BrainCoDe: Electroencephalography-based Comprehension Detection during Reading and Listening

Christina Schneegass[1], **Thomas Kosch**[1], **Andrea Baumann**[1],
Marius Rusu[1], **Mariam Hassib**[2], **Heinrich Hussmann**[1]
[1]LMU Munich, Munich, Germany, {firstname.lastname}@ifi.lmu.de
[2]Bundeswehr University Munich, Munich, Germany, mariam.hassib@unibw.de

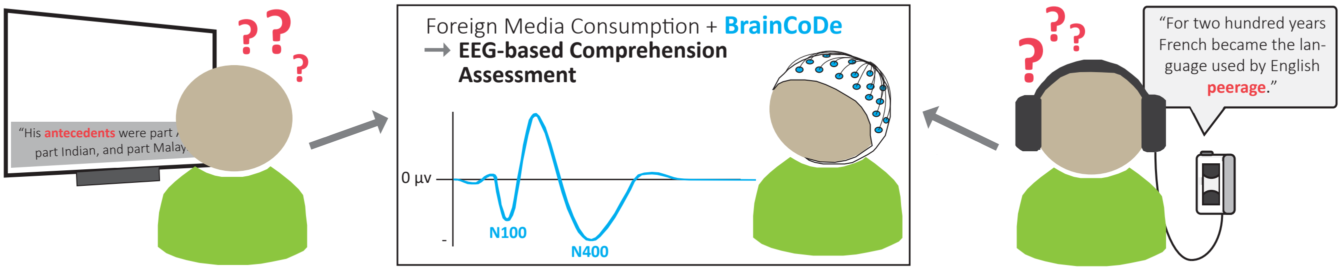

"His antecedents were part
part Indian, and part Malay
Foreign Media Consumption + BrainCoDe
→ EEG-based Comprehension Assessment
0 µv
-
N100
N400
"For two hundred years French became the language used by English peerage."

Figure 1. Our *BrainCoDe* approach utilizes Electroencephalography to assess users' second-language vocabulary comprehension during text reading on screen and listening to narrated audio content. In the future, our approach may be applied for comprehension analysis during subtitle reading in movies (left) as well as for audiobooks (right).

ABSTRACT

The pervasive availability of media in foreign languages is a rich resource for language learning. However, learners are forced to interrupt media consumption whenever comprehension problems occur. We present *BrainCoDe*, a method to implicitly detect vocabulary gaps through the evaluation of event-related potentials (ERPs). In a user study (N=16), we evaluate *BrainCoDe* by investigating differences in ERP amplitudes during listening and reading of *known* words compared to *unknown* words. We found significant deviations in N400 amplitudes during reading and in N100 amplitudes during listening when encountering *unknown* words. To evaluate the feasibility of ERPs for real-time applications, we trained a classifier that detects vocabulary gaps with an accuracy of 87.13% for reading and 82.64% for listening, identifying eight out of ten words correctly as *known* or *unknown*. We show the potential of *BrainCoDe* to support media learning through instant translations or by generating personalized learning content.

Author Keywords

EEG; Implicit Comprehension Detection; Language Learning

CCS Concepts

•**Human-centered computing** → **Empirical studies in HCI;** *Interaction techniques;*

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376707

INTRODUCTION

With the rise of the internet, the availability of media content in a variety of languages has increased tremendously. This content is often used inside and outside the classroom to learn and improve second-language skills [14, 45]. Users can choose movies or audiobooks along with their interests and proficiency level, making media content an efficient tool for second-language learning [18]. Encountering *unknown* vocabulary during media consumption tempts users to look up the respective translation, hence ensuring the overall comprehension of the content. However, looking up translations while watching movies or listening to audiobooks causes interruptions and decreases comprehension, leading to a negative user experience [51]. To support interruption-free language learning during listening and reading, we propose to implicitly assess the gaps in learners' vocabulary knowledge without requesting active user intervention. This implicit assessment is a first step towards the implementation of proficiency-aware interfaces, which could offer real-time support for foreign language learning during media consumption in diverse contexts.

Our approach utilizes Electroencephalography (EEG) to assess users' vocabulary comprehension. EEG has been used to evaluate language processing (cf. [25, 26]) and proved its potential as implicit input for HCI applications (cf. [17, 48]). In the last decade, EEG has become increasingly robust and easier to handle with the availability of prototypes embedded in caps or glasses to enable evaluation in real-world scenarios [6, 9, 52].

Our contribution is *BrainCoDe*, an EEG-based approach for second-language comprehension detection. *BrainCoDe* classifies the users' neural responses to detect *known* and *unknown* vocabulary during English reading as well as listening. Prior work utilized EEG to differentiate between meaningful and pseudo-words [57] and presented text as individual words on the screen (rapid serial visualization presentation) [47], while *BrainCoDe* enables the evaluation of full-sentence text presentation. In comparison to prior work, this paper assesses language comprehension during reading and listening and compares the accuracy of *BrainCoDe* for both modalities.

We present the results of a user study (N=16), showing that we can detect vocabulary gaps during second-language reading with 87.13% accuracy and during listening with 82.64% accuracy, respectively. For our analysis, we only use one electrode centrally located on the scalp, highlighting *BrainCoDe's* potential for replication with consumer EEG devices, such as NeuroSky MindWave[1] or a single electrode attached to regular headphones. We discuss *BrainCoDe's* applicability to provide real-time or post hoc feedback in real-world scenarios, for example, by extracting *unknown* vocabulary to create a personalized vocabulary list or recommending content according to users' language proficiency.

RELATED WORK AND BACKGROUND

Language Learning

The prevalence of the internet gave rise to a steady growth of media content available for everyone in a variety of languages. Especially for studying English, movies, TV series, or audiobooks which are available at streaming services such as Netflix[2] or Amazon[3] are a common tool to improve one's language skills. By changing the audio track of a movie and enabling subtitles, media content can support effective learning [18, 56].

Besides being a convenient tool for learning, which is accessible anytime and anywhere, media content also represents the user's interest and hence can increase learning motivation [36]. This is in contrast to the concept of language learning classes, which predetermine learners' schedules and learning content. Media content ensures a high degree of language exposure and can provide interactivity as in pausing and rewinding certain scenes [44]. This interactivity is, in particular, necessary when learners encounter vocabulary they do not understand. However, requesting translations during reading or listening, even within the application, interrupts the task and leads to "media multitasking". When engaging with more than one medium at once, the effort of multitasking can lead to a decreased recall of the presented content and a worse understanding due to higher cognitive load [51]. Thus, it is likely that *unknown* words are skipped to continue watching the movie or listening to the audiobook, trying to ensure the overall text comprehension.

If we want to support learning of new vocabulary with media content, it is necessary to implicitly assess a person's knowledge gaps [35] without active user intervention. We can assess those knowledge gaps and use them to provide effective learning support by monitoring a user's understanding while engaging with second-language content. Through adaptations in the interface (e.g., lowering the speed of a speaker in an audiobook) we could provide technical support to facilitate comprehension and learning. Moreover, by evaluating a person's comprehension, we can generate personalized learning content for additional post hoc repetition, targeting exactly those vocabulary the user is struggling with.

In the last two decades, the implicit assessment of comprehension by the use of physiological sensing became increasingly researched [3, 4, 12]. For the estimation of comprehension during reading, eye tracking can give insights on people's understanding. In the context of HCI, eye-gaze analysis has been previously evaluated to assess a learner's language proficiency (cf. [1, 3, 46]). Although eye-gaze analysis already presents a feasible approach for language proficiency assessment [22], this method is limited to visual content presentation. Hence, the evaluation of comprehension during the perception of audio content is not possible. A mechanism that gained popularity in the last decade and has the potential to be applied for implicit assessment of comprehension across multiple modalities is EEG.

Electroencephalography in HCI

By measuring electric potentials through electrodes on the scalp, EEG can give insights on a plethora of users' internal processes, such as engagement, workload, attention, fatigue, emotions, flow, or immersion [2, 11]. The evaluation of EEG signals can be performed based on frequency bands or Event-Related Potentials (ERPs) [31]. The latter refer to changes in signal amplitudes occurring at a precise and consistent time after the presentation of a stimulus [8, 16]. The stimulus triggering an ERP can be motory, visual, auditory, or of any other sense (e.g., hand movements or perceiving audio).

