estimate term specificity in a general case, a large document collection representing broad topical and lexical variety is required.

We utilized the largest openly available encyclopedic document collection: the Wikipedia. Document models of 30 documents, shown in the left columns of Table 1, were generated to estimate the occurrences of terms in documents, as well as a corpus model consisting of all of Wikipedia's documents to estimate the occurrences of terms in the entire collection. Prior to constructing these models punctuation marks were removed from the text, and the words were stemmed using the Porter stemming algorithm [40].

To avoid zero probabilities, a smoothed term likelihood model was constructed based on the aforementioned document and collection models by using the linear interpolation model [53]. Formally, a smoothed estimate for a term given a document was computed as:

$$P_t(t|M_d) = (1 - \lambda)P(t|M_d) + \lambda P(t|M_{corpus}).$$

The specificity estimate and the likelihood in our case was computed for each term separately so the term length is one and $\lambda = 0.1$ was chosen as the smoothing parameter. That λ value was chosen as it has been shown to produce the best results on short text when using likelihood models in information retrieval [53]. Finally, the terms with an estimated specificity in the 25th percentile were labelled as specific terms, and terms with estimated specificity greater than the 25th percentile non-specific terms. This distinction was made for visualization and validation purposes. The effects between term specificity and brain potentials were further evaluated using a statistical model with continuous variables. A histogram of the occurrences of specificities can be seen in Figure 2. The theoretical maximum entropy for the set of 30 documents is $\log_2(30) \approx 3.401$, and it is reached when $P(d|t)$ is uniform.

3.3 Term-specificity estimate validation

The split of specific and non-specific terms using the 25th percentile of the entropy distribution was initially based on manual inspection of the data and an experiment in which independent assessors estimated the relevance of terms for documents. We asked three assessors to label each term in the 30 document collection for their relevance to the document in which the term appears.

The terms were labelled by the three assessors to be either relevant (specific) or irrelevant (non-specific). The annotating setting was single-blind, so the assessors were unaware of the entropies of the terms at the time of labelling. Approximately 25% of the words were labelled relevant, so in order to match the class sizes of the human annotated labels and specificity labels, the 25th percentile cutoff was found to be valid. The assessors annotations for relevance and the specificity labels were found to be in substantial agreement (Cohen's $\kappa = 0.612$).

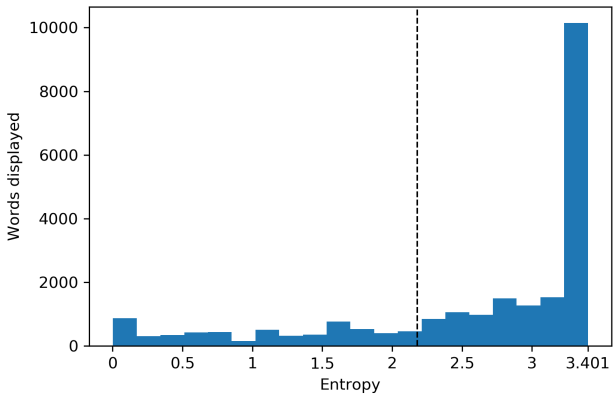

Figure 2: Occurences of entropies for all words presented to the participants, with the dashed line marking the 25th percentile of the entropies.

4 METHODOLOGY

An empirical study was conducted to investigate natural brain responses associated with terms occurring in documents. Typically, IR experiments are designed to measure the goodness of terms in retrieving some target documents that are relevant for a task. However, in order to study whether humans naturally react differently to specific than non-specific terms occurring in the target documents, we designed an empirical study that inverted the typical research

setting: the participants read natural language text from the target documents sampled from the English Wikipedia document corpus that was used to compute the term specificity estimates. EEG was then recorded in response to reading each individual term occurrence from the documents. This ensured that the recorded brain activity did not reflect previously acquired query formulation skills, but solely the natural responses to terms occurring in the documents.

4.1 Participants

Seventeen participants were recruited from Aalto University and the University of Helsinki. They were screened for health (no neuropathological history), handedness (right-handed), and English fluency (scoring high on the Cambridge English "Test your English - Adult Learners" online test[1]. The data of two participants were discarded due to technical issues. Of the fifteen remaining, 8 were female and 7 male, and their English fluency was assessed as high (Mean = 23.53, SD = 1.23; maximum value is 25). They were fully briefed as to the nature and purpose of the study prior to the experiment. Furthermore, and in accordance with the Declaration of Helsinki, they signed informed consents and were instructed on their rights as participants, including the right to withdraw from the experiment at any time without fear of negative consequences. They received two movie tickets as compensation for their participation.

4.2 Task

During the study, participants read 16 documents, randomly drawn from a pool of 30, while their EEG was measured. The thirty documents were drawn from the English Wikipedia (see left column of Table 1). All participants completed eight reading tasks, each of which consisted of reading two documents. Each document comprised six sentences (the first six sentences of the Wikipedia entry), which were read, one word at a time, across six sentences in alternating document order. In other words, in the first trial, the first sentence of the first document was read, followed by the first sentence of second document, in the second trial, the second sentence of each document, and so on. Participants were instructed to passively read what was displayed to them, and not to engage in any additional tasks or mental imagery.

