Robust Gaze Features for Enabling Language Proficiency Awareness

Jakob Karolus
University of Stuttgart
Stuttgart, Germany
jakob.karolus@vis.uni-stuttgart.de

Paweł W. Woźniak
University of Stuttgart
Stuttgart, Germany
pawel.wozniak@vis.uni-stuttgart.de

Lewis L. Chuang
Max Planck Institute for Biological Cybernetics
Tübingen, Germany
lewis.chuang@tuebingen.mpg.de

Albrecht Schmidt
University of Stuttgart
Stuttgart, Germany
albrecht.schmidt@vis.uni-stuttgart.de

ABSTRACT

We are often confronted with information interfaces designed in an unfamiliar language, especially in an increasingly globalized world, where the language barrier inhibits interaction with the system. In our work, we explore the design space for building interfaces that can detect the user's language proficiency. Specifically, we look at how a user's gaze properties can be used to detect whether the interface is presented in a language they understand. We report a study (N=21) where participants were presented with questions in multiple languages, whilst being recorded for gaze behavior. We identified fixation and blink durations to be effective indicators of the participants' language proficiencies. Based on these findings, we propose a classification scheme and technical guidelines for enabling language proficiency awareness on information displays using gaze data.

ACM Classification Keywords

H.5.m Information interfaces and presentation (e.g., HCI): Miscellaneous.; H.1.2 Model and Principles: User/Machine Systems—Human information processing

Author Keywords

Eye-tracking; machine learning; language-aware interfaces; adaptive interfaces.

INTRODUCTION

In an increasingly globalized world, users are constantly exposed to unfamiliar cultures and languages. Concurrently, our lives are increasing reliant on digital technology as our environments - in public, at work or at home - are pervaded by digital artifacts such as public displays and shared mobile devices. When visiting foreign countries, especially those that are multilingual, users are often faced with the challenge of navigating an interface in an unfamiliar language. While alternative language versions are often available, accessing them (usually through a poorly visible button or a submenu) tend to be cumbersome. The problem is compounded by screen space limitations. In this paper, we present work that contributes towards a new generation of systems that will detect a user's language proficiencies and can subsequently adapt their interface language to their users' perceived proficiency, as depicted in Figure 1.

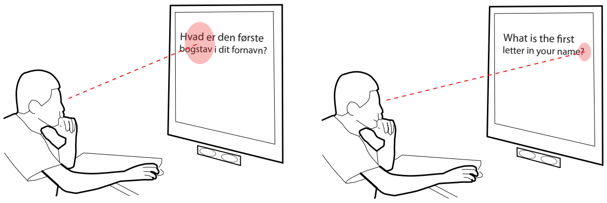
Hvad er den første bogstav i dit fornavn?
What is the first letter in your name?

Figure 1. User interacting with a language-aware interface.

Our daily communications afford us interpersonal cues that allow us to moderate our language and responses to our conversation partners. Telltale facial expressions and body language can indicate the interest and engagement levels of our conversation partners. Gaze can readily indicate whether they are focused on the conversation or distracted by something else, such as looking at their phone or recognizing another person in the background [15, 19]. We are typically able to interpret these implicit signals and adapt our conversation strategy by changing the topic or requesting attention. Here, we explore how user interfaces can use information, such as its user's gaze, to recognize its user's state.

CHI 2017, May 06 - May 11, 2017, Denver, CO, USA

ACM 978-1-4503-4655-9/17/05...$15.00
DOI: http://dx.doi.org/10.1145/3025453.3025601

In this paper, we contribute to the understanding about human gaze properties to build interfaces that recognize the lack of understanding of the user. Specifically, we investigate whether we can effectively detect if the user is confused by the interface language. We examined users' gaze properties (gathered through an embedded eye tracker) when reading short sentences in multiple languages. Afterwards, we evaluated how gaze properties can be interpreted to detect if the interface is presented in a language that the user understands.

Eye tracking is widely used in the domain of reading analysis. However, existing approaches have usually focused on one specific language and are used for post-hoc analysis [31, 50]. We propose an approach that relies on characteristic gaze properties to infer a user's requirement for real-time assistance. The goal is to recognize a user's likelihood for comprehending displayed text, given limited gaze data that is based on reading a short sentence in a given language. We anticipate that users are unlikely to persist in using a system beyond a single sentence, in a language that they lack proficiency in. In this regard, the current work contrasts with eye movement research that is based on reading entire documents. When working with eye gaze interaction, prior calibration often is a hindrance for readily available interaction. We will discuss the aforementioned limitations and analyze their impact on user experience given our approach. To the best of our knowledge, this work represents the first approach to rapid detection of language comprehension for the design of information displays.

In our study with 21 participants we recorded their eye movements while presenting them simple questions in varying languages. The participants were tasked to answer the shown questions to the best of their abilities. Depending on their language proficiency, some questions could be answered easily while others were impossible.

We evaluate the feasibility of language-aware interfaces, based on our experimental results. First, we identified that users exhibited shorter average fixation duration as well as longer average and summed blink duration, when presented with languages that they were proficient in. Based on these results, we developed a classification strategy and provide some technical guidelines to facilitate real-time language proficiency detection on information displays.

This paper is organized as follows. First, we review past efforts in using eye-tracking for assessing language comprehension and data processing methods used to that end. Next, we provide the detail of the experiment we conducted. Following this, we present the analysis of our experimental results, which form the basis of building an automated classifier for rating our users' language comprehension. We discuss the details of the classifier and its alternative designs in the following sections. Finally, we provide technical guidelines and minimum requirements for implementing language comprehension detection using gaze properties on information displays.

RELATED WORK

A user's gaze can be used as an explicit input to computing interfaces, albeit to varying levels of effectiveness. Most prominently, gaze has been used as a substitute or supplement for manual cursor control [45, 40], e.g. in MAGIC-Pointing [51] and typing [29]. Interfaces have also been developed that respond to distinctive and contrived gaze gestures [11] as a form of explicit interaction.

Besides this, gaze can also serve as an implicit input for context-aware applications, such as activity recognition [6] while wearing custom made EOG[1]-glasses [5]. *iTourist* [36] successfully plans city trips based on recorded gaze patterns. Gaze contingent displays [12] use the gaze point of the user to define a region of interest. The information is used to highlight specific elements in the user's view [2] or to selectively render foveated regions at high resolution in order to save computing costs [17, 32].

Other work in the domain of implicit interaction focuses on human-like interaction in virtual environments, e.g. for conversational agents [47, 49]. It has been shown that gaze directional cues can serve as a predictor for conversational attention. Moreover, virtual agents that respond to users' gaze have been shown to increase their users' emotional response and allocated attention [30]. This suggests that gaze-responsive systems are perceived as being more human-like in their interactions and, hence, elicit more user-attention.

Computer systems can also rely on gaze inputs to assist users in accomplishing their task. In tutoring systems, the user's gaze information can be relied on by the system to determine when it is necessary to provide feedback and guidance to the user [10]. For example, patterns in gaze behavior could reveal that the user is confused by a given topic, which prompts the system to provide further guidance [34]. Learning new topics often involves reading provided material, such as a book, a document or even a simple time-table of a bus station. The connection between eye movements and reading has already been researched thoroughly [37, 38]. There exist psychologically plausible models that describe many phenomena in reading [24, 39].

In recent years, researchers built upon these findings to examine language proficiencies [31, 50], as well as develop systems that have educational purposes. These include e.g. real-time annotations [7, 3, 26] or translations during reading [48]. In this scenario the users' eye movements provide contextual information about what they are reading, how fast [28] or even how much they understand [21] and provide assistance accordingly [44, 22]. These systems showcase the capabilities of gaze-assisted language detection and/or translation. A vital part of real-time assistive systems is the exact point in time when to assist, e.g. by displaying a translation. Besides manual activation via gestures or being reliant on user profiles, we explore an implicit method that can be used for real-time interaction.

