Pupilometric and font effects in a lexical decision task

The primary aim of this study was to determine whether font characteristics (serif vs. sans serif) affected performance in a lexical decision task (LDT) under low-contrast conditions. A secondary aim was to explore possible pupilometric effects of the task. Subjects were required to fixate the centre of a computer display after which a word or an orthographically regular non- word was presented 1.5 degrees to the right of centre. The word could be in either a serif or sans-serif font (specifically, Times New Roman or Arial). The subjects task was to make a rapid lexical decision. Overall performance accuracy was significantly greater for Arial as opposed to Times New Roman. There were significant pupilometric effects for word frequency, word length, and font type across a number of pupilometric Parameters. Average pupil dilation tended to vary more as a function of visual features of the task such as font type and word length, whereas latency to peak dilation was sensitive to more cognitive task aspects such as word frequency.

Methods and Materials

Subjects

Eight female and 9 male undergraduate computer science students (M=22.08) from the National University of Ireland Maynooth volunteered in this experiment. All seventeen participants were right handed who reported no physical condition that would interfere with the recording of their pupil diameter. The number of hours of sleep along with their caffeine and tobacco consumption was also recorded. All participants were screened for at least $\frac{20}{30}$ visual acuity (noncorrected or corrected) by using a Snellen wall chart. The sample of participants was recruited using a learning management system called Moodle used by the university. The sample originally comprised of thirty-five participants but due to technical difficulties or not passing the initial phase of the experiment, only data from seventeen could be used as part of this paper.

Stimulas Material

As mentioned above the experiment consisted of two phases, training phase of 60 trials and a main phase of 120 trials. Words used for this lexical decision task were taken from the English Lexicon Project (Burnage & Dunlop, 1992). One hundred and eighty stimuli were used, 90 words and 90 non-words. The non-words were orthographically regular misspelt words of the 90 words; the words were divided into 6 groups of 15 words following an orthogonal combination of the following variables;length (4,6,8) and frequency (High, Low). Within both the training and main phase of the experiment the font of the presentation of the words was randomised between a Serif(Times New Roman) and Sans-Serif font(Arial).

Aparatus

Experimental materials were presented on a 21-inch Samsung SyncMaster 1100MB CRT monitor controlled by an nVidia Geforce FX5200, 256MB RAM graphics card with a refresh rate of 100 Hz non-interlaced. The resolution of the monitor was 1024 x 768 pixels and the screen subtended 42.8 (41.4 cm) horizontally and 32.1 (30.3 cm) vertically. This results in 24 pixels per deg at a viewing distance of 52.6 cm. Eye movements were recorded using an EyeLink II (SR Research Ltd.) head-mounted, video-based eye tracker.The EyeLink II has a maximum sampling rate of 500 Hz. They viewed the display binocularly, though only their right eye was calibrated and recorded. Stimuli were presented and controlled by a display PC and eye movement data were collected using a separate recorder PC. Responses were recorded using a Microsoft Sidewinder Joystick. The experiment was designed and programmed using MATLAB's Psychtoolbox, a graphical user interface was developed using MATLAB GUIDE to record information about the participant. The experiment took place in a room with no windows and the illuminance was measured using a 4 in 1 multifunctional sensor. Luminance was calculated using the sensor of a Samsung WB550 using the following formula

$$L = \frac{12.4 \times \alpha^2}{\epsilon} \times S$$

where

- L is the luminance in cd/m^2
- α is the apperature
- ϵ is the exposure time
- S is the sensitivity in ISO units

Procedure

Participants seated themselves in a comfortable chair with their heads stabilised in a chin rest approximately 52.6cm from the computer screen. The experiment involved two phases, a training phase consisting of 60 trials and a main phase consisting of 120 trials. Prior to each phase the eye tracker was calibrated by instructing them to fixate a series of nine dots (0.4 diameter) displayed in a 3x3 grid pattern on the screen. Stimuli were presented 1.5° to the right of a centre fixation point. The overall objective of the participant was to make a lexical decision. Participants had to judge the lexicality of a presented letter string; the task is to determine if the presented string is a word or non-word as quickly as possible. The fixation point was a cross which was colored red(RGB value=[255, 0, 51]), the participant had to make a lexical decision whilst focusing on the centre fixation point without fixating or making a saccade towards the word. The difficult nature of the task meant that a sizeable training phase be used prior to the main trial phase. In order to proceed to the main trial phase each

participant needed to make thirty correct lexical decisions within the training phase.