Although EEG has been initially developed for medical applications and required high precision and accuracy, technological advancements within the last decade of both software and hardware have made it attractive for HCI applications [33]. While we still rely on medical-grade hardware and software to explore the feasibility of EEG for specific problems or approaches, researchers have already built a variety of increasingly small, wireless, and low-cost sensing devices for specific applications in everyday scenarios [10]. With research prototypes using printed electrodes connected to portable EEG devices such as the ones used by Debener et al. [9] or Bleichner et al. [6], the integration of EEG in users' everyday context does not seem out of reach anymore. For example, Bleichner et al. showed that they could achieve reliable measurements of specific ERPs. They were able to detect P300s, negative potentials that are reactions often occurring after surprising and unexpected events [38] and related to memory and attention processes [37]. To achieve this, they integrated miniaturized EEG electrodes into a baseball cap and an additional customized earpiece [6]. In a different approach, Vourvopoulos et al. modified a regular pair of glasses to include a low-cost EEG device, the OpenBCI[4], for future use in head-mounted displays. Their work shows promising first results in the investigation of cognitive and sensorimotor tasks by evaluation of frequency bands [52]. Finally, Kosch et al. [24] investigated the efficiency of EEG frequency bands for interface evaluations.

[1] www.store.neurosky.com, last accessed January 8th, 2020

[2] www.netflix.com, last accessed January 8th, 2020

[3] www.amazon.com, last accessed January 8th, 2020

[4] www.openbci.com, last accessed January 8th, 2020

ERPs for Language Processing

The application of ERPs to analyze problems during language processing has already been researched extensively in the neuroscience community. Syntactic and semantic problems during reading characteristically elicit N400 ERPs, whereas the N100 ERP is often indicating responses to auditory stimuli. Changes in amplitudes of these ERPs provide insights on various language processing problems based on individual words or sentence structures.

N400

An ERP component that is interesting for the evaluation of language processing and semantic relationships of words is the N400 component. An N400 is a negative deflection of the EEG signal around 250-500 ms (i.e., peaking at about 400 ms) after the presentation of a stimulus [26]. The N400 has shown to reflect on problems during semantically integrating a word into a sentence during reading (cf. [15, 25, 32]) during both visual and auditory word pair and sentence processing [19, 20]. Kutas & Hillyard showed participants reasonable sentences, containing either a word fitting the context or a word that was syntactically correct but semantically incongruous Examples included "They wanted to make the hotel look more like a tropical resort, [...] so they planted [tulips/ palms].". When reading the word "tulips", the authors report higher N400 in participants' neural responses [26]. Additionally, Holcomb & Neville showed higher N400 amplitudes for the processing of non-words or pseudowords (e.g., "jank", "grusp", "kcsrt") as compared to regular words. For the evaluation of foreign language reading comprehension, Schneegass et al. [47] employ N400 analysis and show significant differences between *known* and *unknown* words during reading. However, this approach is limited to presenting one stimulus at the time and to visual text presentation.

N100

The N100 component is frequently evaluated for the processing of auditory stimuli [43]. It is a typical component responding to the onset of a perceived sound with a negative deflation around 100 ms after the stimulus. It can occur in combination with a P200, an increased amplitude of the signal around 200 ms after a stimulus [43, 54]. The N100 is *known* to be an indicator of the auditory "oddball" phenomenon, which occurs when participants are presented with a set of familiar stimuli, followed by an unexpected stimulus [34]. Zhang et al. [57] investigated the N100-P200 complex for the audio presentation of pseudowords and were able to show significantly stronger negative responses as compared to regular words.

EVALUATING BRAINCODE

We hypothesize that the ERP components evaluated in related work can be transferred to second-language vocabulary comprehension assessment. When encountering an *unknown* or non-translatable word, we expect to measure similar N400 potentials during reading as Holcomb & Neville were able to show for non-words or pseudo-words [20]. We will focus on the evaluation of N100 as an indicator of unexpectedly presented stimuli during listening [34]. The objective of this work is to show that we can assess second-language vocabulary incomprehension while reading and listening using EEG. We conduct a user study, presenting participants with foreign language content using (1) text on screen (i.e., visual presentation) and (2) verbal narrations (i.e., auditory presentation). Participants read and listen to English texts which we manipulated to contain several potentially *unknown* words while recording their neural responses. We hypothesize that this manipulation will provoke a measurable neural reaction through greater amplitudes in the N400 reading and N100 while reading or listening, respectively. Furthermore, we show our process of classifying the neural responses we collected while participants encountered *known* and *unknown* words. By applying this classifier to a subset of our data, we calculate the accuracy and assess the potential of *BrainCoDe* for further use as a real-time comprehension detection tool.

Apparatus & Setup

We placed participants in a quiet, dimly lit room to reduce the risk of potential distractions. Participants sat at a fixed distance in front of a 24-inch desktop screen and we recorded their neural activities with a 32-channel EEG. We presented the text content on the screen and used a supplementary eyetracker to assess the users' focus of attention for the reading trials. Thus, we were able to match the EEG responses to individual words on the screen. We manipulated the texts carefully to contain several potentially *unknown* words. After each condition, we confirmed the manipulation (i.e., the comprehension of the vocabulary) using comprehension and translation questionnaires.

EEG & EOG Recording

To record the electric potentials generated by participants' brains, we used a Brain Products Live AMP[5], an EEG device with a 32 channel wireless electrode setup. The sampling rate was set to 500 Hertz (Hz) and the signal was automatically bandpass filtered between 0.1 and 1000 Hz. The electrodes were placed according to the 10-20 layout [21] (ground electrode: Fpz; reference electrode: FCz; see Figure 2). Conductive gel was used to reduce the impedance between electrodes and scalp. We ensured that the impedance was set to below 10 $k\Omega$ before starting the experiment.

For the evaluation of the EEG signals, we set four different markers in the software during the EEG recording to map the neural responses to the particular word shown on the screen. Those markers encoded the beginning of a text (marker "1") and a word's estimated difficulty. We differentiated *known* words (marker "2"), potentially *unknown* words (marker "3"), and words that were excluded from the analysis (marker "4"). The latter category contained words to be excluded from our evaluation for two reasons: (A) words that are shorter than three characters as users often skip them during reading [41, 42] and (B) proper names that do not necessarily have a translation since these are difficult to understand in the audio presentation. All markers were encoded into the EEG signal as a simulated keyboard input with a frequency of 8000 Hz.

[5] www.brainproducts.com/productdetails.php?id=63, last accessed January 8th, 2020

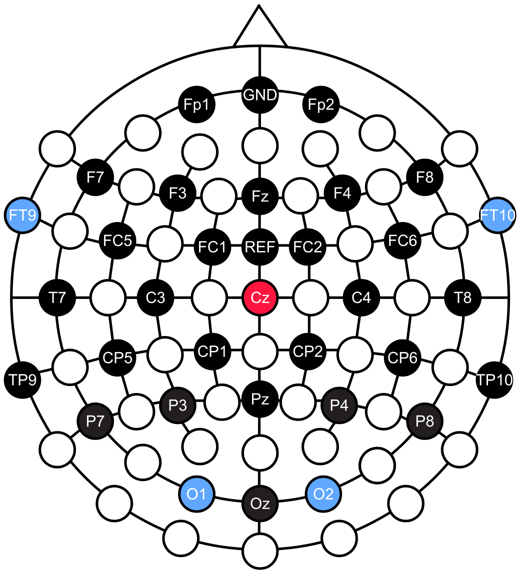


Figure 2. For the analysis, we used the Cz electrode (red) as positioned in the 10-20 layout [21]. For the measurement of EOG, we utilized the FT9, FT10, O1, and O2 electrodes (blue) and placed them around the participants' left and right eye.