In our work, we adopt the paradigm of real-time gaze analysis to enable interactive systems that react upon the user's gaze as well as utilize the research findings on reading processes. We aim to combine the real-time properties of gaze-contingent displays with the inherently longer lasting process of recognizing the users understanding based on their reading patterns.

[1] Electrooculography

In contrast to past work on proficiency analysis, we propose a method than only relies on a short sentence in the respective language, hence enabling interactive interfaces. In our study, we evaluate the feasibility and accuracy of language proficiency detection on information displays.

METHOD

For our study design, we considered previous research findings on eye movements and reading to identify gaze characteristics that could be viable candidates for inferring language proficiency in our application context.

Research reports an average duration for fixations of about $200ms$ to $250ms$ and an average saccade length of 7 to 9 letters, but fixation durations can range from $50ms$ to $500ms$ depending on the context of the task and user state. Pertinently, the fixation durations increase with conceptually more difficult text, leading us to hypothesize that fixation durations should vary with a user's language proficiency. In addition, text difficulty is reported to correlate with saccade length as well as the frequency of saccade regressions[2] and refixations. Hence, "difficult" languages should exhibit lower saccade length and a higher frequency[3] of regressions and refixations. [37, 16]

Moreover, blink rate and pupil diameters have been reported to be associated with cognitive load and human information processing. For example, Siegle et al. [46] demonstrated a phasic increase of blinks prior and pursuant to the anticipation of an increase in the load of information processing, which was manipulated with basic psychological tests (i.e., Stroop task, digit-sorting task). Pupil dilation, on the other hand, reflected sustained information processing, over longer periods.

Based on these previous findings, we formulate our first research question:

RQ1: Can we determine whether a user is able to use the interface in a given language from gaze properties available by a gaze tracker?

To implement a real-time and language-aware user interface, it is vital that the system is able to tell a user's proficiency in the display language in a mere seconds. This implies that the users should only be presented with at most a few sentences of an unknown language before the system decides on the respective proficiency. This is the basis for the second research question:

RQ2: Can we determine whether a user understands the current interface language fast enough to build efficient language-aware interfaces? If so, what are the technical requirements?

Based on these questions, we investigate the following hypotheses in our study:

H1: Increased language proficiency level will result in a lower average gaze fixation duration.

H2: Increased language proficiency level will decrease the number of refixations in a given time period.

Lower average gaze fixation duration as well as lower refixation ratio is connected to text difficulty when reading [37]. In our study we aim to vary text difficulty by changing the display language, hence creating several different text difficulty levels. Participants should therefore find it easier to answer questions in their proficient languages.

H3: Increased language proficiency level will increase the average blink duration.

H4: Increased language proficiency level will increase the total blink ratio.

Proficient users of a language should experience lower cognitive load during reading than non-proficient users. We assume that this increases blink duration and total blink ratio when presented with a language in which the user is proficient. During unknown languages cognitive load increases as the user is trying to figure out the question, hence blinks are less frequent.

If we reflect on one primary use case – public displays in commuting areas – the user usually looks for predefined information (e.g. bus departure, flight details). The method of presenting this information is often ordered and predictable, such as a time-table, and does not change across languages. Yet this is of no help for the user if he cannot locate the name of his destination or instructions for the payment process.

Information like this is expressed in simple sentences as well as simple language, to ensure readability. Hence, we chose our questions to be of "Basic User" ([8], p. 23) level with regard to the *Common European Framework of Reference for Languages* (CEFR) [8]. This framework defines six levels of foreign language proficiency: basic (A1, A2), independent (B1, B2) and proficient user (C1, C2) [8]. For our study we added an extra non-proficient level (X).

We collected questions in 13 different languages (15 questions each). By choosing to display simple questions to the participants, we provide an engaging task [20] as reading the text is required to answer correctly. The respective translations were provided by either native or highly proficient users of that language. Table 1 shows a few example questions. Most are part of the Indo-European language family [9], yet we included some outliers such as Finnish and Hungarian as well as languages not using the Latin alphabet (Greek, Arabic) to analyze their effect. See Table 2 for a complete overview on the used languages and the respective proficiency levels exhibited by our participants.

Participants

We recruited twenty-nine participants from the University of Stuttgart and the Stuttgart Media University via mailing lists. The data of 21 participants (12 females, age: 19-36 years) were used for further analysis. The eye-tracking data of three participants were removed because glasses and make-up interfered with the reliability of eye-movement recording. We excluded the data of five more participants, as the recording was not stable enough during the whole study[4]. Two participants had prior experience with eye-tracking studies. Out of the 21 participants, 17 were native German readers, two were native

[2] Saccades opposite to the reading direction

[3] Standard frequency is about 10 to 15%.

[4] Participants started to move more towards the end of the study.

Language	Example questions	
English	How many days are within a week?	What is the first letter of your first name?
French	Combien de jours y a-t-il dans une semaine?	Quelle est la première lettre de votre prénom?
Danish	Hvor mange dage er der i en uge?	Hvad er det første bogstav i dit fornavn?
Finnish	Kuinka monta päivää on viikossa?	Mikä on etunimesi ensimmäinen kirjain?

Table 1. Example questions used in our study in four different languages.

Language	Proficiencies of participants
English	B2, C1, C2
German	X, B2, C1, C2
Danish	X
Dutch	X, A1
Finnish	X
French	X, A1, A2, B1, B2
Greek	X, A1, C1
Romanian	X
Spanish	X, A1, A2, B1, C2
Turkish	X
Slovenian	X
Arabic	X, A1
Hungarian	X

Table 2. Languages used in this study and respective proficiencies present in our participant base.

English readers as well as one native Spanish and Indonesian participant. All participants had normal or corrected-to-normal vision. Each participant was paid an allowance of 10 Euros.

Apparatus

Our setup consisted of a 22 inch LCD display (resolution: 1680x1050) and a remote eye tracker (SMI RED 250; $250 Hz$ sampling frequency) that was positioned below the display. Our participants were seated at a distance that felt comfortable for them and well within the reliable tracking range of the system ($0.5m - 0.7m$) in an enclosed cubicle. Figure 2 shows a picture of the apparatus.

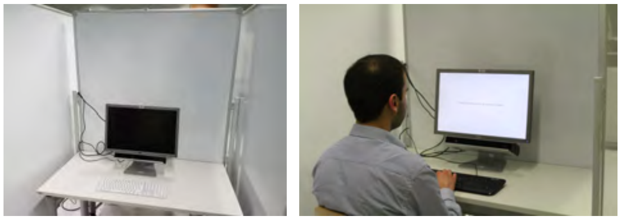

Figure 2. Apparatus showing LCD monitor with attached eye tracking device and participant during the study.

Procedure

After introducing the prospective participants to our study, we handed them a detailed study description. It stated to follow on-screen instructions worded as simple questions, that each require a single key-press to answer. The study description sheet additionally contained an example question. Prospective participants were made aware to expect these questions in different languages and to answer to the best of their knowledge, yet favoring correctness over speed. The respective answer could be given in any preferred language[5]. If a question was impossible to answer, prospective participants were made aware not to press anything and wait for the next question. Timings between each question (10 seconds) and the overall study time (two session of 12.5 minutes each) were provided on the information sheet. The available answer time of 10 seconds per question was conservatively chosen, preventing unnecessary pressure due to lack of time. After providing informed consent, the participants were asked to complete a demographic questionnaire, polling sex, age, work field, highest educational qualification, their native language and their eyesight. They also rated their reading level for specific languages based on the provided CEFR's self-assessment test[6] ([8], pp. 26-27). Additionally, the participants were asked to provide their proficiency level for languages that were not listed. The experimenter was present to answer any questions that might arise.