Pupil Diameter Conversion Method

Samples outputted from the Eyelink II during experiments contain information relating to the eye movement recorded. A sample also contains a measurement of the participants pupil diameter. This value has a set range but is arbitrary. Pupil diameter is measured in millimetres; therefore there is a need to convert this arbitrary values into meaningful ones. To do this we created an artificial eye with a pupil diameter of 5 millimetres. By collecting samples using this artifical eye it is possible to calculate a scaling factor for each participant. The method involved recording the pupil diameter for 5 seconds prior to each experiment. The scaling factor was then used to convert the participants outputted pupil diameter values into the correct unit.

The training phase had five luminance conditions, they were modulated every twelve trials. Illuminance values for these screens was (24.5,25.0,25.5,26.0,26.5)lx and luminance values were $(1.35{,}193{,}2.7{,}4.05{,}5.4)\frac{cd}{m^2}$. The purpose of this was to calculate each participants accuracy for the various levels of contrast and to determine the level with which 50% accuracy was achieved. It was calculated using the following formula

$$A = \frac{TP+TN}{TP+FP+FN+TN}$$

where

- A is accuracy
- TP is the number of true positives; if the instance is positive and it is classified as being positive
- TN is the number of true negatives; if the instance is negative and it clasified as negative
- FP is the number of false positives; if the instance is negative and it clasified as positive
- FN is the number of false negatives; if the instance is positive and it clasified as negative

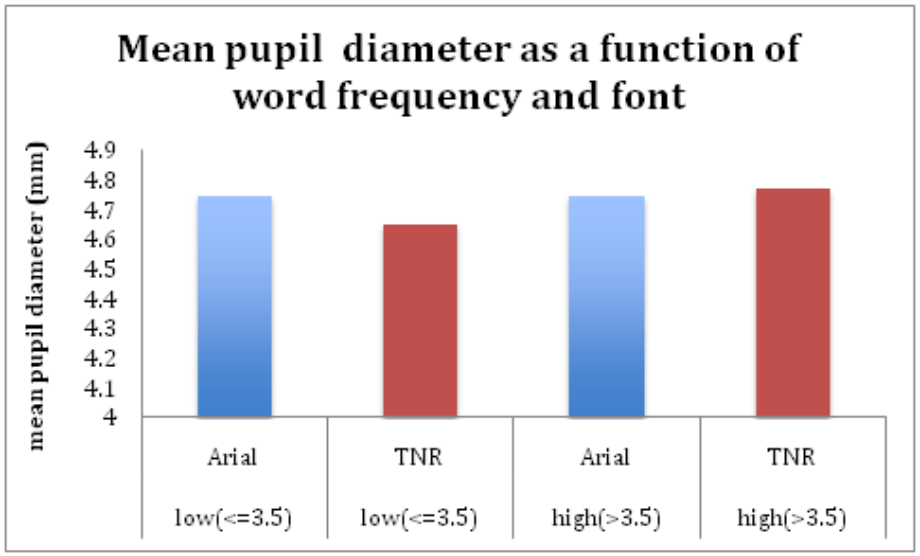
The contrast with which the participant achieved 50% accuracy was used then for that participants main phase. After the training phase, participants rested for five minutes before continuing with the main phase.

Results

Pupil Diameter Results

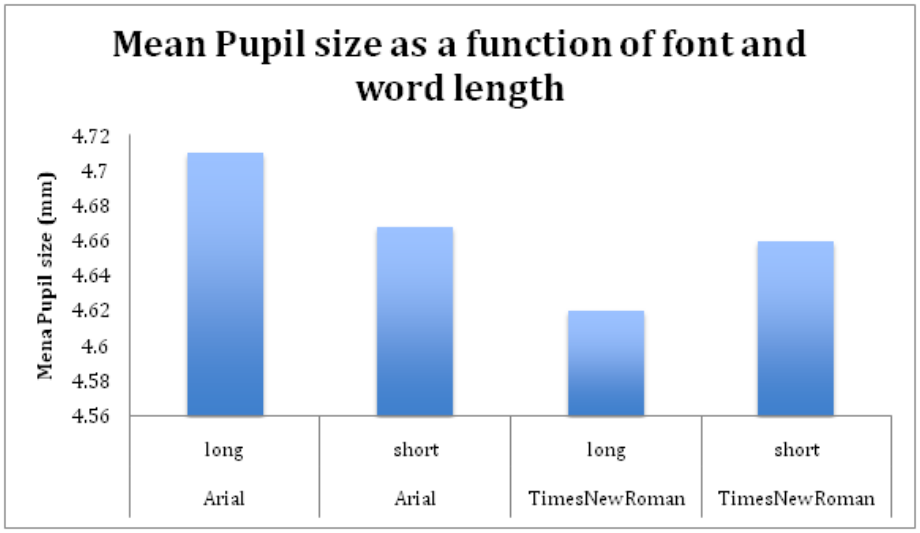
LME analysis was carried out on the data from this experiment and the following significant effects were found. Font had a highly significant effect on overall

accuracy of response ($p<0.001$) with Arial affording more accurate responses than TNR. There was a significant effect for latency to peak pupil dilation for type of font ($t = 2.12$) and for word length ($t = 2.55$). TNR gives rise to longer latencies to peak than Arial as also do longer words.