From our 32-channel EEG setup, four electrodes were used for Electrooculography (EOG). EOG is used to record electric signals caused by muscles around the eye and works as an indicator of eye movements. Since eye movements are inevitable during reading, EOG enables us to filter the noise generated by muscles to create a clean recording of the actual brain responses. To record EOG, the electrodes were placed on the right and left canthi as well as above and below the left eye as suggested by prior work [13] using adhesive tape for medical use. We chose four electrodes from our setup (right eye: FT10; left eye: above: FT9, side: O2, below: O1), which are least likely to show responses to language processing, i.e., with the great distance to the central parietal area [25]. The remaining 28 electrodes were used for EEG recording.

Gaze Tracking

The EOG enables us to filter signals generated by muscles during the reading movements of the eye. However, it does not tell us the user's focus of attention. Knowing which word the user is focusing on is necessary to precisely map the resulting EEG response to the word that caused it. When presenting more than one stimulus at the same time, such as when presenting multiple words on the screen, eye-tracking can be used to map the gaze and, thus, the brain's focus to an individual word. In our setup, we used an EyeLink 1000+[6] which utilizes a video-based recording of eye gaze at 1000 Hz and was calibrated for each participant. A chin-rest is used to avoid re-calibration of the eye-tracker and maintain a fluent reading experience. Considering real-life settings, we expect reading to generate few to no head movement and are confident that valid gaze detection can be achieved without a chin-rest (e.g., use of a head-mounted eye-tracking device). Our implementation tracked participants' gaze and annotated the EEG signal accordingly.

Text Presentation

In the first part of our study, we assessed the participants' comprehension during *reading* on a computer screen. Texts were presented as a single centered line of black text on a grey background with a font size of 25. We set the maximum number of characters presented in a row to 40 to approximate subtitles in a movie while still being easily readable. If a sentence exceeded the character limit, the sentence was split. If a word would have been split due to the character limit, we pushed the whole word to the next line presentation instead. The system was designed to adapt to participants' reading speed: The next sentence was presented on the screen as soon as the eye-tracker recognized a short fixation of every word with at least three characters. By choosing a fixation time of only 1 ms, we ensured the recognition of each word while maintaining a natural reading flow.

In the second part of our study, we analyzed comprehension during *listening*. For the listening trials, we created narrated speech files using the Google Cloud text-to-speech (TTS) engine[7]. In contrast to a human reader, the engine ensured the creation of comparable texts and easy manipulation of individual words. For the speech presentation, we chose the default options suggested by Google's TTS engine, namely a female voice and American pronunciation. We reduced the speaking speed to 90% of the default for easier comprehension and less overlap across the neural reactions in the EEG signal. After downsampling the data to 1000 Hz, we analyzed the amplitudes of the resulting audio file to detect the beginning and end of the spoken words and list their timestamps. The resulting file format contained a list of all words, a timestamp for start and end, and the duration in ms (e.g., "cucumber", 2.258, 2.954, 0.696). In analogy to our process of encoding word markers into the EEG signal for the reading materials, we annotated the narrated words also in the list of spoken words, with the markers 1 (start of text), 2 (*known* word), 3 (*unknown* word), and 4 (excluded from analysis). To validate the accuracy of the timestamps, we visually analyzed the resulting audio file with an open-source audio software. During the listening trials, the screen displayed a red dot for participants to focus their gaze on to reduce unnecessary eye movement [50]. We used consumer earbud headphones to deliver the narrated texts.

Text Materials & Manipulation

To ensure the comparability of the two modalities reading and listening, we used standardized textual materials by Quinn & Nation [39, 40]. Besides a native language baseline, which we used to familiarize participants with our setup and text presentation, we assessed neural responses to individually presented English words and four English full texts, two for each modality (further explained in section "Procedure"). All texts used in this study were taken from a corpus designed for English Second-Language (ESL) Learners that includes 15 texts on

[6] www.sr-research.com/products/eyelink-1000-plus, last accessed January 8th, 2020

[7] https://cloud.google.com/text-to-speech, last accessed January 8th, 2020

Table 1. We evaluated two native texts, a set of individual words, and four full texts. This table outlines the number of words and sentences per condition as well as the number of difficult words we induced.

Language	Trial Phase	Total Number of Sentences	Total Number of Words	Total Number of Difficult Words	Duration Audio Narration (in min)
German	Native Baseline Ge 1	15	214	0	1:42
	Native Baseline Ge 2	15	238	0	1:28
English	Individual Words IW1	-	30	15	3:18
	Individual Words IW2	-	30	15	3:13
	Full Text En 1	41	514	15	3:12
	Full Text En 2	46	517	15	3:24
	Full Text En 3	46	542	15	3:21
	Full Text En 4	44	540	15	3:10

various topics and complementary multiple-choice comprehension questionnaires [39]. The texts are designed to have easy grammar and feature frequent words [40]. Therefore, they are supposed to be easily understandable by participants with low-level English skills. For the listening trials, the written texts were transformed into narrations with a TTS engine as explained above.

To evaluate vocabulary gaps during reading and listening, we manipulated the texts to include a number of rare and potentially *unknown* words. Hence, we incorporated words from lists containing rare or uncommon words in the English language, which are difficult even for a native speaker (e.g., Oxford Lexico's "Weird and Wonderful Words" List[8] or the "Archaic Words" List[9]). Whether a word was actually *unknown* to the user was later confirmed through translation questionnaires as explained in our procedure. With our setup of easily comprehensible texts we aim to eliminate potential word or grammar difficulties that could interfere with our study's manipulations.

We randomly chose four *Full Texts* from the ESL corpus [39]. Two texts were used for the visual presentation (En 1, En 2) and two for the auditory evaluation (En 3, En 4). The texts had a mean of 44.24 sentences ($SD = 2.05$) and 528.25 words per text ($SD = 12.81$) (cf. Table 1). We manipulated 15 sentences (i.e., one out of three) of each text to contain one potentially *unknown* word. The following sentences are an excerpt of the English text En 1, including two manipulated words:

> "They want to remember their culture and teach their ***progeny*** the old ways."
>
> "Sometimes the Inuit ***seethed*** their food but often it was not cooked at all."

The percentage of manipulations remained low so that they did not affect overall text comprehension and created a realistic scenario as it could occur during the use of media content. We randomized the presentation of texts within the conditions, to avoid content-related effects.

For the *Native Baselines*, we translated one text carefully into German, the participants' native language. This text was later split into two shorter parts, Ge 1 and Ge 2, with each 15 sentences to serve as baseline for both the reading and listening condition and familiarize the participants with the text and auditory presentation. We decided to split the text to create two short comprehensive texts to not strain the user's attention before starting the presentation of the English content. The native baseline was presented as the first condition in both modalities and did not include any manipulations.

Furthermore, we presented the participants with 60 random *Individual Words* to assess the neural reaction time on a single word basis without influences of overall text comprehension. The individual words can provide insights on potential offsets of response time induced by our setup. The words include verbs, nouns, and adjectives, half of them easily understandable, and half of them supposedly difficult (cf. procedure full-text manipulation).

Comprehension Questionnaires and Translations

We assessed the participant's text understanding using 10 prevalidated multiple-choice comprehension questions provided in the corpus by Quinn & Nation [39] (five for the native baselines). For example, for the text "Life in the South Pacific Islands", the following question is given:

> *Thousands of years ago, people came to the Pacific Islands from*
>
> *a) South America.*
> *b) Asia.*
> *c) Australia.*
> *d) Europe.*

In addition to the comprehension questions, we asked the participants to translate the manipulated and potentially difficult words. After each full text and the individual word presentation, we provided the participants with a translation test, containing a list of the difficult English words with blanks next to them to fill in the correct translation.