Sentences were displayed on a computer screen, one word at a time, against a black rectangle with a grid-like pattern designed to minimise luminance differences between individual words. Following an initial warning signal ("Starting trial"), and a pattern-mask drawn from numerals and other non-literal characters (e.g. @@@@ in Figure 3), the first word from the first sentence was shown. Each single word was shown for 700 ms (SD = 0.3 ms), and replaced the preceding one without intermittent flash. After the last word of the first sentence, another pattern-mask was shown to indicate the separation between the previous and next document. The reading pace (of ca 1.43 Hz) was based on pilot studies suggesting this speed was fast enough for fluent reading, but not so fast as to cause interference between EEG potentials.

Following each reading task, we presented two validation tasks to ascertain that participants had not forgotten the topic, and had read both sentences. First, they were asked to indicate the name of one of the topics by typing it on the keyboard. Second, they were presented with a recall task, in which one of the sentences (chosen randomly from the two) was shown on the screen, with one of the nouns or verbs substituted by question marks. To indicate successful reading, participants were asked to fill out the missing word by typing it in. They were then presented with feedback regarding their performance on these two tasks in order to maintain concentration.

After this, the trial would be complete and the next two sentences would be shown following an intertrial interval of 1 s. Three 1 minute minimum breaks were provided to separate each quarter of the experiment, and ample possibility for further self-timed breaks was provided throughout. Completing the experiment took approximately 100 minutes, including instruction and psychophysiological preparation.

4.3 Apparatus and stimuli

Stimuli were words, presented in 18-point Lucida Console black typeface. They were displayed against a light-grey (RGB 82%, 82%, 82%) screen background, in the centre of a 300 x 100 pixel pattern mask. This mask showed a grid-like pattern, with a variable-width brighter window in the middle in which the words appeared. Words were each 1-15 characters in length, and punctuation was omitted from presentation. The sentences were clearly demarcated by word-like character repetitions of 4-9 numbers (e.g. 3333333) or other non-alphabetic characters (&&&&&&). These were designed to deliver similar low-level visual information as words, to avoid low-level mismatch effects, without provoking linguistic processing.

The LCD display used to present stimuli was positioned at approximately 60 cm from the participants, running at a refresh rate of 60 Hz, and a resolution of 1680 by 1050 pixels. Psychology Software Tools E-Prime Professional 2.0.10.353 stimulus presentation software was used to optimise timing of display and EEG amplifier trigger control. EEG was recorded from 32 Ag/AgCl electrodes, positioned on F7, F3, Fz, F4, F8, FT9, FC5, FC1, FC2, FC6, FT10, T7, C3, Cz, C4, T8, TP9, CP5, CP1, CP2, CP6, TP10, P7, P3, Pz, P4, P8, O1, Oz, and O2, using EasyCap elastic caps (EasyCap GmbH, Herrsching, Germany). Hardware amplification, filtering and digitisation was done via a QuickAmp USB (BrainProducts GmbH, Gilching, Germany) amplifier running at 2000 Hz. The electro-oculographam was also recorded, using two pairs of bipolar electrodes, situated 1 cm lateral to the outer canthi of the left and right eye, and 2 cm inferior and superior to the right pupil.

4.4 Pilot experiments

Preliminary versions of the final experimental procedure and design were piloted with four separate participants. In these experiments, we tested and evaluated, for example, the stimulus duration and the task. The data of these pilot experiments were not used in the final analysis.

4.5 EEG preprocessing

Electrophysiological data recorded using EEG commonly contain sources of noise related to eye blinks, head movements, power line noise, and so on. To improve the signal-to-noise ratio of the EEG

[1] https://www.cambridgeenglish.org/in/test-your-english/adult-learners/

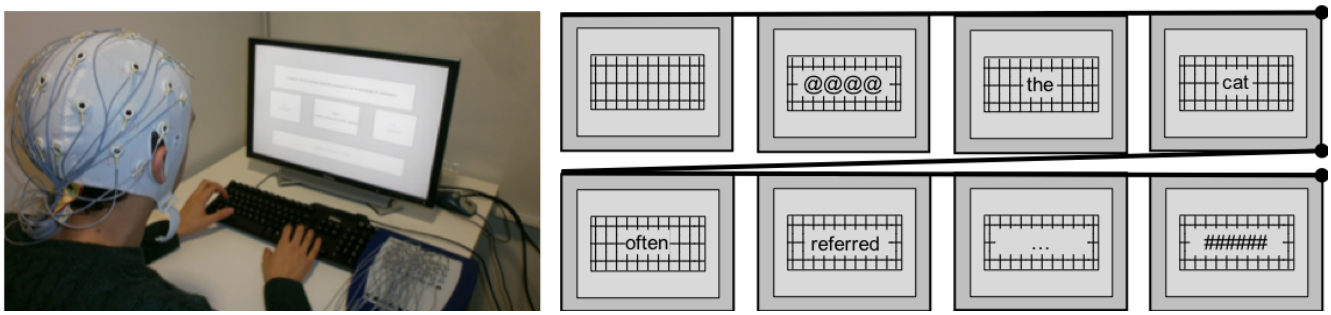

Figure 3: Experimental setup (left) showing the apparatus and EEG measurement setup, and step-by-step sequence (right) of a reading task showing farming, masks, sentence separator, and example terms presented one term at a time for 700ms per word in a sequence.