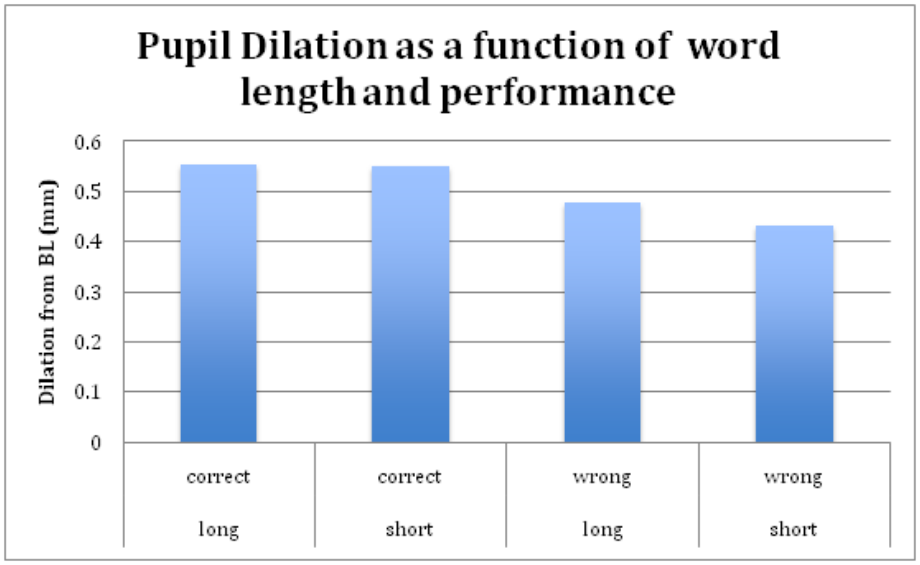
Mean pupil diameter as a function of font and word frequency

Another significant effect was observed for pupil dilation compared to a baseline (BL) measure as a function of the type of font ($t = -2.466$). TNR seems to generate smaller dilations as compared to BL Arial. Several significant interactions were also observed. Response latencies were significantly different for type of stimuli (Word vs. Non-Word) as a function of the performance of the task ($t = 5.80$).

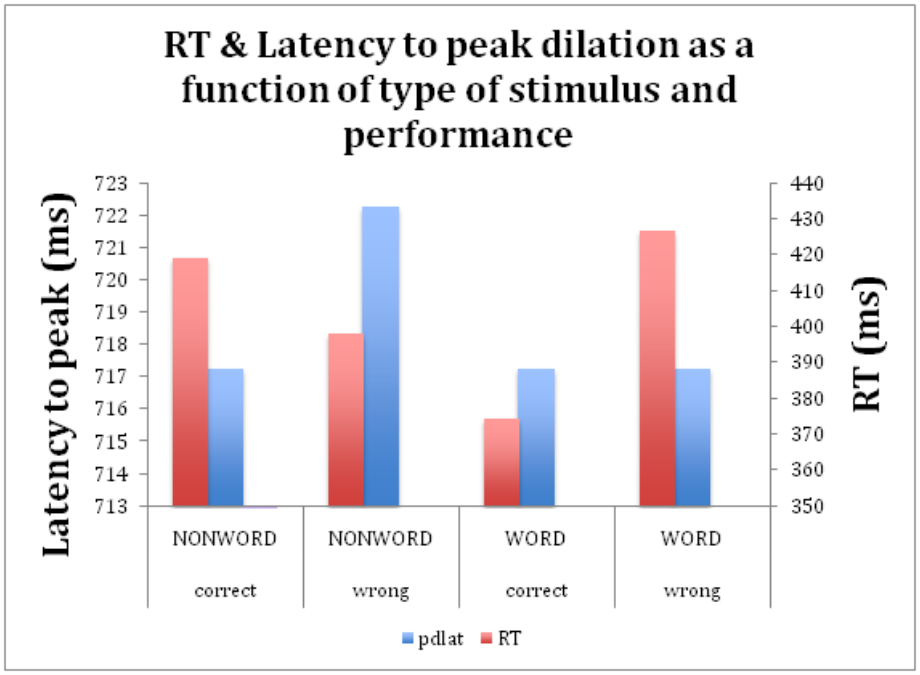


Mean pupil diameter as a function of font and word length

For example, it took more time to reach the peak dilation for the non-word condition when the participant gave a wrong answer. See adjacent graph. Additionally, a high correlation was observed between the RT and the latency to peak dilation is 0.9 for words and non-words and 0.93 for words alone.