Before starting the experiment, the eye tracking device was calibrated using a five-point calibration. We only accepted calibration accuracy below one degree of visual angle. During the experiment, we sequentially displayed 150 questions to the participants, in random order. Each question was visible for exactly ten seconds and could be answered by a single key press, such as one letter or one number (see Table 1). Participant keystrokes were collected during this period. The experiment was conducted in two sessions of 12.5 minutes with an intervening rest period. It was possible to ask questions during the experiment, which occurred two times. The respective experiment question was hence marked for deletion. Before resuming with session two, another calibration analog to the start of the experiment was performed. Ethical approval for this study was obtained from the Ethics Committee at the University of Konstanz.

Post-processing

We applied the following post processing steps to the obtained eye-tracking data including event detection[7] using a velocity-based fixation algorithm [41] with a velocity threshold of 35 degrees of visual angle per second and blink detection based on pupil diameter change. Eye movement events were allocated to the respective question and any data after the point in time when the participants provided an answer for the respective question, as indicated by the recorded keystrokes, was discarded. Note that we did not distinguish between wrong

[5] E.g. answering in one's native language or in English.

[6] If needed. Most participants were familiar with the framework and provided information based on language tests.

[7] utilizing the SMI Event Detector

and correct answers, since language proficiency as stated by the participants was used as ground truth. We observed a mean answer time for all participants of 4.9 seconds with a standard deviation of 0.2 seconds. Thus, we conservatively limited the overall observation time for each question to a maximum of 4.5 seconds. Additionally, we discarded fixations with durations, less than $50ms$ and more than $600ms$, that were outside the range of reported values in reading research [37, 38, 43].

Measures

All measures were collected on a per question basis. We derived the following metrics from the eye tracking data. For fixations, we examined the average fixation duration. We also recorded refixations that occurred when the user's gaze revisited the location of a previous fixation. This area was bounded by a 30 pixel radius[8] around the previous fixation's location. The chosen radius is a conservative interpretation of the distance of consecutive fixations during reading [37]. The amount of refixations was normalized by the total fixation count for the respective question (refixation ratio).

For blinks, we calculated average duration as well. Total blink ratio relates the total blink duration to the respective answer time of each question.

All the measures were grouped by the reported language proficiency based on the CEFR [8]. Hence, we end up with seven groups: non-proficient (X), basic user (A1, A2), independent user (B1, B2) and proficient user (C1, C2).

Results

As mentioned before, we looked at data available after having interacted with the screen for a maximum of 4.5 seconds or the user's respective answer time, whichever was shorter. Sample size did vary for different metrics, e.g. if the user did not blink during one particular question.

Average Fixation Duration (AFD)

The grand mean of AFD was $239.86ms$ ($SD = 53.22ms$). Participants with C2 proficiency in the tested language exhibited the shortest average fixation durations ($M = 194.02ms$, $SD = 37.24ms$) while no knowledge of a given language produced the longest fixations ($M = 249.19ms$, $SD = 53.04ms$). As all metrics in our experiment are unevenly distributed in terms of proficiencies (participants have varying levels of proficiency in different languages), we decided to compute the slope coefficient for each of the participants using ordinal logistic regression. We then conducted a t-test comparing the obtained slope coefficients with a constant function. The test shows that language proficiency has a significant effect on AFD, $t(20) = -3.78$, $p < 0.01$.

Refixation Ratio (RFR)

The grand mean of RFR was 0.24 ($SD = 0.15$). Participants with C2 proficiency in the tested language exhibited the lowest refixation ratio ($M = 0.15$, $SD = 0.13$) while participants who did not know the given language had the highest ratio ($M = 0.26$, $SD = 0.15$). Analogously to AFD, slope coefficients were computed and compared with $y = 0$ using a t-test. The test showed that language proficiency had a significant effect on RFR, $t(20) = -2.48$, $p < 0.05$.

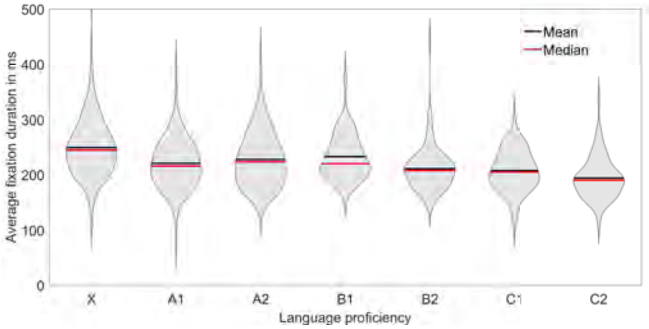

Figure 3. Violin plot showing the distribution of the average fixation duration (AFD) grouped by proficiency according to CEFR [8].

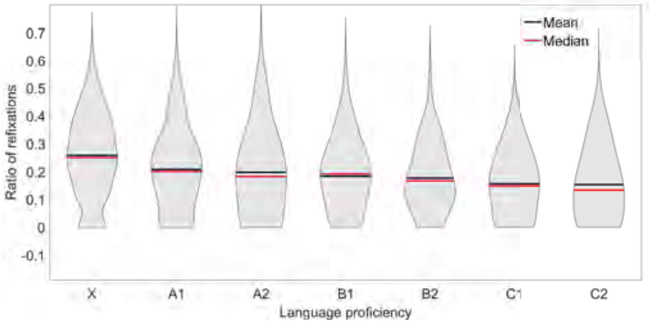

Figure 4. Violin plot showing the distribution of refixation count per overall fixation count (RFR) grouped by proficiency according to CEFR [8].

Average Blink Duration (ABD)

The grand mean of ABD was $413.47ms$ ($SD = 395.52ms$). Participants with C2 proficiency in the tested language exhibited the longest average blink time ($M = 752.05ms$, $SD = 391.00ms$) while participants who did not know the given language blinked the shortest ($M = 306.41ms$, $SD = 332.29ms$). Again, we calculated slope coefficients for each participant and used a t-test to compare with no slope yielding a result of $t(20) = 2.71$, $p < 0.05$. Figure 5 shows the distribution of average blink durations.

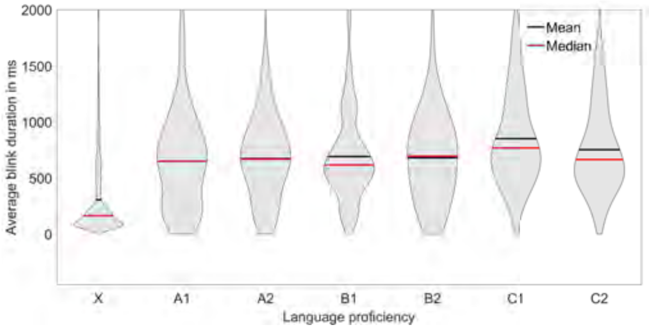

Figure 5. Violin plot showing the distribution of the average blink duration (ABD) grouped by proficiency according to CEFR [8].

[8] Covers roughly 1-2 letters.

Total Blink Ratio (TBR)

The grand mean of TBR was 0.15 ($SD = 0.14$). Participants with C2 proficiency in the tested language exhibited the highest total blink ratio ($M = 0.30$, $SD = 0.14$) while participants who did not know the given language had the lowest ratio ($M = 0.11$, $SD = 0.11$). Using a method analogous to the previous metrics, we obtained $t(20) = 2.75$, $p < 0.05$. Figure 6 shows the distribution of total blink ratio.