recordings, the EEG data was preprocessed according to standard procedures [28].

To remove low-frequency signal fluctuations and high-frequency line noise the data were band-pass filtered at the frequency range of 0.25 - 35 Hz with a Firwin1 filter. After this, the data were split to epochs spanning -200 - 1000 ms relative to the onset of each stimulus. Per-participant thresholds were computed to identify bad channels and epochs. For the computation of these thresholds, the epochs of the data were absolute baseline corrected using the average of the whole epoch -200 - 1000 ms, and the following subset of channels was picked: F3, Fz, F4, FC1, FC2, C3, Cz, C4, CP1, CP2, P3, Pz, P4. This subset of channels was chosen, because these channels reside near the top of the head, where less noise is present.

Using the aforementioned channels, a maximum absolute voltage was calculated for each epoch for the time interval -200 to 700 ms. The 80th percentile of the absolute max voltages was assigned as the voltage threshold. The threshold values ranged from 25 to 67 μV between participants. Epochs with an absolute maximum voltage over the threshold were marked as bad. In other words, 20% of the epochs with the highest absolute maximum values were deemed bad.

To find bad channels, the absolute maximum voltage was computed separately for each channel and epoch. Channels with an invalid epoch rate of over 20% of all epochs were marked as bad. Epochs were marked invalid if their absolute maximum voltage exceeded the voltage threshold or if their voltage variance was less than 0.5 μV.

Finally, the following modifications were made to the final data set, which at this point included all the channels and was not baseline corrected: the bad epochs were dropped, and the bad channels were interpolated using spherical splines [36], and each epoch was absolute baseline corrected using the pre-stimulus period -200 - 0 ms. After the preprocessing the average number of epochs per participant was 1550.

4.6 Statistical model

The experimental setup in the present study provides many factors which make the observations non-independent. Examples of the factors causing non-independencies are individual differences in observations between participants, the lengths of words that correlate with each other, and the documents displayed which vary between participants. Independence of observations is an assumption in the popular Analysis of Variance (ANOVA) models, and breaking this assumption may lead to overconfidence of the test (high Type I error rate). In order to avoid excessive Type I errors, Linear Mixed Models (LMMs) were used for modelling the dependence of brain activity and term specificity. These models allow for partial relaxation of the independence assumption by random effects. To avoid resulting to Type I errors, the LMM models were designed using the "keep it maximal"-principle presented by Barr et al. [4].

The mean voltage in the Pz channel was computed for each ERP component (components and their time windows specified at the beginning of the results section), and LMMs were fit for the data corresponding to each of the components. In other words, the brain activity was studied overall for the entire ERP as well as for the following components separately: P200, P300, N400, and P600.

Formally, the model was specified as follows:

$$Y_{pi} = (\beta_1 + P_{1p})X_i + \beta_2 Z_i + P_{0p} + L_i + D_i + \beta_0 + e_{pi},$$

and the Null model respectively as follows:

$$Y_{pi} = P_{1p}X_i + \beta_2 Z_i + P_{0p} + L_i + D_i + \beta_0 + e_{pi}.$$

Fixed effects in the models were term specificity (X_i) and document interest preference (Z_i), for term i. Their corresponding slopes were β_1 and β_2, respectively. The random intercepts were the participant ($P_{0p} \sim N(0, \tau^2)$ for participant p), the length of a term ($L_i \sim N(0, \chi^2)$), and the document from which the term was from ($D_i \sim N(0, \omega^2)$). Additionally, the model had a random by-participant slope for the effect of term specificity ($P_{1p} \sim N(0, \phi^2)$). Finally, β_0 is the overall intercept and $e_{pi} \sim N(0, \sigma^2)$ represents the general error term. The null model was the same as the alternative hypothesis model, except that the fixed effect of term specificity was omitted.

Document interest preference, as specified by the participant at the beginning of each reading task, was included to control the participants' potential experiment-independent pre-interest in one of the topics. The word length was added because a significant inverse correlation between term specificity and word lengths (Spearman's ρ -0.74, p < .0001) was observed. Although it is not clear whether the effect of word length can be separated from that of specificity; Longer words tend to carry more meaning than short words [37].