Procedure

After welcoming the participants and explaining the process of the user study, they gave informed consent for participation and data handling following the European GDPR. Next, participants chose a random ID from a sheet of prepared user IDs

[8] www.lexico.com/en/explore/weird-and-wonderful-words, last accessed January 8th, 2020

[9] www.lexico.com/en/explore/archaic-words, last accessed January 8th, 2020

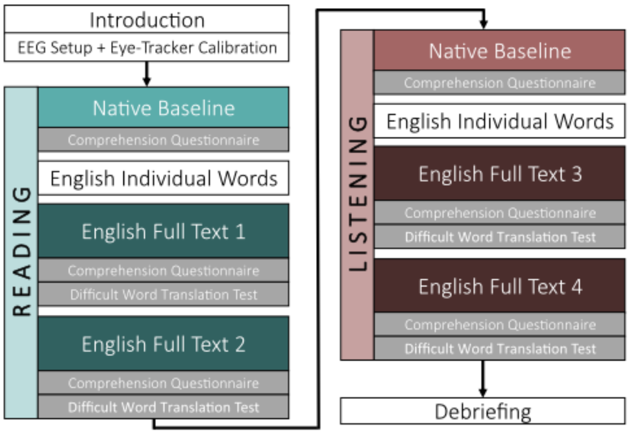

Introduction
EEG Setup + Eye-Tracker Calibration
READING
Native Baseline
Comprehension Questionnaire
English Individual Words
English Full Text 1
Comprehension Questionnaire
Difficult Word Translation Test
English Full Text 2
Comprehension Questionnaire
Difficult Word Translation Test
LISTENING
Native Baseline
Comprehension Questionnaire
English Individual Words
English Full Text 3
Comprehension Questionnaire
Difficult Word Translation Test
English Full Text 4
Comprehension Questionnaire
Difficult Word Translation Test
Debriefing

Figure 3. After welcoming our participants and preparing our study setup, the participants took part in the evaluation of four reading and four listening trials. Each trial consisted of the presentation of a native baseline, a set of individual words, and two full texts.

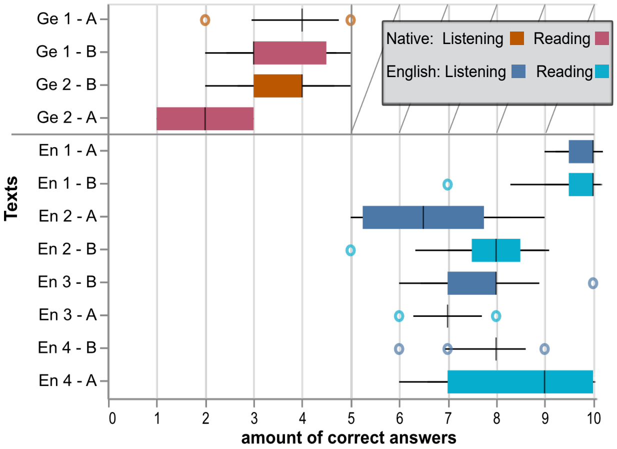


Figure 4. Amount of correct answers in the comprehension questionnaires for German (Ge) and English (En) texts across participants of the randomized groups A and B. We compare the results for listening (orange/dark blue) and reading (red/light blue).

to ensure the anonymization of the data and we introduced them to the EEG and eye-tracking setup. Participants filled in a questionnaire asking about demographic data, including age, highest education level, gender, vision impairment, and history of neurological diseases. We assigned the participants randomly to two conditions, resulting in a changed sequence of text presentation within the two modalities. Afterward, they passed through the reading and listening phase, each including one native baseline text, a condition presenting 60 individual words, and two full texts with additional questionnaires (cf. Figure 3). Following the text presentation, the participants were asked to fill in the respective comprehension and vocabulary translation questionnaire. Overall, the participation in our user study took around 110 minutes (including electrode setup, debriefing, and cleaning the electrode caps). As a study compensation, participants could choose between a 20€ Amazon voucher or study participation points offered by our university.

EEG Data Processing

To analyze the recorded data, we used the Python MNE library[10] and resampled the raw EEG data to 250 Hz. Afterward, the data was high pass filtered at 1 Hz and low pass filtered at 125 Hz. The data was then re-referenced to the average of all channels which included the original reference electrode FCz. To clean our data, we used a notch filter to remove the 50 Hz powerline noise. We then extracted the epochs and rejected every epoch with an amplitude of higher than 200 μv around the EOG electrodes to remove ocular artifacts from the analysis. Afterward, the data was high pass filtered with 0.2 Hz and low pass filtered with 35 Hz. Finally, we sliced the epochs into blocks of -0.3 ms and 0.7 ms, where 0.0 ms denotes the onset of the stimulus. We automatically extracted the ERP negativity peaks and their latencies. For detection of the N100 during listening, we located the minimum peak in a 50 ms to 150 ms time window after stimulus onset, using a 10 Hz low pass filter. For the N400 ERP detection during reading, we chose a 350 ms to 450 ms time window, respectively.

RESULTS

In the following, we elaborate on the evaluated sample and investigate differences in the ERP amplitudes for *known* and *unknown words*. Furthermore, we assess the feasibility of classifying ERPs to detect vocabulary gaps in real-time.

Sample

We recruited 16 participants (nine identifying as female, seven as male) through our university's internal mailing lists, Facebook page, and Slack channel. The age range of our participants was 20 to 55 ($M = 24.25$, $SD = 8.09$), with 13 participants having a high school degree, two having a master's degree, and one having a secondary school degree. Five participants stated to wear glasses. They were asked to remove the glasses to increase the eye-tracking accuracy for optimal recognition of reading behavior. All participants reported being able to read the text shown on the screen without any problem. Due to an issue in the study setup, the first four participants had to be excluded from the reading trials since the EEG signal was incorrectly mapped to the words the participants were reading. The mapping worked accurately for the subsequent twelve participants after resolving this issue. This error did not have any influence on the listening trials. Thus, the full sample size for the listening trials remained 16.

Text and Vocabulary Comprehension

The evaluation of the *overall comprehension questionnaires* showed a medium to high text comprehension rate across all texts. Figure 4 sums up the results for the two conditions. Randomization group A included seven participants from which two had to be excluded from the reading trials. Randomization group B contained nine participants, excluding two from reading. For the German baseline, participants achieved a median of four correct answers out of five questions during listening and a slightly lower median of three correct answers during reading for all comprehension scores.

For the English full texts, we notice only minor differences in comprehension scores between the four texts and the modal-

[10] www.mne.tools/stable/index.html, last accessed January 8th, 2020

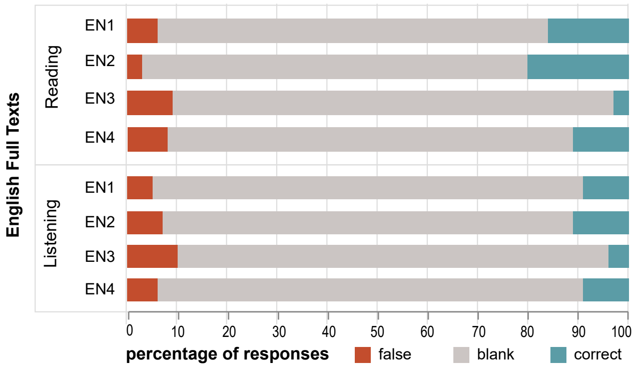


Figure 5. Amount of false, blank, and correct answers in the vocabulary translation questionnaires for both modalities.

ities. For text En 1, participants achieved a median of ten out of ten correct answers in both modalities; for En 2, they reached eight correct answers for reading and 6.5 for listening. Within texts En 3 and En 4, participants scored a median of eight correct answers for listening, seven and nine for reading respectively. The minimum amount of correct answers across all participants for the English text comprehension questionnaires was five. Thus, we can assume an appropriate level of comprehension for all study participants across the English language texts.

Additional to the overall comprehension, we presented participants with *vocabulary translation tests*, asking to fill in the German word for the potentially *unknown* English words we used as manipulation. The participants were able to translate 7.76% of all difficult words correctly. The majority of questions were left blank, as can be seen in Figure 5. Based on the translation tests, we excluded all potentially difficult words, which the participants were able to translate, from our analysis. Thus, we can clearly differentiate between *known* and *unknown* words.

Evaluating Event-Related Potentials

We statistically analyze the amplitudes of the ERPs generated by *known* and *unknown* English words. Accordingly, we investigate ERPs across both text presentation conditions, individual words, and the full texts for the respective reading and listening trials. We focus our evaluation on the electrode Cz. It is frequently reported in related work that the N400 is larger over the central region of the scalp [26–28]. For the N100, the Cz electrode also shows higher amplitudes for unexpected stimuli and is also called "vertex potential" [49].