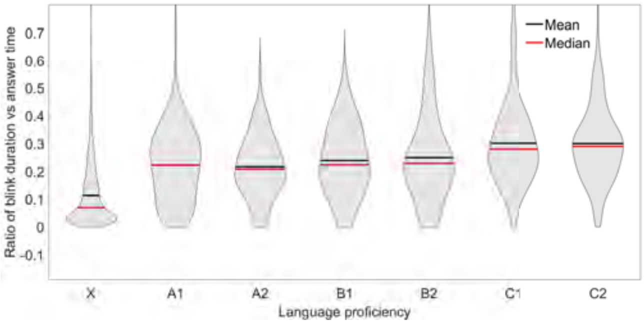

Figure 6. Violin plot showing the distribution of the ratio given total blink duration and answer time (TBR) grouped by proficiency according to CEFR [8].

Result Interpretation and Implications

Statistics on average fixation duration and refixation ratio showed a significant effect of language proficiency on all metrics. We conclude that increased language proficiency levels will result in a lower AFD and a lower RFR (as t-scores are negative). This confirms **H1** and **H2**.

Our results indicate that average blink duration and total blink ratio were significant metrics as well, confirming **H3** and **H4**. However, standard deviation was also quite high for these metrics, which might lead to unstable classification results. Especially for proficient levels (A to C), blink duration is quite high compared to values found in literature [20]. This could be due to participants "relaxing" during easy questions, which relates to longer blink durations [20]. Nevertheless, simple blink detection (e.g. via a RGB camera) could be an alternative for systems where eye tracking is not feasible.

Our statistical analysis indicates that we can provide a positive answer to **RQ1**. All metrics showed a significant difference for non-proficient vs proficient users (except the B1 level), indicating that our approach is feasible. In other words, pur experiment shows that the eye gaze metrics we chose do carry information about language comprehension.

The second research question (**RQ2**) concerned the technical requirements of applying such an approach to language-aware interfaces. In the next section, we pursue **RQ2** further by implementing classifiers for language proficiency detection.

BUILDING A LANGUAGE-AWARE SYSTEM

In the previous sections, we evaluated the statistic relevance of certain metrics extracted from our eye-tracking data. AFD, ABD, TBR and RFR showed promising results in discerning proficiency levels. Yet, to realize systems that support real-time detection of language proficiency it is necessary to find a discriminative feature set that defines certain proficiency levels. Ideally, we want the margin between neighboring levels as large as possible and choose these levels based on the application scenario. In this section, we provide answers to our second research question and evaluate the needed technical requirements for language-aware interfaces.

We highlight specific factors that need to be considered when implementing such an adaptive display. From an HCI perspective, we want to maximize the user experience when interacting with the system such as providing assistance when needed and only when needed. Furthermore, an interactive system should be able to provide feedback in a reasonable time span. Ideally, it should not take longer than a few seconds for the system to make a decision. From a technical standpoint, we might need to consider limitations that may arise with the application scenario, such as low-cost hardware that will result in noisy data. As such we tested our findings on an artificially downsampled dataset.

Baseline Classifier

Based on our previous results, we decided to train a binary classifier that predicts whether people are not proficient or proficient (levels A to C) in the displayed languages. This choice conforms with our aspired application scenario in two ways. Firstly, public displays such as timetables exhibits "predictable information" [8] that can be read by a basic language user (A level proficiency). Hence, the choice to include all proficiency levels into one class seems reasonable. Secondly, overeager assistance systems tend to be rejected by the users. To achieve a good user experience, minimizing false proficiency classification on proficient users should therefore be a primary goal.

We constructed a dataset containing the following features from our eye-tracking data as described in the previous sections: average fixation duration (AFD), refixation ratio (RFR), average blink duration (ABD) and total blink ratio (TBR). We used the same maximum observation time as before (4.5 seconds). After balancing the class distribution, we evaluated the dataset using 10-fold cross validation on four different common classifiers[9]. All classifiers were executed with standard parameters. An overview on achieved accuracy is given in Table 3.

	J48	SVM	NN	BayesNet
Accuracy	78.3%	64.9%	78.8%	77.6%
F-Measure	78.3%	64.3%	78.8%	77.6%

Table 3. Accuracies and F-Measures for selected classifiers.

While most classifiers exhibited a similar average accuracy, it is vital in our case to report the false positive rate (FPR) of each class. As mentioned before, our primary goal was to keep the ratio of falsely classifying proficient users (classified as non-proficient) to a minimum, while our secondary goal was

[9] J48: Java derivate of C4.5 [35]
SVM: LibLINEAR package. L2-loss (dual form) [13]
Neural network: MultiLayer Perceptron [18]
BayesNet: Bayesian network [18]

to maximize true positive rate (TPR) of either class. Table 4 shows more detailed descriptive statistics on the decision tree J48, including false and true positive rate for both classes. This allows us to spot irregularities that may arise between classification accuracy of the two classes.

Class	TPR	FPR	Precision	Recall
NP	75.4%	18.8%	80.0%	75.4%
P	81.2%	24.6%	76.8%	81.2%

Table 4. Classifier statistics on J48 given proficient users (class P) and non-proficient users (class NP).

True positive rate was higher for the class of proficient language users (class P), yet this came with a higher false positive rate as well. In other words, in 24.6% of cases we wrongly classify a non-proficient user (class NP) as being proficient. Of course this is inconvenient for the user as he will not be provided with the needed assistance. However, our primary goal was to keep the false positive rate of the class of non-proficient users to a minimum. In the standard configuration J48 misclassified 18.8% of all instances as non-proficient when they were in fact instances of proficient users. Hence, the system would attempt to provide assistance although the user would be able to read the given language, which is undesirable and has a negative impact on user experience.

Figure 7 illustrates the top three levels of the resulting tree model[10]. Average blink duration is used as a first separation step. After just two tests (on ABD and AFD) almost 32%[11] of instances are correctly assigned to the non-proficient class (3.6% are wrongly classified in this step). The built tree exhibits several of these "heavy" leaves that carry most of the instances. This indicates that pruning may be a meaningful strategy.

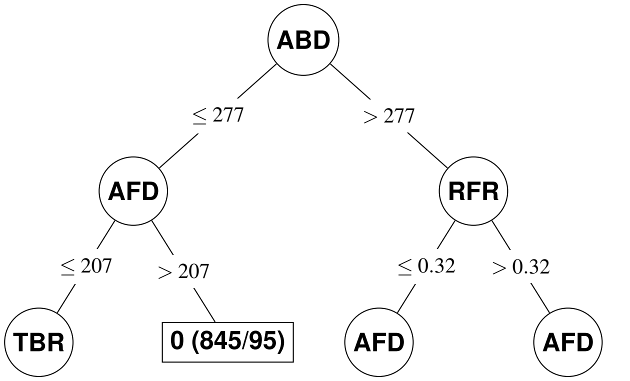

Figure 7. Top three levels of the built decision tree using J48 with standard parameters. ABD and AFD are given in milliseconds; TBR and RFR are given as ratios. Leaf nodes (rectangles) indicate predicted class (0 for NP; 1 for P) and show (correctly classified/misclassified) instances. Values are rounded for visual clarity.

A more heavily pruned tree is shown in Figure 8, basing its decision on only three remaining features (ABD, RFR and AFD), while still achieving 77.64% accuracy. Additionally, true and false positive rate of each class do not change by more than two percentage points, indicating that heavy pruning is indeed a valid option. This avoids overfitting to the training set and may generalize better for unseen data.