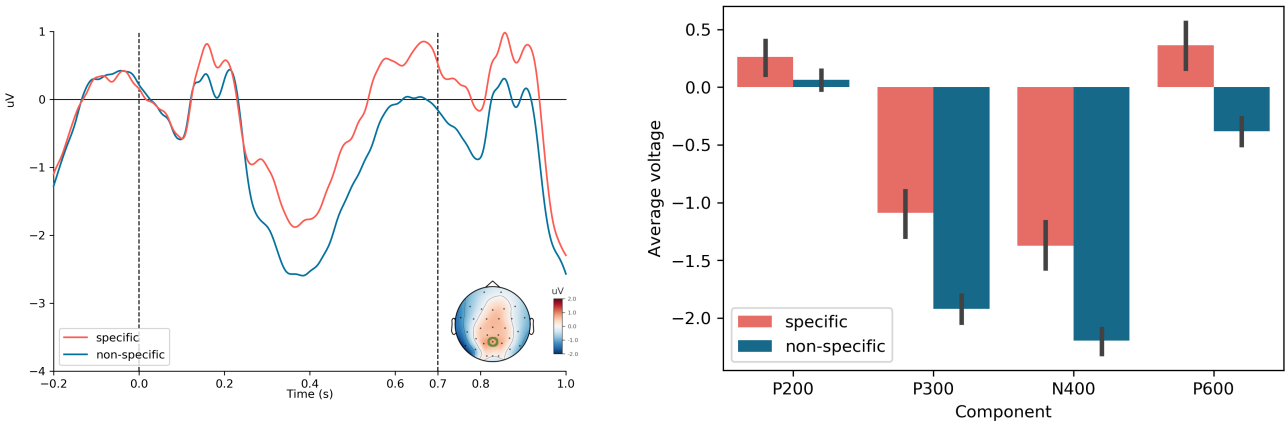

Figure 4: Term specificity is associated with amplified brain activity. Left: Grand average voltages for each ERP-component for the specific and non-specific stimuli terms at the Pz channel. The term specificity neural correlates are significantly different between 200ms to 800ms after the term has been presented. Right: Significant differences between the specific and non-specific terms are found for P200, P300, and P600 components.

However, in the model we wanted to control for this effect as it could have been possible that longer words could have caused more cognitive processing independent from the term specificity. By-participant random intercepts were included to control for the participant-wise variance of the base-level of the brain responses. Additionally, by-participant random slopes for the effect of term specificity were included, to control for the possible individual variations in brain responses to term specificity. Visual inspection of residual plots did not reveal obvious deviations from normality or homoscedasticity.

The models were evaluated with likelihood ratio tests between the alternative hypothesis and the null hypothesis model.

5 RESULTS

Three types of results are reported: the overall effect of term specificity on brain activity, the effect of term specificity on particular ERP components, and terms exemplifying the terms with low and high specificity.

5.1 Overall term specificity effect

Term specificity was found to have a significant effect on brain activity associated with reading ($df = 8$, $\chi^2 = 12.22$, Bonferroni corrected $p < .01$).

The greatest difference between the two classes was found in the Pz electrode, and it was therefore chosen for further inspection. Figure 4 (left) displays the grand average voltages for specific and non-specific terms at the Pz electrode. The averaged difference in voltages between the two term classes for the 200-800 ms post-stimulus time range is displayed as a scalp topography in the lower right corner of the plot. The location of the Pz electrode is highlighted with a green circle in the scalp topography. Specific terms are associated with a long-lasting positivity starting at roughly 250ms and lasting until 800ms.

Our results are also intuitive. The ERPs are amplified for example on terms, such as *nucleus, atomic, atom,* and *neutrons* for the document *atom*; *democrat, student, clinton,* and *president* for the document *bill clinton*; and *supply, coins,* and *currency* for the document *money*. These terms are specific and match the intuition of appropriated query terms for the corresponding topics. Conversely, the ERPs are declined for example on terms, such as *used, equal, address, into, nearly, consist,* and *of* which are intuitively non-specific and would not count for good query terms independently of the target document or topic.

5.2 Term specificity effect on ERP components

Figure 4 (right) displays the average voltages for ERP-components. The ERP-components were defined with time-windows post-stimuli: P200 [100, 250] ms, P300 [250, 350] ms, N400 [350, 500] ms, and P600 [500, 800] ms. These time intervals were chosen based on visual inspection of the ERPs and correspond to the P200, P300, N400, and P600 components that have been previously shown to be associated with language processing.

P200. The difference in voltages for terms in the specific and non-specific classes was found to be significant in terms of the P200 component ($df = 8$, $\chi^2 = 6.74$, Bonferroni corrected $p < .05$). The P200 component is affected by the physical features of the stimuli [35], so the difference in the amplitudes for the two term specificity classes may be explained by the effects that word length has on the visual characteristics of the stimuli.