Using the extracted negativity peaks for the N400s and N100s, we average the resulting amplitudes for the full-text conditions of reading and listening. A Shapiro-Wilk test shows a deviation from normality when listening to narrated sentences or individual words ($p < .05$). Thus, we proceed with the non-parametric Wilcoxon-Signed rank test for the analysis of the listening trials. The test reveals a statistically significant difference in the N100 amplitudes between *known* and *unknown* words for individual narrated words ($Z = 78$, $p < .001$; see Figure 6a) as well as listening to narrated full text ($Z = 300$, $p < .001$; see Figure 7a). The results indicate that there are measurable differences in the auditory processing of *known* and *unknown* words.

A Shapiro-Wilk test does not indicate a deviation from normality when reading sentences ($p > .05$). Therefore, we submit the N400 amplitudes of the reading trials to a t-test for statistical analysis. We find a statistically significant difference in the N400 amplitudes when reading individual words ($t(15) = 14.327$, $p < .001$, $d = 1.531$; see Figure 6b) as well as full texts ($t(23) = 23.307$, $p < .001$, $d = 0.758$; see Figure 7b) with a large effect size. These results suggest that there are measurable differences in N400 amplitudes, indicating differences when processing *known* and *unknown* words during reading.

Predicting Vocabulary Gaps

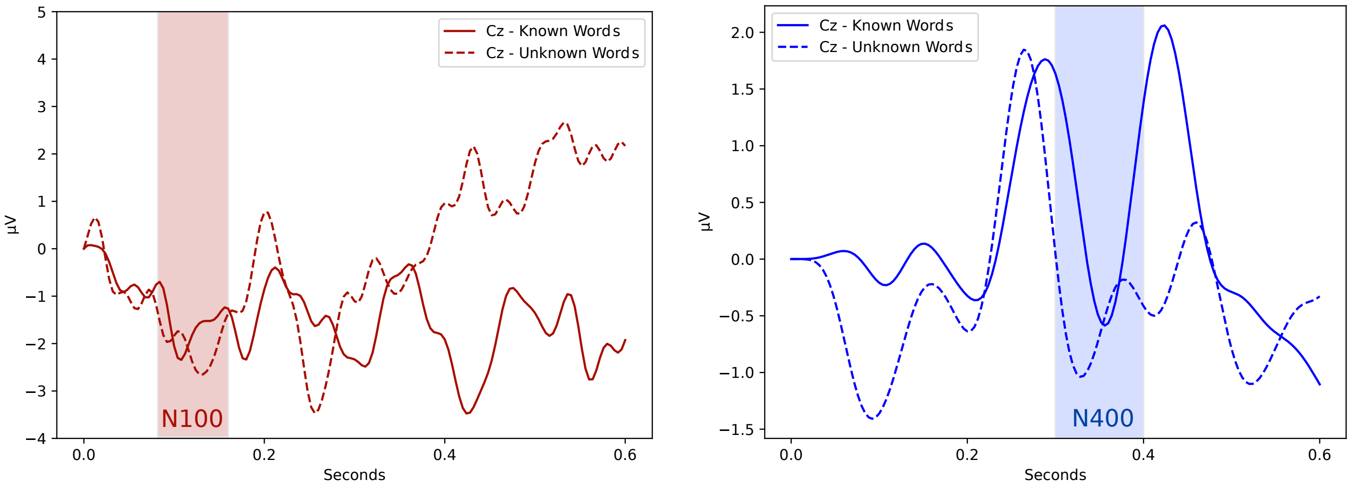
The results from our study show significant differences between the N400 and N100 ERPs when reading and listening to *known* or *unknown* words. We investigate the performance of person-dependent classification based on the extracted ERP amplitudes.

Features, Instances, and Classifier Performance

We apply the same data processing as in the analysis mentioned before to extract the ERP amplitudes. Thereby, we focus on the N400 amplitude to detect word comprehensions during the reading of sentences and on the N100 amplitudes when detecting vocabulary gaps in auditory narrated text. Separately for reading sentences and audio listening, the N400 amplitudes were used for the reading trials and N100 amplitudes for auditory trials. We labeled words afterward as a *known* word or *unknown* word. If *unknown* words were translated correctly after each trial, they were labeled *known* words. The number of epochs was different for each participant since a different number of epochs was rejected. Therefore, per participant between 450 and 510 epochs ($M = 482.75$, $SD = 18.9$, $N = 11586$) with *known* words and between 12 to 15 epochs ($M = 13$, $SD = 1.3$, $N = 312$) with *unknown* words were taken into account during the reading trial. Between 487 and 509 ($M = 498.5$, $SD = 16.3$, $N = 11964$) epochs for *known* words and 13 to 15 ($M = 14.2$, $SD = 1.2$, $N = 341$) epochs for *unknown* words per participant were considered for further analysis for the auditory trials. According to the remaining epochs, the N400 and N100 amplitudes were labeled for classification and were the only features used.

To sustain a natural scenario, the number of the attributes with the label *known* word is much higher than the number of attributes with the label *unknown* word. Therefore, we slice the number of attributes with the label *known* word to the same number as *unknown* words for each participant. Using scikit-learn[11], we train a Support Vector Machine (SVM) with a radial kernel. We perform a cross-validation on the trained instances with $k = 5$, where $k-1$ folds were iteratively used for training while the remaining fold was used for evaluation. This process was repeated until *known* words were evaluated with an equal set of *unknown* words. Finally, we calculate the mean

[11] www.scikit-learn.org, last accessed January 8th, 2020



(a)

(b)

Figure 6. ERP responses measured at the Cz electrode for hearing individual narrated words and reading written words. (a): *Unknown* words generate greater N100 amplitudes during auditory presentation as compared to *known* words. (b): For visual presentation, *unknown* words elicit greater N400 amplitudes as *known* words.

from the resulting classification scores to aggregate an overall accuracy. The average overall accuracy in discriminating *known* words and *unknown* words during reading was 87.13% ($SD = 4.5\%$). Auditory trials resulted in an overall accuracy of 82.64% ($SD = 9.6\%$). In both modalities, the majority of wrongly classified words were false positives (FP) compared to false negatives (FN) (listening FP = 10.96%, FN = 6.4%, reading FP = 9.35%, FN = 3.52%; see Figure 8). Hence, in case of wrong classification, words are more likely to be classified as *unknown* words, although the word is already *known* during both listening and reading.

Limitations

Like many EEG-based studies, our evaluation was done in a quiet environment with no distractions. This was suitable to obtain data with limited noise from outside sources to be able to explore the EEG data. However, the discrepancies between the experimental conditions applied in our study and real-life situations can not be neglected. We need further evaluation to explore the feasibility of *BrainCoDe* to detect comprehension problems in everyday scenarios with potentially influencing environmental factors such as noise or complex media such as movies.

DISCUSSION

In this work, we investigated the neural responses during vocabulary comprehension. We found significant differences in the size of amplitudes in N400 ERPs during reading and in N100 ERPs during listening as neural responses to *unknown* words. The classification of the responses performed with an accuracy of above 80%. The *BrainCoDe* concept worked effectively for the detection of vocabulary comprehension problems in second-language reading and listening. We discuss the opportunities of our approach to be utilized in real-world scenarios.