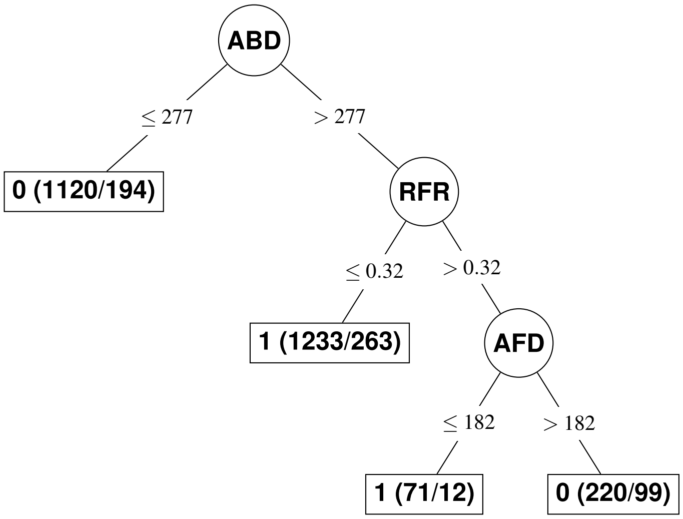

Figure 8. Complete decision tree using J48 and heavy pruning. ABD and AFD are given in milliseconds; RFR is given as ratio. Leaf nodes (rectangles) indicate predicted class (0 for NP; 1 for P) and show (correctly classified/misclassified) instances. Values are rounded for visual clarity.

Cost-Sensitive Classifier

To further reduce the false positive rate of the non-proficient class, we introduced a different cost function for the classifier. We penalized classifying proficient users as non-proficient more severely. This is a compromise between maximizing overall classification accuracy (results shown in Table 3 and Table 4) and minimizing the false positive rate of one class. We decided that not more than one out of ten users should be wrongly classified as non-proficient. To achieve this goal, misclassifying a proficient user had to be three times more expensive than misclassifying a non-proficient user. Table 5 shows the results after applying such a function.

Class	TPR	FPR	Precision	Recall
NP	58.9%	9.90%	85.6%	58.9%
P	90.1%	41.1%	77.1%	74.5%

Table 5. Classifier statistics on cost-sensitive J48 (standard parameters) given proficient users (class P) and non-proficient users (class NP).

Overall accuracy dropped only slightly to 74.5%, yet the gap between the true positive rate of both classes increased. While TPR of the proficient class increased[12], the TPR of the non-proficient class decreased to 58.9%. It was now more expensive to classify as non-proficient, hence it was safer to classify as proficient. The same applied for the false positive rates. We successfully pushed the FPR of the non-proficient class to below 10%, yet this also meant that the FPR of the proficient class increased and fewer users were to get the needed assistance.

[10] using J48 and standard parameters

[11] 845 out of a total of 2643 instances

[12] It was more likely to classify an instance as proficient.

Class Distribution

Generally, it is important to adjust the cost function based on the scenario and think about how to penalize wrong classifications. Besides the cost function, the distribution of instances into respective classes is vital. In a different scenario, such as displaying advertisements on public displays, it is more favorable to emphasize highly proficient users. Since using puns in advertisements to spark the interest of potential customers is a common technique, aiming for a binary classifier that separates C-level language users from other users is more beneficial. If the users do not understand the display pun due to their proficiency level being too low, it seems reasonable to combine them with the non-proficient users. Table 6 shows classification results for a grouping into the classes highly-proficient HP (C1 and C2) and less-proficient LP (X, A1, A2, B1 and B2). Overall accuracy was 79.3%. The false positive rate for both classes was around 21%. Again, we can ask ourselves whether misclassifying one class is more severe than the other and adjust a cost function as we see fit.

Class	TPR	FPR	Precision	Recall
LP	79.2%	20.6%	79.3%	79.2%
HP	79.4%	20.8%	79.2%	79.4%

Table 6. Classifier statistics on J48 with different class distribution: highly-proficient (class HP: C1 and C2) and less-proficient (class LP: X, A1, A2, B1 and B2) users. Levels according to [8].

Recording Duration

As outlined previously, fast responsiveness and interactivity are vital for an adaptive information display. Hence, we evaluated a minimum recording duration that allowed for a sensible classification of non-proficient users. For this purpose, we limited the maximum observation time for our dataset in one second intervals from one to five seconds[13]. Figure 9 shows accuracy as well as true and false positive rate over time.

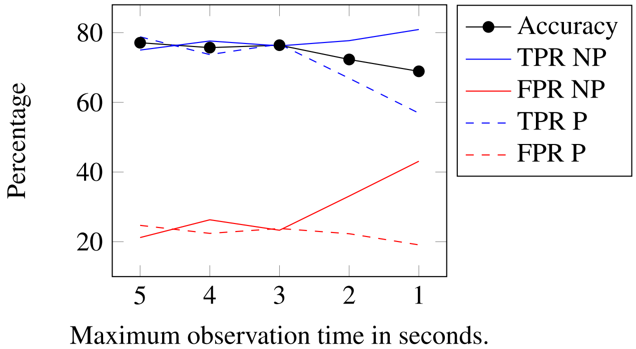

Figure 9. Accuracy, true and false positive rate of proficient (class P: A1 to C2) and non-proficient (class NP: X) class given different maximum observation times of recorded gaze data.

The classifier metrics did not change much when reducing the maximum observation time to only three seconds. After that, we saw a strong increase in the false positive rate of class NP. Further investigation revealed that the classes exhibit a strong overlap, thus making accurate classification difficult. We believe that this is due to the reduced sample size, hence allowing noise in the recording to have a larger influence.

Hardware Limitations

In this study, we used a high-quality eye-tracking device with a sampling rate of $250Hz$. However, eye-trackers for commercial product integration can be expected to have lower sampling rates. To evaluate the viability of our approach for its stated purpose, we downsampled[14] our recording data to the following sampling rates $125Hz$, $62.5Hz$ and $31.25Hz$, which approximated the sampling rates of low-cost eye-tracking devices.

Reducing sampling rate introduces noise in the event (e.g. fixation, saccades) detection of eye-movement behavior. Here, we showed the effects for the average fixation duration. A classification accuracy of 64.7% was achieved from the original sampled data (i.e., $250Hz$). Figure 10 illustrates how classification accuracy, as well as true and false positive rate for each class, varied when decreasing the sampling rate.

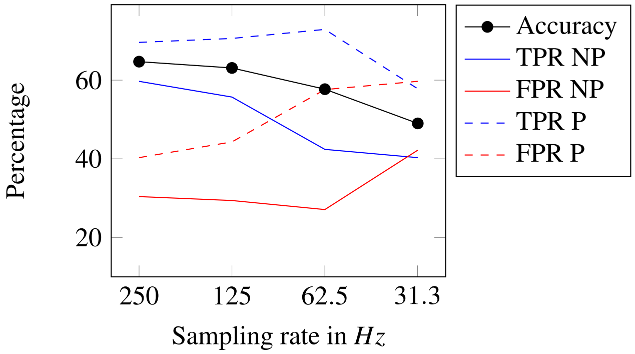

Figure 10. Accuracy, true and false positive rate of proficient (class P: A1 to C2) and non-proficient (class NP: X) class given different sampling rates of the underlying gaze dataset (logarithmic scale).

Sampling at a rate of $125Hz$ had only minor influence on the classification results. Further downsampling drastically reduced true positive rate of class NP (FPR of class P increases), yet the false positive rate of the non-proficient class stayed constant. Hence, at a sampling rate of $62.5Hz$ detecting a non-proficient speaker correctly was less likely. For the lowest sampling rate $31.25Hz$, FPR for the non-proficient class strongly increased. At this rate, a sample was obtained every 32ms, which was about the mean difference that we found between the two classes (NP and P) at high sampling rates. Thus, temporal downsampling rendered information between categories less discriminable and sensible classification was not possible at this resolution.