Pupil dialation length performance



Pupil dialation reaction time performance

Receiver Operating Characteristics Curves

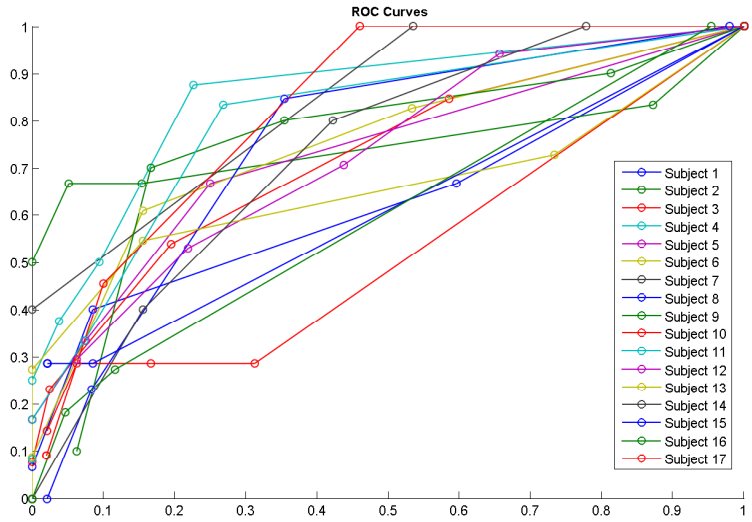
A ROC graph is a technique for visualising classifiers based on their performance. Historically ROC graphs have been used in signal detection theory(SDT) to illustrate the trade off between hit rates and false alarm rates of classifiers (Fawcett, 2006). It has also been widely used in radiology research to express diagnostic accuracy (**Eng2005?**).

Since one of our experiments was a LDT, a binary classification model is used. That is, each instance $\boldsymbol{I}$ is mapped to and element of the set {**p**,**n**} of positive and negative class labels. (Fawcett, 2006) describes a *classification model* as a mapping from instances to predicted classes. Once a subject is presented with a stimulus there are four possible outcomes based on their classification. Since the instance could be a word or non-word, a word is described as being positive and a non-word is negative. If the instance is positive and it is classified as being positive, it is counted as a *true positive*; if it is classified as being negative, it is counted as being a *false negative*. If the instance is negative and it clasified as negative, it is counted as a *true negative*; if it classified as positive, it is counted as a *false postive*. When creating ROC curves the two metrics of interest are the **true positive rate** and the **false positive rate**. They are plotted on the Y axis and the X axis accordingly. Theses are defined to be

$$tp\ rate \approx \frac{\text{Positive instances correctly classified}}{\text{Total Positives}}$$

$$fp\ rate \approx \frac{\text{Negatives incorrectly classified}}{\text{Total Negative}}$$

The points of the ROC are plotted each between 0 and 1 on each axis accordingly. The points of the graph are representative of the strategy used to classify the instances. Entries along the diagonal x=y are indicative of a guessing strategy. If points are located close to the left hand side of the X-axis, they are thought to be conservative since they make few false positive classifications but make positive classifications only with strong evidence. Classifiers located in the up left are thought of as *liberal* since they make positive classifications with weak evidence (Fawcett, 2006).



ROC Curve for Subjects

Discussion

TNR gives rise to longer latencies to peak dilation than Arial as do longer words suggesting that it takes longer to gather the cognitive resources needed for the task while performing under more challenging circumstances. Paradoxically, TNR seems to generate smaller dilations, as compared to a baseline measure, than Arial suggesting that more cognitive effort is devoted to perform the task with the latter font. That is also shown in the mean pupil size when taking into account the word length. Longer words made the pupil dilate more but only with Arial font, indicating increased cognitive demands. Longer dilations

were observed for correct trials and in incorrect trials dilations were smaller for short versus long words. This seems to indicate that the more cognitive resources you devote to task, the more accurate your response. Complementarily, it also suggests that the longer the stimuli, the harder they are to process. Less frequent words, took more time to reach the maximum pupil dilation suggesting, again, that the cognitive resources needed for the task are greater for harder task conditions (processing low frequency words is harder than processing very common words). Additionally, low frequency words seem to be more easily processed when displayed in TNR but high frequency words seem to be harder to process.

References

Burnage, G., & Dunlop, D. (1992). *Encoding the british national corpus.*

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*, 861–874.