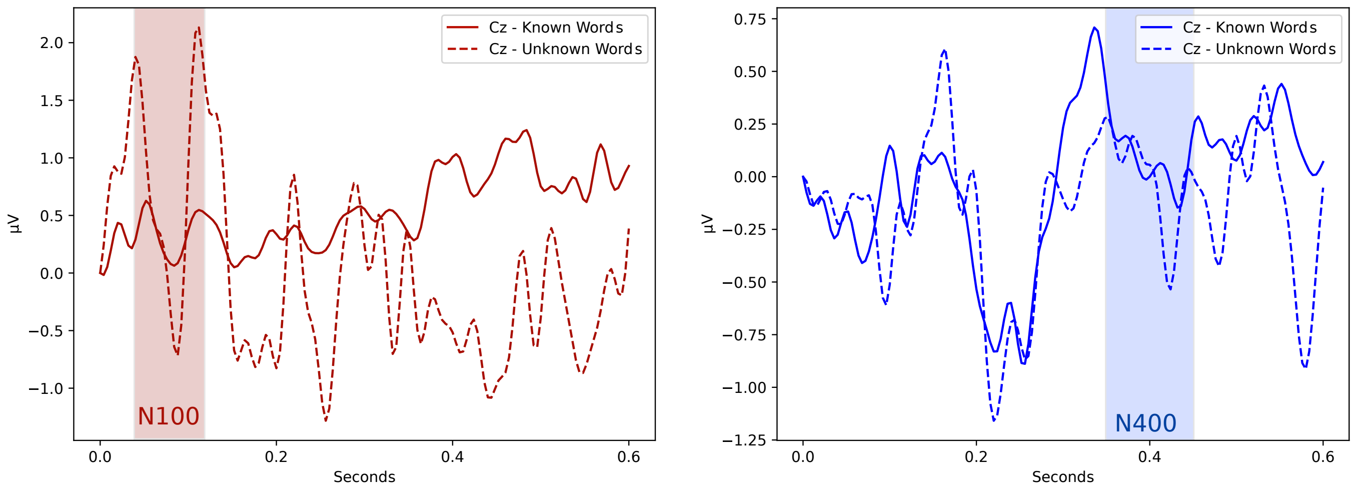
Exploring Language Learning Modalities

In our study, we focused on reading and listening as modalities important for language learning. We achieved a classification accuracy of 87.13% when differentiating between reading *known* and *unknown* words on screen. This can be used for situations such as reading an e-book in a new language on a tablet. The analysis of EEG responses in such a situation would still require eye-tracking to be able to pinpoint the exact word the user is looking at at the moment. However, this technology is currently finding its way to commodity tablets, PCs, and smartphones [23, 53].

The *BrainCoDe* approach achieved a classification accuracy of 82.64% for the listening modality. This approach can be used in a real-world learning scenario, such as listening to an audiobook in a second-language with slow narration speed or listening to recorded conversations in an online language class. However, finding comprehension problems using our approach in more complex media such as movies that contain multiple stimuli and modalities would require further investigation. Fortunately for these scenarios, a perfect classification accuracy is not necessary to build successful learning applications, since including false positives, thus, repeating already *known* content, does not hinder learning.

Person-dependent Learning

Currently, the evaluation of ERP data only creates expressive results through averaging over many trials and is highly user-specific [30]. In order to use EEG responses as input for applications, a training phase is required [29]. Similar to most state-of-the-art systems utilizing physiological sensors, *BrainCoDe* requires a user-dependent training phase to be able to detect vocabulary incomprehension due to unique manifestations of ERPs for every individual user. Single-trial ERP classification is currently advancing and shows promising



(a)

(b)

Figure 7. ERP responses measured at the Cz electrode for *known* and *unknown* words perceived during narrated full texts and during reading full texts. (a): *Unknown* words generate greater N100 amplitudes during auditory presentation as compared to *known* words. (b): For visual presentation, *unknown* words elicit greater N400 amplitudes as *known* words.

results [5] that would make our approach feasible for real-time classification of *unknown* vocabulary in the future.

In addition to inter-person variations of ERPs, the momentary cognitive and physiological state of each user may have an effect on their neural responses per session. The response to repeated *known* and *unknown* words over time may also be different as the user starts to learn the language. Related work in the use of ERPs for language learning also suggests that the neural responses (e.g., N400) of listening stimuli are susceptible to habituation effects, for example, when the user listens to an unexpected stimulus frequently [7]. These factors must be taken into consideration when designing a language learning application utilizing neural responses.

In this work, we focused our analysis on the N400 and N100 ERP components based on prior work. Nonetheless, the closer investigation of other ERP features or potential feature interactions can yield further insights into comprehension problems during listening and reading and improve the classification accuracy.

BrainCoDe Application Scenarios

We envision a personal learning application that provides both real-time and post-hoc personalized feedback. The users would start by wearing a brain-computer interface (BCI) headset, such as the Emotiv EPOC[12] or the OpenBCI[4], both of which have support for the used electrode setups verified by BrainCoDe. The users would go through a training session by reading and listening to phrases with *known* and *unknown* words to assess their current language level and collect training data, then the media content (e.g., audiobook, movie, e-book, or language learning chat session) would be started. The application can provide real-time or post hoc feedback. Real-time feedback could recommend the user to pause, rewind a scene of the media content, or repeat a section of the audiobook to increase exposition to the new vocabulary and enhance learning efficiency. As a post hoc analysis tool, the *BrainCoDe* approach easily facilitates the extraction of *unknown* vocabulary to create a personalized list of contents for the user to learn. Based on a continuous monitoring of the users' language comprehension, *BrainCoDe* could be applied to provide recommendations on media contents adapted to users' proficiency as explored by Yuksel et al. for learning the Piano [55].

Integrating our approach into applications such as Duolingo[13], we could implement a tool to adapt the content according to the user's knowledge. Thus, the app could present content, which is currently *unknown* to the user, with higher frequency to support learning.

CONCLUSION AND FUTURE WORK

The detection of gaps in a learner's vocabulary knowledge is a critical step to facilitate effective second-language learning. Since media consumption in different languages is nowadays a common tool to learn and improve on a language, such as in the form of movies, e-books, or audiobooks, the implicit detection of comprehension problems becomes increasingly important. Avoiding interruptions and distractions while consuming media helps to improve the user experience. Understanding what words are *unknown* to the user allows proactively presenting translations or explanations without interrupting the media consumption.

In this work, we present BrainCoDe, an EEG-based method to implicitly detect *unknown* words during foreign language reading and listening. In a user study ($N = 16$), we show the feasibility of evaluating N100 and N400 event-related potentials using the single Cz electrode, to detect participants' English vocabulary knowledge gaps. By building a classifier

[12] www.emotiv.com/epoc, last accessed January 8th, 2020

[13] www.duolingo.com, last accessed January 8th, 2020

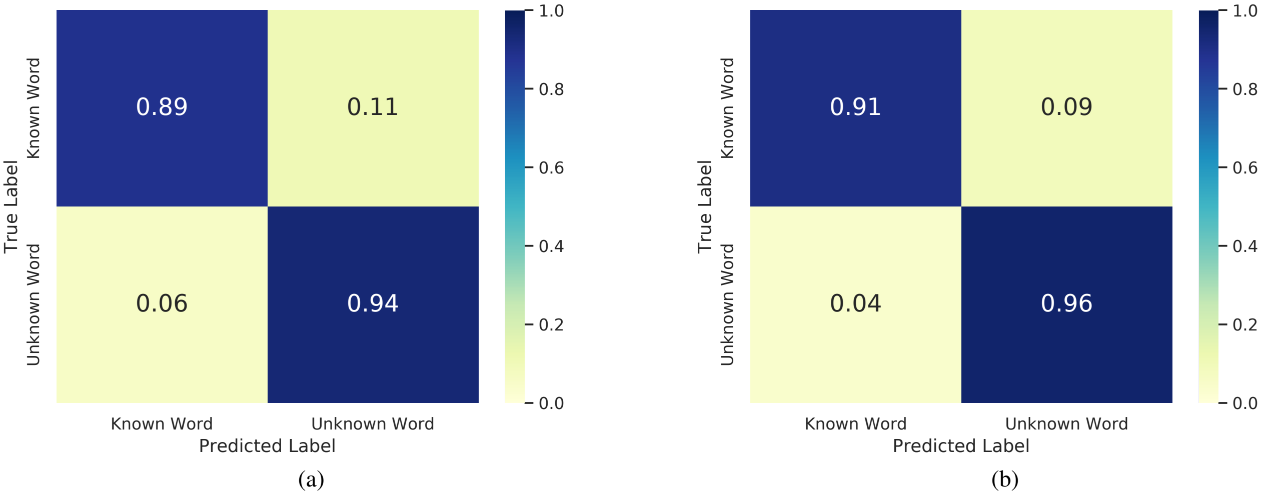


Figure 8. Normalized confusion matrix of classifying *known* words and *unknown* words between the (a) auditory trials and (b) reading trials.

trained to identify those gaps, we can successfully recognize *unknown* vocabulary in eight out of ten situations. Thereby, the accuracy of reading (87.13%) exceeds the accuracy of detecting *known* and *unknown* words during listening to narrated content (82.64%).