DISCUSSION

Through the initial statistical analysis and our investigation of how to build classifiers using our data set, we have shown that rapid automated detection of language comprehension is possible. The data we extracted from the user's gaze enables

[13] Before then our dataset used a maximum observation time of 4.5 seconds

[14] By omitting every n-th sample.

us to reliably predict their proficiency in a given language. It needs to be noted that our proposed classifiers operate on a coarse level. Their primary purpose is to determine whether the user is able to understand the interface presented in a given language. More sophisticated methods and larger gaze pattern samples are still required to estimate exact proficiency levels, as shown in previous work.

We can conclude that the answer to **RQ2** is positive — it is possible to build a reliable automated system that will promptly detect whether a user is able to comprehend the content presented in a given interface. An emergent question is how these systems can be used in practice and what technical prerequisites are needed for deployment. Below, we further elaborate on **RQ2** by providing a set of four technical guidelines for implementing language-aware systems that use gaze data.

Guidelines for Language-Aware Interfaces

Based on our classification analysis, we propose the following technical guidelines for interfaces that are to be designed to be aware of the user's language proficiencies.

Firstly, we recommend that designers of future systems **adapt the cost function of the classifier as well as the class distribution to their particular application**. The application scenario highly influences classification accuracy, as we have shown when grouping different proficiency levels for binary classification. Fine-grained detection down to the proficiency levels of the CEFR [8] did not deliver satisfying accuracy. Cost functions can be introduced to better fit the use case scenario. Here, a compromise has to be made between a more eager assistance system and a more rigorous one. Depending on the application scenario, one might opt for either direction. Besides adjusting the cost function, changing the class distribution is also a possibility, e.g. when considering advertisement banners instead of public displays in an airport. Hence, determining whether a user is able to use an interface in a given language from gaze properties is entirely possible. Achieved accuracy depends on the application scenario and the selected proficiency levels. If fine-grained detection and high accuracy is needed, e.g. in an language education software that observes the learning process of a student, other sources of context should be used in conjunction, while eye tracking can provide an auxiliary input.

Secondly, we suggest **a minimum interaction period of three seconds and a sampling rate not less than 100 Hz**, while lower rates might still be applicable for specific scenarios. In our study, we evaluated how long and how often a language-aware interface should record eye movements of interacting users. Higher resolution or longer recording duration only slightly improve classification results. These findings confirm the feasibility for language-aware interfaces in public displays, as "calibration time" is short enough for user interaction and sampling rate is not too high for low-cost eye-tracking devices.

Especially for public displays or similar interfaces, we recommend to **ensure that light sources do not interfere with the recording**. Locations of public displays are diverse in lighting and environmental conditions. Bad lighting can lead to artifacts during gaze recording and hinder classification. This hardly poses a problem in buildings or underground, yet displays positioned outside can be problematic due to the infrared radiation from the sun. Gaze recording that does not rely on infrared video-based eye trackers can solve this problem, if the technique is accurate enough. In combination with regression-based models [4], it is possible to explicitly learn for distinct gaze characteristics such as fixations, circumventing the need for accurate gaze data in the first place.

Finally, we remind designers to **guide the user into a suitable interaction range**. Reliable tracking is possible only within a rather limited sweet spot. Solutions to solve this problem have already been shown in *GazeHorizon* [52] and *GravitySpot* [1]. If eye-tracking is to be used in a public display scenario, additional challenges need to be considered [33].

Limitations

The scope of the current work addresses how gaze features are discriminable for language proficiency. An adaptive language-aware interface should consider additional probabilistic priors. In this regard, we feel that modifications to our current classification approach (e.g., by adding a "location prior"[15]) holds promise for subsequent work. However, we verified that proficiency in a given language could not generalize to proficiency in related languages[16]. Hence, further research has to be conducted to evaluate different language-switching mechanisms with regard to user experience and accuracy. Design guidelines can serve to constrain the diversity of interaction behavior, increase the predictability of desired behavior, which will render interactions more predictable. If gaze behavior is predictable based on screen content, e.g. predominantly horizontal saccades when reading text vs a search pattern when viewing pictures, the constraints on calibration accuracy are less severe and interaction might be possible without prior per-user calibration. Detecting blinks and fixations – as used in our metrics – is feasible using only relative gaze data.

In our study, we used a set of languages that reflected the possible linguistic abilities of our participant base. We mostly selected languages that are spoken in Europe. However, we believe that as long as reading direction and alphabet are kept the same, the results of our study are reproducible with other languages. Even reading direction might have only a minor influence [37]. Statistics on whether the alphabet was a confounding factor (e.g. Arabic, Greek) have shown that there was no significant difference on the four metrics discussed previously. Nevertheless, this was a subjective result based on the subset of our used languages.

Another limiting factor was the uneven distribution of proficiency levels. Naturally, we had a lot of data samples for non-proficient users, but missed out on proficient levels. B1-level proficiency was especially rare among the participants.

Based on the application scenario, participants might react differently. For example, when on a stressful trip figuring out

[15] E.g. emphasizing predominant languages at the system's deploy location.

[16] By comparing metrics for German C2 speakers between German and Dutch.

their next connection on an information display versus reading questions during a calm lab study. This will likely introduce another noise factor that influences the resulting data. The nature of this noise and its impact in an in-the-wild system is to be evaluated. Consequently, before deployment, extensive pre-testing needs to be conducted for a particular interface to determine what comprehension levels are required in a given interaction scenario.

CONCLUSION AND FUTURE WORK

In this paper, we examined the perception of different languages and their influence on a person's eye movements. Contrary to common approaches in reading comprehension, which focused on post-hoc evaluations, we evaluated the feasibility of real-time language proficiency detection suitable for language-aware interfaces.

In our study, we presented 150 questions in different languages to our participants and analyzed the effect of language proficiency on certain gaze characteristics. We found that proficiency had a significant effect on the average fixation duration and refixation ratio as well as average blink duration and total blink ratio measured over a maximum timespan of 4.5 seconds.

To effectively utilize these results in real-time adaptive information displays, we proposed a classification scheme that realizes language proficiency detection and recommended technical guidelines on recording duration and sampling rate of the eye-tracking device. Moreover, if gaze behavior is predictable based on screen content, constraints on calibration accuracy are less severe. Hence, it is well suited to be used in public displays where "walk-up" interaction is required [33]. Additionally, it enables the usage of less accurate and cheap eye tracking devices as exact gaze position is not necessary. We contributed a set of four technical guidelines for deploying gaze-based language-aware systems. We hope that these guidelines will help designers build real-life systems that detect when the user does not understand the interface language.

Further work that can benefit from our results includes techniques for language switching. New approaches can attempt providing a broader selection of initial languages, i.e. displaying multiple languages at once or traversing a language tree based on probabilistic "location priors" that allows fine-grained proficiency classification. The user's initial language choice – given by the first gaze position – and the reading direction enables alphabet detection. Yet, utilizing exact gaze position would require a per-user calibrated eye-tracking device. Here, new methods on implicit calibration relying on predicted gaze targets [25] or smooth pursuit as shown in *TextPursuits* [27] are applicable. If access to fully calibrated gaze data is available, more elaborate comparing methods are possible, such as directly comparing scanpaths [14, 23] or clustering approaches [42]. To foster research in this direction, the data set is available to the research community for further analysis and improvement on our institute's homepage[17].

We envision that our idea of real-time adaptive information displays is not only applicable to language awareness, but also holds true for higher level concepts, such as different forms of graph visualizations, schematics or maps. This does imply that such a system recognizes a user's level of understanding and it is able to switch to an appropriate visualization to assist the user. We hope that our work will inspire further developments that will yield more understanding on how gaze properties can be used to build more efficient adaptive interfaces.