P300. The difference in voltages for terms in the specific and non-specific classes was found to be significant in terms of the P300 component ($df = 8$, $\chi^2 = 6.71$, Bonferroni corrected $p < .05$). The P300 is sensitive to a person's response to a stimuli, rather than its physical attributes. More specifically, the P300 is thought to reflect processes involved in stimulus evaluation or categorization.

document	high specificity	low specificity
atom	microscope, nucleus, atomic, atom, neutrons	used, equal, observed, containing, of
automobile	automobile, benz, car, passengers, rail	rather, an, was, 500, a
bank	monte, bank, liabilities, dei, liquid	due, and, nations, had, of
bicycle	society, frame, automobiles, 1885, safety	more, 19th, and, the, had
bill clinton	democrat, student, clinton, president, peacetime	foundation, address, united, states, into
brain	animals, comparison, brain, synapses, neurons	a, on, complex, other, the
cat	indoor, cats, quick, teeth, bred	includes, of, known, when, species
communism	interpretations, characterised, recycling, marxism, communist	first, 20th, developed, theory, understanding
euro	banking, debt, reserve, banknotes, dollar	of, for, official, making, value
football	feet, soccer, football, ball, besides	then, may, commonly, association, only
india	india, vast, society, struggle, wildlife	diversity, identified, valley, became, home
learning	awareness, animals, skills, consciously, learn	of, species, personal, be, result
machine learning	statistics, data, computer, artificial, vision	and, applications, in, rule, example
michael jackson	solo, complicated, music, michael, philanthropist	is, professional, made, for, through
money	supply, coins, currency, tender, account	nearly, consists, record, country, unit
ocean	oceanographers, 97, earth, unknown, climate	world, because, component, believed, of
painting	sponges, brush, pigment, paint, painting	used, of, numerous, dominated, and
plato	teach, plato, abstract, socratic, ethics	a, his, along, work, of
politics	civic, society, discourse, exercising, political	control, people, given, particularly, wide
rome	michelangelo, peninsula, vatican, lazio, sistine	is, cities, famous, country, example
savanna	rainfall, hemisphere, savannah, savannas, water	regularly, by, season, than, the
schizophrenia	medication, receptor, disorder, triggered, abnormal	thinking, are, which, single, appear
school	teachers, students, learning, formal, seminary	countries, most, institution, economics, be
society	society, insofar, artificial, societies, otherwise	in, ways, extensively, or, organism
star	luminous, nearest, star, energy, plasma	white, to, determine, most, space
telephone	portable, user, device, distant, keypad	enabled, talk, with, radio, such
time	durations, astronomy, arts, intervals, events	or, science, major, felt, from
volcano	troposphere, vicinity, cool, gases, volcanoes	affect, a, not, found, escape
wife	heterosexual, spouse, obligations, female, marital	cultures, is, applied, may, relation
wine	acids, fermented, nutrients, wine, fruits	closely, history, with, discovered, played

Table 1: Random five terms sampled from the high specificity and low specificity classes for each document used in the experiment. The sampled terms illustrate that the brain activity is significantly amplified for terms that can be considered descriptive and good query terms for the documents.

N400. The difference in voltages for terms in the specific and non-specific classes was not found to be significant in terms of the N400 component ($df = 8$, $\chi^2 = 4.89$, Bonferroni corrected $p > .2$). While we did not observe a significant difference in the voltages between the low and high term specificity classes, it is notable that the difference between the two classes is visually present. The difference, however, does not grow after the P300 component. Thus, the N400 component seems to be unaffected by the term specificity.

P600. The difference in voltages for terms in the specific and non-specific classes was found to be significant in terms of the P600 component ($df = 8$, $\chi^2 = 21.757$, Bonferroni corrected $p < .0001$). P600 is a language-relevant ERP and is thought to be elicited by various grammatical and syntactic processing tasks. On the other hand, it has been suggested that P600 is associated with the cognitive effort that it takes to interpret language. Our results support the latter and specifically suggest that P600 is associated with the amount of expected information carried by a term.

6 CONCLUSIONS

We set out to study the association between term specificity and human brain activity in a scenario involving real data from a large document corpora. We asked the following research question: Is the specificity of a term associated with the amplitude of its evoked brain activity? Our results show that term specificity has a neural basis and specificity of a term is associated with its evoked brain activity. In particular, term specificity is associated with human brain activity in ERP-components P200, P300, and P600, which we show to differ significantly between specific and non-specific terms.

6.1 Empirical findings

The results suggest that cognitive processing of terms is associated with term specificity, and that humans seem to have capability to naturally recognize terms that are specific and discriminative to a particular document.

While this ability may reflect a behavior that is learned as a function of performing successful and unsuccessful searches – i.e. a result from learning good terms for a particular information need – one would have expected to observe cognitive equanimity. Conversely, our experiments revealed amplified potentials at first exposure already. This suggests that the ability is not learned, but a natural function of human cognition.