For future work, we plan to further evaluate two components of our approach: (1) we aim to test the feasibility and robustness of our approach when deployed with low-cost EEG sensors. We imagine using consumer-grade hardware such as the OpenBCI kit for augmenting headphones and analyzing the Cz electrode for the occurrence of the N400s and N100s. Furthermore, we (2) aim to develop our approach further to be applicable for real-time user support. After a short training phase, we expect that our classification can support live detection of *unknown* words, thus aiming to make *BrainCoDe* a promising tool for being integrated into learning applications.

With the help of *BrainCoDe*, we envision to use EEG-based implicit comprehension detection to build language proficiency aware interfaces. With only a minimum number of electrodes, we see great potential for the identification of vocabulary gaps during everyday media consumption. With our evaluation of *BrainCoDe* we take the first steps toward enabling vocabulary learning while users can enjoy their favorite media contents. We believe that in the long run, progress in BCI research will transform learning tasks into a joyful experience.

REFERENCES

[1] Olivier Augereau, Hiroki Fujiyoshi, and Koichi Kise. 2016. Towards an automated estimation of English skill via TOEIC score based on reading analysis. In *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 1285–1290. DOI: http://dx.doi.org/10.1109/ICPR.2016.7899814

[2] Sylvain Baillet, John C Mosher, and Richard M Leahy. 2001. Electromagnetic brain mapping. *IEEE Signal processing magazine* 18, 6 (2001), 14–30. DOI: http://dx.doi.org/10.1109/79.962275

[3] Yevgeni Berzak, Boris Katz, and Roger Levy. 2018. Assessing Language Proficiency from Eye Movements in Reading. *arXiv preprint arXiv:1804.07329* (2018).

[4] Yevgeni Berzak, Chie Nakamura, Suzanne Flynn, and Boris Katz. 2017. Predicting native language from gaze. *arXiv preprint arXiv:1704.07398* (2017).

[5] Benjamin Blankertz, Steven Lemm, Matthias Treder, Stefan Haufe, and Klaus-Robert Müller. 2011. Single-trial analysis and classification of ERP components - a tutorial. *NeuroImage* 56, 2 (2011), 814–825. DOI: http://dx.doi.org/10.1016/j.neuroimage.2010.06.048

[6] Martin G Bleichner, Micha Lundbeck, Matthias Selisky, Falk Minow, Manuela Jäger, Reiner Emkes, Stefan Debener, and Maarten De Vos. 2015. Exploring miniaturized EEG electrodes for brain-computer interfaces. An EEG you do not see? *Physiological reports* 3, 4 (2015). DOI: http://dx.doi.org/10.14814/phy2.12362

[7] Timothy W Budd, Robert J Barry, Evian Gordon, Chris Rennie, and Patricia T Michie. 1998. Decrement of the N1 auditory event-related potential with stimulus repetition: habituation vs. refractoriness. *International Journal of Psychophysiology* 31, 1 (1998), 51–68.

[8] Doug J Davidson. 2012. Brain Activity During Second Language Processing (ERP). *The Encyclopedia of Applied Linguistics* (2012). DOI: http://dx.doi.org/10.1002/9781405198431.wbeal0106

[9] Stefan Debener, Reiner Emkes, Maarten De Vos, and Martin Bleichner. 2015. Unobtrusive ambulatory EEG using a smartphone and flexible printed electrodes around the ear. *Scientific reports* 5 (2015), 16743. DOI: http://dx.doi.org/10.1038/srep16743

[10] Stefan Debener, Falk Minow, Reiner Emkes, Katharina Gandras, and Maarten De Vos. 2012. How about taking a low-cost, small, and wireless EEG for a walk? *Psychophysiology* 49, 11 (2012), 1617–1621. DOI: http://dx.doi.org/10.1111/j.1469-8986.2012.01471.x

[11] Jérémy Frey, Christian Mühl, Fabien Lotte, and Martin Hachet. 2014. Review of the Use of Electroencephalography as an Evaluation Method for Human-Computer Interaction. (2014), 214–223. DOI: http://dx.doi.org/10.5220/0004708102140223

[12] Katsuya Fujii and Jun Rekimoto. 2019. SubMe: An Interactive Subtitle System with English Skill Estimation Using Eye Tracking. In *Proceedings of the 10th Augmented Human International Conference 2019*. ACM, 23. DOI: http://dx.doi.org/10.1145/3311823.33118650

[13] Christiane Glatz, Stas S Krupenia, Heinrich H Bülthoff, and Lewis L Chuang. 2018. Use the right sound for the right job: verbal commands and auditory icons for a task-management system favor different information processes in the brain. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 472. DOI: http://dx.doi.org/10.1145/3173574.3174046

[14] Ramazan Goctu. 2017. Using movies in EFL classrooms. *European Journal of Language and Literature* 3, 2 (2017), 121–124. DOI: http://dx.doi.org/10.5539/elt.v9n3p248

[15] Peter Hagoort. 2007. The fractionation of spoken language understanding by measuring electrical and magnetic brain signals. *Philosophical Transactions of the Royal Society B: Biological Sciences* 363, 1493 (2007), 1055–1069. DOI: http://dx.doi.org/10.1098/rstb.2007.2159

[16] Peter Hagoort, Lea Hald, Marcel Bastiaansen, and Karl Magnus Petersson. 2004. Integration of word meaning and world knowledge in language comprehension. *science* 304, 5669 (2004), 438–441. DOI:http://dx.doi.org/10.1126/science.1095455

[17] Mariam Hassib, Stefan Schneegass, Philipp Eiglsperger, Niels Henze, Albrecht Schmidt, and Florian Alt. 2017. EngageMeter: A system for implicit audience engagement sensing using electroencephalography. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 5114–5119. DOI: http://dx.doi.org/10.1145/3025453.3025669

[18] Abdolmajid Hayati and Firooz Mohmedi. 2011. The effect of films with and without subtitles on listening comprehension of EFL learners. *British Journal of Educational Technology* 42, 1 (2011), 181–192. DOI: http://dx.doi.org/10.1111/j.1467-8535.2009.01004.x

[19] Phillip J Holcomb, Sharon A Coffey, and Helen J Neville. 1992. Visual and auditory sentence processing: A developmental analysis using event-related brain potentials. *Developmental Neuropsychology* 8, 2-3 (1992), 203–241. DOI: http://dx.doi.org/10.1080/87565649209540525

[20] Phillip J Holcomb and Helen J Neville. 1990. Auditory and visual semantic priming in lexical decision: A comparison using event-related brain potentials. *Language and cognitive processes* 5, 4 (1990), 281–312. DOI:http://dx.doi.org/10.1080/01690969008407065

[21] Herbert H Jasper. 1958. The ten-twenty electrode system of the International Federation. *Electroencephalogr. Clin. Neurophysiol.* 10 (1958), 370–375.

[22] Jakob Karolus, Paweł W Wozniak, Lewis L Chuang, and Albrecht Schmidt. 2017. Robust gaze features for enabling language proficiency awareness. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2998–3010. DOI: http://dx.doi.org/10.1145/3025453.3025601

[23] Mohamed Khamis, Florian Alt, and Andreas Bulling. 2018. The past, present, and future of gaze-enabled handheld mobile devices: Survey and lessons learned. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 38. DOI: http://dx.doi.org/10.1145/3229434.3229452

[24] Thomas Kosch, Markus Funk, Albrecht Schmidt, and Lewis L Chuang. 2018. Identifying Cognitive Assistance with Mobile Electroencephalography: A Case Study with In-Situ Projections for Manual Assembly. *Proceedings of the ACM on Human-Computer Interaction* 2, EICS (2018), 11.