[17] http://www.hcilab.org/publications/

ACKNOWLEDGEMENTS

We thank all colleagues and friends that provided us with the necessary language translations. This research is financially supported by the German Research Foundation (DFG) within the projects C02, C03 and task force TF3 of SFB/Transregio 161 and supported by the SimTech Cluster of Excellence. Part of this research was supported by the European Union's Horizon 2020 Programme under ERCEA grant no. 683008 AMPLIFY.

REFERENCES

1. Florian Alt, Andreas Bulling, Gino Gravanis, and Daniel Buschek. 2015. GravitySpot: Guiding Users in Front of Public Displays Using On-Screen Visual Cues. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology (UIST '15)*. ACM, New York, NY, USA, 47–56. DOI: http://dx.doi.org/10.1145/2807442.2807490

2. Florian Alt, Alireza Sahami Shirazi, Albrecht Schmidt, and Julian Mennenöh. 2012. Increasing the User's Attention on the Web: Using Implicit Interaction Based on Gaze Behavior to Tailor Content. In *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design (NordiCHI '12)*. ACM, New York, NY, USA, 544–553. DOI: http://dx.doi.org/10.1145/2399016.2399099

3. Ralf Biedert, Georg Buscher, Sven Schwarz, Jörn Hees, and Andreas Dengel. 2010. Text 2.0. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems (CHI EA '10)*. ACM, New York, NY, USA, 4003–4008. DOI: http://dx.doi.org/10.1145/1753846.1754093

4. Björn Browatzki, Heinrich H. Bülthoff, and Lewis L. Chuang. 2014. A Comparison of Geometric- and Regression-Based Mobile Gaze-Tracking. *Frontiers in Human Neuroscience* 8 (April 2014). DOI: http://dx.doi.org/10.3389/fnhum.2014.00200

5. Andreas Bulling, Daniel Roggen, and Gerhard Tröster. 2008. It's in Your Eyes: Towards Context-Awareness and Mobile HCI Using Wearable EOG Goggles. In *Proceedings of the 10th International Conference on Ubiquitous Computing (UbiComp '08)*. ACM, New York, NY, USA, 84–93. DOI: http://dx.doi.org/10.1145/1409635.1409647

6. Andreas Bulling, Jamie A. Ward, Hans Gellersen, and Gerhard Tröster. 2009. Eye Movement Analysis for Activity Recognition. In *Proceedings of the 11th International Conference on Ubiquitous Computing*

(UbiComp '09). ACM, New York, NY, USA, 41–50. DOI: http://dx.doi.org/10.1145/1620545.1620552

7. Shiwei Cheng, Zhiqiang Sun, Lingyun Sun, Kirsten Yee, and Anind K. Dey. 2015. Gaze-Based Annotations for Reading Comprehension. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 1569–1572. DOI: http://dx.doi.org/10.1145/2702123.2702271

8. Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press.

9. Guy Deutscher. 2006. *The Unfolding of Language: An Evolutionary Tour of Mankind's Greatest Invention*. Henry Holt and Company.

10. Sidney D'Mello, Andrew Olney, Claire Williams, and Patrick Hays. 2012. Gaze Tutor: A Gaze-Reactive Intelligent Tutoring System. *International Journal of Human-Computer Studies* 70, 5 (May 2012), 377–398. DOI:http://dx.doi.org/10.1016/j.ijhcs.2012.01.004

11. Heiko Drewes and Albrecht Schmidt. 2007. Interacting with the Computer Using Gaze Gestures. In *Proceedings of the 11th IFIP TC 13 International Conference on Human-Computer Interaction - Volume Part II (INTERACT'07)*. Springer-Verlag, Berlin, Heidelberg, 475–488. DOI: http://dx.doi.org/10.1007/978-3-540-74800-7_43

12. Andrew T. Duchowski, Nathan Cournia, and Hunter Murphy. 2004. Gaze-Contingent Displays: A Review. *Cyberpsychology & Behavior: The Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society* 7, 6 (Dec. 2004), 621–634. DOI: http://dx.doi.org/10.1089/cpb.2004.7.621

13. Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR - A Library for Large Linear Classification. (2008). http://www.csie.ntu.edu.tw/~cjlin/liblinear/ The Weka classifier works with version 1.33 of LIBLINEAR.

14. Matt Feusner and Brian Lukoff. 2008. Testing for Statistically Significant Differences Between Groups of Scan Patterns. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications (ETRA '08)*. ACM, 43–46. DOI:http://dx.doi.org/10.1145/1344471.1344481

15. Tom Foulsham, Joey T. Cheng, Jessica L. Tracy, Joseph Henrich, and Alan Kingstone. 2010. Gaze Allocation in a Dynamic Situation: Effects of Social Status and Speaking. *Cognition* 117, 3 (Dec. 2010), 319–331. DOI: http://dx.doi.org/10.1016/j.cognition.2010.09.003

16. Joseph H Goldberg and Xerxes P Kotval. 1999. Computer Interface Evaluation Using Eye Movements: Methods and Constructs. *International Journal of Industrial Ergonomics* 24, 6 (Oct. 1999), 631–645. DOI: http://dx.doi.org/10.1016/S0169-8141(98)00068-7

17. Brian Guenter, Mark Finch, Steven Drucker, Desney Tan, and John Snyder. 2012. Foveated 3D Graphics. *ACM Trans. Graph.* 31, 6 (Nov. 2012), 164:1–164:10. DOI: http://dx.doi.org/10.1145/2366145.2366183

18. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.* 11, 1 (Nov. 2009), 10–18. DOI: http://dx.doi.org/10.1145/1656274.1656278

19. Lotta Hirvenkari, Johanna Ruusuvuori, Veli-Matti Saarinen, Maari Kivioja, Anssi Peräkylä, and Riitta Hari. 2013. Influence of Turn-Taking in a Two-Person Conversation on the Gaze of a Viewer. *PLoS ONE* 8, 8 (Aug. 2013). DOI: http://dx.doi.org/10.1371/journal.pone.0071569

20. Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost van de Weijer. 2011. *Eye Tracking: A Comprehensive Guide to Methods and Measures*. OUP Oxford.

21. Aulikki Hyrskykari, Päivi Majaranta, Antti Aaltonen, and Kari-Jouko Räihä. 2000. Design Issues of iDICT: A Gaze-Assisted Translation Aid. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications (ETRA '00)*. ACM, New York, NY, USA, 9–14. DOI:http://dx.doi.org/10.1145/355017.355019

22. Aulikki Hyrskykari, Päivi Majaranta, and Kari-Jouko Räihä. 2003. Proactive Response to Eye Movements. In *INTERACT*, Vol. 3. IOS press Amsterdam, 129–136.

23. Halszka Jarodzka, Kenneth Holmqvist, and Marcus Nyström. 2010. A Vector-based, Multidimensional Scanpath Similarity Measure. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications (ETRA '10)*. ACM, 211–218. DOI: http://dx.doi.org/10.1145/1743666.1743718

24. Marcel A. Just and Patricia A. Carpenter. 1980. A Theory of Reading: From Eye Fixations to Comprehension. *Psychological review* 87, 4 (1980), 329.