The present study was based on natural reading rather than experimentally constructed linguistic stimuli, and we therefore cannot with certainty distinguish which potentials precisely are associated with term specificity. However, significant effects with

word specificity and the P300 and P600 ERP-components are in line with the context-updating theory [9] of the P300 and the P600-as-P300 theory: Informative words within a sentence are likely to impact the context, the representation of the text, more than nonspecific terms. The observed correlation on the P200 component is more unexpected, as the early onset precludes the effect from being explained by the word's semantic content, as semantic processing is commonly theorised to occur later [25]. We surmise therefore that it may derive from physical differences related to term specificity, for example due to specific words generally having a longer word-length, which affects the P200 [35]. Alternatively, it may also be a consequence of certain top-down processes related to preceding context and expecation, affecting memory [12] and attention [34]. Similarly, the P200 is sensitive to repetition suppression [16], a reduction in neural activity when a stimulus is repeated. This could explain the lower amplitude of the P200 component for non-specific terms, as they naturally occur more frequently across documents.

6.2 Limitations

The EEG preprocessing was conducted in accordance with data preprocessing standards for BCI systems [51]. We also used natural stimuli from a general encyclopedic source and constructed the reading experiment to ensure that any intentionality in selecting stimuli text or individual terms would not affect the experiment. Participants' pre-existing knowledge may nevertheless have affected the results. However, this is unlikely to account for the present results, as topics were of a diverse nature.

Moreover, we chose to use linear mixed models to avoid statistical methods the assumptions of which the data did not fulfil. However, including all of the factors possibly affecting the results as random effects in an LMM is infeasible for two reasons: not all of such factors may be determined, and, on a more technical note, an LMM is known to have convergence problems caused by the model running out of degrees of freedom due to constraints imposed on the data by fixed and random effects. For the aforementioned reasons, we picked the factors that we deemed caused most of the variance in the measurements as effects in the LMMs.

While the pre-processing, the recruitment of participants, the source and language of stimuli text, and the decisions made in statistical modelling were carefully defined, a possibility of bias cannot be fully excluded. For example, all of the factors causing non-independencies in the measurements could not be included as random effects in the LMMs. This is mainly due to the fact that when studying natural language, there is little control over the stimuli presented to participants. Factors such as sentence length or word context cannot be fully controlled without compromising the natural document context assumption.

6.3 Implications for information retrieval

In essence, decisions on query formulation and term selection happen in the human brain, but so far, our understanding as to whether the query formulation is simply learned by trial-and-error when interacting with search engines, or whether it is independent of previous searching experience on the topic, has been limited. Our work shows evidence that term selection process is based on fundamental cognitive abilities that determine which terms are specific enough for an information need and which terms are less specific.

To our knowledge, this is the first study showing that adult human ability to detect specific terms has neural origins. As participants had never seen the presented documents before, they should not have been influenced by a prior understanding of the documents, nor a particular search goal. Learning term specificity specifically for these documents cannot therefore account for the observed effects, suggesting a natural origin that generalizes to new terms and documents.

The results have a fundamental implication for research in information science and retrieval: the assumption that adult people can recognize and issue queries that are discriminative has neural basis and this ability naturally generalizes to new documents and information needs.

ACKNOWLEDGMENTS

The research was supported by the Academy of Finland under the project NEUROADAPT: Neuro-adaptive Intention Learning (Decision No. 313610).

REFERENCES

[1] 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)*. ACM, New York, NY, USA, 719–722. https://doi.org/10.1145/2766462.2767811

[2] 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)*. ACM, New York, NY, USA, 395–402. https://doi.org/10.1145/1390334.1390403

[3] Anne Aula, Rehan M. Khan, and Zhiwei Guan. 2010. How Does Search Behavior Change As Search Becomes More Difficult?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 35–44. https://doi.org/10.1145/1753326.1753333

[4] Dale J Barr, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language* 68, 3 (2013), 255–278.

[5] Oswald Barral, Ilkka Kosunen, Tuukka Ruotsalo, Michiel M. Spapé, Manuel J. A. Eugster, Niklas Ravaja, Samuel Kaski, and Giulio Jacucci. 2016. Extracting relevance and affect information from physiological text annotation. *User Modeling and User-Adapted Interaction* 26, 5 (2016), 493–520. https://doi.org/10.1007/s11257-016-9184-8

[6] N. J. Belkin, D. Kelly, G. Kim, J.-Y. Kim, H.-J. Lee, G. Muresan, M.-C. Tang, X.-J. Yuan, and C. Cool. 2003. Query Length in Interactive Information Retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval (SIGIR '03)*. ACM, New York, NY, USA, 205–212. https://doi.org/10.1145/860435.860474

[7] Huanhuan Cao, Daxin Jiang, Jian Pei, Qi He, Zhen Liao, Enhong Chen, and Hang Li. 2008. Context-aware Query Suggestion by Mining Click-through and Session Data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08)*. ACM, New York, NY, USA, 875–883. https://doi.org/10.1145/1401890.1401995

[8] Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. 2002. Predicting Query Performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '02)*. ACM, New York, NY, USA, 299–306. https://doi.org/10.1145/564376.564429

[9] Emanuel Donchin and Michael Coles. 1988. Is the P300 component a manifestation of context updating? *Behavioral and Brain Sciences - BEHAV BRAIN SCI* 11 (09 1988), 357–427. https://doi.org/10.1017/S0140525X00058027

[10] Emanuel Donchin and Michael GH Coles. 1998. Context updating and the P300. *Behavioral and brain sciences* 21, 1 (1998), 152–154.