[25] Marta Kutas and Kara D Federmeier. 2000. Electrophysiology reveals semantic memory use in language comprehension. *Trends in cognitive sciences* 4, 12 (2000), 463–470. DOI: http://dx.doi.org/10.1016/S1364-6613(00)01560-6

[26] Marta Kutas and Steven A Hillyard. 1980. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science* 207, 4427 (1980), 203–205. DOI: http://dx.doi.org/10.1126/science.7350657

[27] Marta Kutas and Steven A Hillyard. 1983. Event-related brain potentials to grammatical errors and semantic anomalies. *Memory & cognition* 11, 5 (1983), 539–550.

[28] Marta Kutas, C Van Petten, and M Besson. 1988. Event-related potential asymmetries during the reading of sentences. *Electroencephalography and clinical neurophysiology* 69, 3 (1988), 218–233.

[29] Fabien Lotte. 2014. A tutorial on EEG signal-processing techniques for mental-state recognition in brain–computer interfaces. In *Guide to Brain-Computer Music Interfacing*. Springer, 133–161.

[30] Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger. 2018. A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update. *Journal of neural engineering* 15, 3 (2018), 031005. DOI: http://dx.doi.org/10.1088/1741-2552/aab2f2

[31] Fabien Lotte, Marco Congedo, Anatole Lécuyer, Fabrice Lamarche, and Bruno Arnaldi. 2007. A review of classification algorithms for EEG-based brain–computer interfaces. *Journal of neural engineering* 4, 2 (2007), R1. DOI:http://dx.doi.org/10.1088/1741-2560/4/2/R01

[32] Jutta L Mueller. 2005. Electrophysiological correlates of second language processing. *Second Language Research* 21, 2 (2005), 152–174. DOI: http://dx.doi.org/10.1191%2F0267658305sr256oa

[33] Tim R Mullen, Christian AE Kothe, Yu Mike Chi, Alejandro Ojeda, Trevor Kerth, Scott Makeig, Tzyy-Ping Jung, and Gert Cauwenberghs. 2015. Real-time neuroimaging and cognitive monitoring using wearable dry EEG. *IEEE Transactions on Biomedical Engineering* 62, 11 (2015), 2553–2567. DOI: http://dx.doi.org/10.1109/TBME.2015.2481482

[34] Risto Näätänen and Terence Picton. 1987. The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology* 24, 4 (1987), 375–425.

[35] P David Pearson, Elfrieda H Hiebert, and Michael L Kamil. 2007. Vocabulary assessment: What we know and what we need to learn. *Reading research quarterly* 42, 2 (2007), 282–296. DOI: http://dx.doi.org/10.1598/RRQ.42.2.4

[36] Paul R Pintrich. 2003. A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of educational Psychology* 95, 4 (2003), 667. DOI: http://dx.doi.org/10.1037/0022-0663.95.4.667

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

[38] Walter S Pritchard. 1981. Psychophysiology of P300. *Psychological bulletin* 89, 3 (1981), 506.

[39] Elizabeth Quinn and Ian Stephen Paul Nation. 1974. *Speed reading: A course for learners of English*. Oxford University Press.

[40] Elizabeth Quinn, Ian Stephen Paul Nation, and Sonia Millett. 2007. Asian and Pacific speed readings for ESL learners. *English Language Institute Occasional Publication* 24 (2007).

[41] Keith Rayner and George W McConkie. 1976. What guides a reader's eye movements? *Vision research* 16, 8 (1976), 829–837. DOI: http://dx.doi.org/10.1016/0042-6989(76)90143-7

[42] Keith Rayner, Timothy J Slattery, Denis Drieghe, and Simon P Liversedge. 2011. Eye movements and word skipping during reading: effects of word length and predictability. *Journal of Experimental Psychology: Human Perception and Performance* 37, 2 (2011), 514. DOI:http://dx.doi.org/10.1037/a0020990

[43] Gerard Remijn, Emi Hasuo, Haruna Fujihira, and Satoshi Morimoto. 2014. An introduction to the measurement of auditory event-related potentials (ERPs). *Acoustical Science and Technology* 35 (01 2014), 229–242. DOI:http://dx.doi.org/10.1250/ast.35.229

[44] Jack C Richards. 2015. The changing face of language learning: Learning beyond the classroom. *RELC Journal* 46, 1 (2015), 5–22. DOI: http://dx.doi.org/10.1177/0033688214561621

[45] Jelisaveta Safranj. 2015. Advancing listening comprehension through movies. *Procedia-Social and Behavioral Sciences* 191 (2015), 169–173. DOI: http://dx.doi.org/10.1016/j.sbspro.2015.04.513

[46] Charles Lima Sanches, Koichi Kise, and Olivier Augereau. 2017. Japanese Reading Objective Understanding Estimation by Eye Gaze Analysis. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers (UbiComp '17)*. ACM, New York, NY, USA, 121–124. DOI: http://dx.doi.org/10.1145/3123024.3123092

[47] Christina Schneegass, Thomas Kosch, Albrecht Schmidt, and Heinrich Hussmann. 2019. Investigating the Potential of EEG for Implicit Detection of Unknown Words for Foreign Language Learning. In *Proceedings of the 2019 INTERACT Conference*. ACM. DOI: http://dx.doi.org/10.1007/978-3-030-29387-1_17

[48] Alireza Sahami Shirazi, Mariam Hassib, Niels Henze, Albrecht Schmidt, and Kai Kunze. 2014. What's on your mind?: mental task awareness using single electrode brain computer interfaces. In *Proceedings of the 5th Augmented Human International Conference*. ACM, 45. DOI:http://dx.doi.org/10.1145/2582051.2582096

[49] Shravani Sur and VK Sinha. 2009. Event-related potential: An overview. *Industrial psychiatry journal* 18, 1 (2009), 70. DOI: http://dx.doi.org/10.4103/0972-6748.57865

[50] Lore Thaler, Alexander C Schütz, Melvyn A Goodale, and Karl R Gegenfurtner. 2013. What is the best fixation target? The effect of target shape on stability of fixational eye movements. *Vision Research* 76 (2013), 31–42. DOI: http://dx.doi.org/10.1016/j.visres.2012.10.012

[51] Anna Van Cauwenberge, Gabi Schaap, and Rob Van Roy. 2014. "TV no longer commands our full attention": Effects of second-screen viewing and task relevance on cognitive load and learning from news. *Computers in Human Behavior* 38 (2014), 100–109. DOI:http://dx.doi.org/10.1016/j.chb.2014.05.021

[52] Athanasios Vourvopoulos, Evangelos Niforatos, and Michail Giannakos. 2019. EEGlass: An EEG-Eyeware Prototype for Ubiquitous Brain-Computer Interaction. DOI:http://dx.doi.org/10.1145/3341162.3348383

[53] Erroll Wood and Andreas Bulling. 2014. Eyetab: Model-based gaze estimation on unmodified tablet computers. In *Proceedings of the Symposium on Eye Tracking Research and Applications*. ACM, 207–210. DOI:http://dx.doi.org/10.1145/2578153.2578185

[54] David L Woods. 1995. The component structure of the N 1 wave of the human auditory evoked potential. *Electroencephalography and Clinical Neurophysiology-Supplements Only* 44 (1995), 102–109.

[55] Beste F Yuksel, Kurt B Oleson, Lane Harrison, Evan M Peck, Daniel Afergan, Remco Chang, and Robert JK Jacob. 2016. Learn piano with BACh: An adaptive learning interface that adjusts task difficulty based on brain state. In *Proceedings of the 2016 chi conference on human factors in computing systems*. ACM, 5372–5384. DOI:http://dx.doi.org/10.1145/2858036.2858388

[56] Noa Talaván Zanón. 2006. Using subtitles to enhance foreign language learning. *Porta Linguarum: revista internacional de didáctica de las lenguas extranjeras* 6 (2006), 4.

[57] Xinlei Zhang, Nataliya Kosmyna, Pattie Maes, and Jun Rekimoto. 2018. Investigating Bodily Responses to Unknown Words: a Focus on Facial Expressions and EEG. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 5211–5215.