25. Pawel Kasprowski and Katarzyna Harezlak. 2016. Implicit Calibration Using Predicted Gaze Targets. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications (ETRA '16)*. ACM, 245–248. DOI: http://dx.doi.org/10.1145/2857491.2857511

26. Dagmar Kern, Paul Marshall, and Albrecht Schmidt. 2010. Gazemarks: Gaze-Based Visual Placeholders to Ease Attention Switching. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 2093–2102. DOI: http://dx.doi.org/10.1145/1753326.1753646

27. Mohamed Khamis, Ozan Saltuk, Alina Hang, Katharina Stolz, Andreas Bulling, and Florian Alt. 2016. TextPursuits: Using Text for Pursuits-Based Interaction and Calibration on Public Displays. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. ACM, New York, NY, USA, 274–285. DOI: http://dx.doi.org/10.1145/2971648.2971679

28. Kai Kunze, Katsutoshi Masai, Masahiko Inami, Ömer Sacakli, Marcus Liwicki, Andreas Dengel, Shoya Ishimaru, and Koichi Kise. 2015. Quantifying Reading Habits: Counting How Many Words You Read. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. ACM, New York, NY, USA, 87–96. DOI: http://dx.doi.org/10.1145/2750858.2804278

29. Päivi Majaranta and Kari-Jouko Räihä. 2002. Twenty Years of Eye Typing: Systems and Design Issues. In *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications (ETRA '02)*. ACM, New York, NY, USA, 15–22. DOI: http://dx.doi.org/10.1145/507072.507076

30. Linda Marschner, Sebastian Pannasch, Johannes Schulz, and Sven-Thomas Graupner. 2015. Social Communication with Virtual Agents: The Effects of Body and Gaze Direction on Attention and Emotional Responding in Human Observers. *International Journal of Psychophysiology* 97, 2 (Aug. 2015), 85–92. DOI: http://dx.doi.org/10.1016/j.ijpsycho.2015.05.007

31. Pascual Martínez-Gómez and Akiko Aizawa. 2014. Recognition of Understanding Level and Language Skill Using Measurements of Reading Behavior. In *Proceedings of the 19th International Conference on Intelligent User Interfaces (IUI '14)*. ACM, New York, NY, USA, 95–104. DOI: http://dx.doi.org/10.1145/2557500.2557546

32. Michael Mauderer, David R. Flatla, and Miguel A. Nacenta. 2016. Gaze-Contingent Manipulation of Color Perception. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 5191–5202. DOI: http://dx.doi.org/10.1145/2858036.2858320

33. Jörg Müller, Florian Alt, Daniel Michelis, and Albrecht Schmidt. 2010. Requirements and Design Space for Interactive Public Displays. In *Proceedings of the 18th ACM International Conference on Multimedia (MM '10)*. ACM, New York, NY, USA, 1285–1294. DOI: http://dx.doi.org/10.1145/1873951.1874203

34. Ayano Okoso, Takumi Toyama, Kai Kunze, Joachim Folz, Marcus Liwicki, and Koichi Kise. 2015. Towards Extraction of Subjective Reading Incomprehension: Analysis of Eye Gaze Features. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '15)*. ACM, New York, NY, USA, 1325–1330. DOI: http://dx.doi.org/10.1145/2702613.2732896

35. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

36. Pernilla Qvarfordt and Shumin Zhai. 2005. Conversing with the User Based on Eye-Gaze Patterns. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '05)*. ACM, New York, NY, USA, 221–230. DOI: http://dx.doi.org/10.1145/1054972.1055004

37. K. Rayner. 1998. Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin* 124, 3 (Nov. 1998), 372–422.

38. Keith Rayner, Timothy J. Slattery, and Nathalie N. Bélanger. 2010. Eye Movements, the Perceptual Span, and Reading Speed. *Psychonomic bulletin & review* 17, 6 (2010), 834–839. DOI: http://dx.doi.org/10.3758/PBR.17.6.834

39. Erik D. Reichle, Alexander Pollatsek, Donald L. Fisher, and Keith Rayner. 1998. Toward a Model of Eye Movement Control in Reading. *Psychological Review* 105, 1 (1998), 125–157. DOI: http://dx.doi.org/10.1037/0033-295X.105.1.125

40. Dario D. Salvucci and John R. Anderson. 2000. Intelligent Gaze-Added Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '00)*. ACM, New York, NY, USA, 273–280. DOI:http://dx.doi.org/10.1145/332040.332444

41. Dario D. Salvucci and Joseph H. Goldberg. 2000. Identifying Fixations and Saccades in Eye-Tracking Protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*. ACM, 71–78. DOI: http://dx.doi.org/10.1145/355017.355028

42. Anthony Santella and Doug DeCarlo. 2004. Robust Clustering of Eye Movement Recordings for Quantification of Visual Interest. In *Proceedings of the 2004 Symposium on Eye Tracking Research & Applications (ETRA '04)*. ACM, 27–34. DOI: http://dx.doi.org/10.1145/968363.968368

43. Sara C. Sereno and Keith Rayner. 2003. Measuring Word Recognition in Reading: Eye Movements and Event-Related Potentials. *Trends in cognitive sciences* 7, 11 (2003), 489–493. DOI: http://dx.doi.org/10.1016/j.tics.2003.09.010

44. John L. Sibert, Mehmet Gokturk, and Robert A. Lavine. 2000. The Reading Assistant: Eye Gaze Triggered Auditory Prompting for Reading Remediation. In *Proceedings of the 13th Annual ACM Symposium on User Interface Software and Technology (UIST '00)*. ACM, New York, NY, USA, 101–107. DOI: http://dx.doi.org/10.1145/354401.354418

45. Linda E. Sibert and Robert J. K. Jacob. 2000. Evaluation of Eye Gaze Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '00)*. ACM, New York, NY, USA, 281–288. DOI: http://dx.doi.org/10.1145/332040.332445

46. Greg J. Siegle, Naho Ichikawa, and Stuart Steinhauer. 2008. Blink before and after You Think: Blinks Occur prior to and Following Cognitive Load Indexed by Pupillary Responses. *Psychophysiology* 45, 5 (Sept. 2008), 679–687. DOI: http://dx.doi.org/10.1111/j.1469-8986.2008.00681.x

47. William Steptoe, Robin Wolff, Alessio Murgia, Estefania Guimaraes, John Rae, Paul Sharkey, David Roberts, and Anthony Steed. 2008. Eye-Tracking for Avatar Eye-Gaze and Interactional Analysis in Immersive Collaborative Virtual Environments. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work (CSCW '08)*. ACM, New York, NY, USA, 197–200. DOI: http://dx.doi.org/10.1145/1460563.1460593

48. Takumi Toyama, Daniel Sonntag, Andreas Dengel, Takahiro Matsuda, Masakazu Iwamura, and Koichi Kise. 2014. A Mixed Reality Head-Mounted Text Translation System Using Eye Gaze Input. In *Proceedings of the 19th International Conference on Intelligent User Interfaces (IUI '14)*. ACM, New York, NY, USA, 329–334. DOI: http://dx.doi.org/10.1145/2557500.2557528

49. Roel Vertegaal, Robert Slagter, Gerrit van der Veer, and Anton Nijholt. 2001. Eye Gaze Patterns in Conversations: There Is More to Conversational Agents Than Meets the Eyes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '01)*. ACM, New York, NY, USA, 301–308. DOI: http://dx.doi.org/10.1145/365024.365119

50. Kazuyo Yoshimura, Koichi Kise, and Kai Kunze. 2015. The Eye as the Window of the Language Ability: Estimation of English Skills by Analyzing Eye Movement While Reading Documents. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*. IEEE, 251–255. DOI: http://dx.doi.org/10.1109/ICDAR.2015.7333762

51. Shumin Zhai, Carlos Morimoto, and Steven Ihde. 1999. Manual and Gaze Input Cascaded (MAGIC) Pointing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '99)*. ACM, 246–253. DOI:http://dx.doi.org/10.1145/302979.303053

52. Yanxia Zhang, Ming Ki Chong, Jörg Müller, Andreas Bulling, and Hans Gellersen. 2015. Eye Tracking for Public Displays in the Wild. *Personal and Ubiquitous Computing* 19, 5-6 (Aug. 2015), 967–981. DOI: http://dx.doi.org/10.1007/s00779-015-0866-8