[11] Doug Downey, Susan Dumais, Dan Liebling, and Eric Horvitz. 2008. Understanding the Relationship Between Searchers' Queries and Information Goals. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM '08)*. ACM, New York, NY, USA, 449–458. https://doi.org/10.1145/1458082.1458143

[12] Bruce R Dunn, Denise A Dunn, Marlin Languis, and David Andrews. 1998. The Relation of ERP Components to Complex Memory Processing. *Brain and Cognition* 36, 3 (1998), 355 – 376. https://doi.org/10.1006/brcg.1998.0998

[13] 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)*. ACM, New York, NY, USA, 425–434. https://doi.org/10.1145/2600428.2609594

[14] Manuel J A 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, 1 (2016), 38580. https://doi.org/10.1038/srep38580

[15] Jan-Eike Golenia, Markus A. Wenzel, Mihail Bogojeski, and Benjamin Blankertz. 2017. Implicit relevance feedback from electroencephalography and eye tracking in image search. *Journal of Neural Engineering* (2017). https://doi.org/10.1088/1741-2552/aa9999 accepted; open access.

[16] E.J Golob, H Pratt, and A Starr. 2002. Preparatory slow potentials and event-related potentials in an auditory cued attention task. *Clinical Neurophysiology* 113, 10 (2002), 1544 – 1557. https://doi.org/10.1016/S1388-2457(02)00220-1

[17] 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. https://doi.org/10.1002/asi.23904

[18] Peter Hagoort, Colin Brown, and Jolanda Groothusen. 1993. The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and cognitive processes* 8, 4 (1993), 439–483.

[19] Bernard J. Jansen, Amanda Spink, and Tefko Saracevic. 2000. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing & Management* 36, 2 (2000), 207 – 227. https://doi.org/10.1016/S0306-4573(99)00056-4

[20] D. A. Jeffreys and J. G. Axford. 1972. Source locations of pattern-specific components of human visual evoked potentials. I. Component of striate cortical origin. *Experimental brain research* 16, 1 (1972), 1–21.

[21] Karen Spark Jones. 1972. A statistical interpretation of term specificity and its application in information retrieval. *Journal of Documentation* 28, 1 (1972), 11–21. https://doi.org/10.1108/eb026526

[22] Edith Kaan and Tamara Y. Swaab. 2003. Repair, revision, and complexity in syntactic analysis: An electrophysiological differentiation. *Journal of cognitive neuroscience* 15, 1 (2003), 98–110.

[23] 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 (may 2015), 288–98. https://doi.org/10.1016/j.neuroimage.2014.12.079

[24] Diane Kelly, Karl Gyllstrom, and Earl W. Bailey. 2009. A Comparison of Query and Term Suggestion Features for Interactive Searching. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09)*. ACM, New York, NY, USA, 371–378. https://doi.org/10.1145/1571941.1572006

[25] Marta Kutas and Kara D. Federmeier. 2011. Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP). *Annual review of psychology* 62 (2011), 621. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4052444/

[26] Marta Kutas and Steven A. Hillyard. 1980. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science* 207, 4427 (1980), 203–205.

[27] S. J. Luck. 2005. *An introduction to the event-related potential technique*. MIT Press, Cambridge, MA.

[28] Steven J. Luck. 2014. Artifact Rejection and Correction. In *An introduction to the event-related potential technique* (2nd ed.). MIT Press, Cambridge, Massachusetts, 185–217.

[29] 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 (SIGIR '13)*. ACM, New York, NY, USA, 133–142. https://doi.org/10.1145/2484028.2484074

[30] Yashar Moshfeghi, Luisa R. Pinto, Frank E. Pollick, and Joemon M. Jose. 2013. Understanding Relevance: An fMRI Study. In *ECIR (Lecture Notes in Computer Science)*, Vol. 7814. Springer, 14–25.

[31] Yashar Moshfeghi and Frank E. Pollick. 2018. Search Process As Transitions Between Neural States. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1683–1692. https://doi.org/10.1145/3178876.3186080

[32] 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 (SIGIR '16)*. ACM, New York, NY, USA, 335–344. https://doi.org/10.1145/2911451.2911534

[33] R Näätänen and Terence Picton. 1986. N2 and automatic versus controlled processes. *Electroencephalography and clinical neurophysiology. Supplement* 38 (02 1986), 169–86.

[34] Helen J. Neville and Donald Lawson. 1987. Attention to central and peripheral visual space in a movement detection task: an event-related potential and behavioral study. I. Normal hearing adults. *Brain Research* 405, 2 (1987), 253 – 267. https://doi.org/10.1016/0006-8993(87)90295-2

[35] Shu Omoto, Yoshiyuki Kuroiwa, Saika Otsuka, Yasuhisa Baba, Chuanwei Wang, Mei Li, Nobuhisa Mizuki, Naohisa Ueda, Shigeru Koyano, and Yume Suzuki. 2010. P1 and P2 components of human visual evoked potentials are modulated by depth perception of 3-dimensional images. *Clinical Neurophysiology* 121, 3 (2010), 386 – 391. https://doi.org/10.1016/j.clinph.2009.12.005

[36] F. Perrin, J. Pernier, O. Bertrand, and J.F. Echallier. 1989. Spherical splines for scalp potential and current density mapping. *Electroencephalography and Clinical Neurophysiology* 72, 2 (1989), 184 – 187. https://doi.org/10.1016/0013-4694(89)90180-6

[37] Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences* 108, 9 (2011), 3526–3529. https://doi.org/10.1073/pnas.1012551108

[38] John Polich. 2007. Updating P300: an integrative theory of P3a and P3b. *Clinical neurophysiology* 118, 10 (2007), 2128–2148. https://doi.org/10.1016/j.clinph.2007.04.019

[39] Jay M. Ponte and W. Bruce Croft. 1998. 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)*. ACM, New York, NY, USA, 275–281. https://doi.org/10.1145/290941.291008

[40] Martin F Porter. 1980. An algorithm for suffix stripping. *Program* 14, 3 (1980), 130–137.

[41] Daniel E. Rose and Danny Levinson. 2004. Understanding User Goals in Web Search. In *Proceedings of the 13th International Conference on World Wide Web (WWW '04)*. ACM, New York, NY, USA, 13–19. https://doi.org/10.1145/988672.988675

[42] Tuukka Ruotsalo, Jaakko Peltonen, Manuel J. A. Eugster, Dorota Głowacka, Patrik Floréen, Petri Myllymäki, Giulio Jacucci, and Samuel Kaski. 2018. Interactive Intent Modeling for Exploratory Search. *ACM Trans. Inf. Syst.* 36, 4, Article 44 (Oct. 2018), 46 pages. https://doi.org/10.1145/3231593

[43] Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 24, 5 (1988), 513 – 523. https://doi.org/10.1016/0306-4573(88)90021-0

[44] Craig Silverstein, Hannes Marais, Monika Henzinger, and Michael Moricz. 1999. Analysis of a Very Large Web Search Engine Query Log. *SIGIR Forum* 33, 1 (Sept. 1999), 6–12. https://doi.org/10.1145/331403.331405

[45] Amanda Spink and Tefko Saracevic. 1997. Interaction in information retrieval: Selection and effectiveness of search terms. *Journal of the American Society for Information Science* 48, 8 (1997), 741–761.

[46] Amanda Spink, Dietmar Wolfram, Major B. J. Jansen, and Tefko Saracevic. 2001. Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology* 52, 3 (2001), 226–234. https://doi.org/10.1002/1097-4571(2000)9999:9999<::AID-ASI1591>3.0.CO;2-R

[47] Samuel Sutton, Patricia Tueting, Joseph Zubin, and E. Roy John. 1967. Information delivery and the sensory evoked potential. *Science* 155, 3768 (1967), 1436–1439. http://www.sciencemag.org/content/155/3768/1436.short

[48] Tung Vuong, Miamaria Saastamoinen, Giulio Jacucci, and Tuukka Ruotsalo. [n. d.]. Understanding user behavior in naturalistic information search tasks. *Journal of the Association for Information Science and Technology* ([n. d.]). https://doi.org/10.1002/asi.24201

[49] M A Wenzel, M Bogojeski, and B Blankertz. 2017. Real-time inference of word relevance from electroencephalogram and eye gaze. *Journal of Neural Engineering* 14, 5 (oct 2017), 056007. https://doi.org/10.1088/1741-2552/aa7590

[50] Ryen W. White, Matthew Richardson, and Wen-tau Yih. 2015. Questions vs. Queries in Informational Search Tasks. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion)*. ACM, New York, NY, USA, 135–136. https://doi.org/10.1145/2740908.2742769

[51] Jonathan Wolpaw and Elizabeth Winter Wolpaw. 2012. *Brain-computer interfaces: principles and practice*. OUP USA.

[52] S. K. M. Wong and Y. Y. Yao. 1992. An information-theoretic measure of term specificity. *Journal of the American Society for Information Science* 43, 1 (1992), 54–61.

[53] Chengxiang Zhai and John Lafferty. 2017. A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. *SIGIR Forum* 51, 2 (Aug. 2017), 268–276. https://doi.org/10.1145/3130348.3130377