Adaptive Fact Learning: Using Pupil Dilation to Measure Item Difficulty

Maarten van der Velde
Department of Artificial Intelligence
University of Groningen, Groningen, The Netherlands
m.a.van.der.velde.2@student.rug.nl

ABSTRACT
Pupil dilation has been shown to be an indicator for the effort needed to retrieve a fact from memory. This paper describes a computer-assisted fact learning system, in which changes in pupil dilation serve as a measure of the difficulty of each study item. The order in which items are presented to the learner is continuously adapted on the basis of their estimated difficulty. Results from an experiment in which participants memorised a set of translations show that the system could identify difficult items by the pupil response they evoked and dedicate more rehearsal time to them. Although more research is needed, these preliminary findings show that pupil dilation could contribute to an efficient fact learning system.

Author Keywords
Pupil dilation; memory retrieval; adaptive learning system.

INTRODUCTION
The ability to recall particular facts is integral to any student's success. Countless hours are spent memorising historical dates, topography, and translations of words from one language to another. Some of these facts are inherently more difficult to master than others, and individual differences can mean that a fact that is easy for one student is hard for another. This means that the process by which a set of facts is memorised is shaped both by the facts themselves, and by the learner.

Simple, commonly used approaches to fact memorisation can be quite inefficient. For example, consider the method of repeatedly going through a stack of flashcards. Some items will be memorised quickly, but will keep getting repeated at the same frequency anyway, while others are not repeated often enough to be remembered. What this shows is that the spacing between repetitions of an item should not be fixed, but should reflect how urgently an item needs repeating.

A number of computer-assisted learning systems have been developed on this principle, dynamically scheduling items in a way that prioritises facts that are close to being forgotten. Van Rijn and colleagues did this by estimating the activation of each item (an indicator for the strength of an item's memory trace that decays over time), and presenting items when their activation was about to drop below a threshold value. Response times and errors were subsequently used to adjust the estimated activation. Retention rates with this system were higher than with a regular flashcard strategy [3].

A different way to determine how urgently an item needs to be rehearsed may be to use the learner's pupil response to it. Pupil dilation has been shown to be an index of the difficulty of individual words with which a participant is presented, with more difficult words eliciting a larger dilation [1]. Previous work has also found that pupil dilation can serve as an indicator for the strength of individual memory traces, and that as an item's memory trace gets consolidated through repetition, the pupil's response to it becomes smaller [2].

Based on these findings, this study investigates the possibility of using pupil dilation to inform item scheduling in a computer-assisted fact learning system. Pupil dilation can provide a continuous measure of the mental effort that is spent retrieving the answer to a prompt, potentially giving a more direct and richer insight into the item's difficulty than traditional behavioural measures can provide. The algorithm that the system uses is described in the following section, after which an experiment in which participants learned translations of Zulu words using the system is discussed.

ALGORITHM
A study session consists of a set of trials, each pertaining to a particular prompt-response pair that is to be learned. In each trial the learner is presented with a prompt, and then attempts to retrieve the corresponding response from memory. The increase in pupil dilation that happens during this retrieval is taken to be indicative of its difficulty: the larger the relative dilation, the larger the mental effort needed to retrieve the response.

The algorithm determines which item needs to be rehearsed most urgently at any point during a learner's study session. It bases this prioritisation on the estimated difficulty of each item at the time, which is represented by a difficulty score. For the duration of the rehearsal, the system will follow these steps to select an item each time a trial starts:

1. If there is an item that has never been rehearsed, select it.
2. If all items have been rehearsed at least once, select the item that has the highest difficulty score and has not been rehearsed within the last two trials.

Whenever an item is presented to the learner, a new difficulty score for that item is calculated that reflects how the learner responded to this presentation. To calculate the difficulty score, the pupil size is measured for the duration of the presentation of the prompt. Each measurement is converted to a percentage that expresses the change relative to a baseline pupil size measured just before the prompt was presented (percentage change pupil size; PCPS). PCPS is a useful metric because it allows us to compare pupil responses in different trials and across parti-

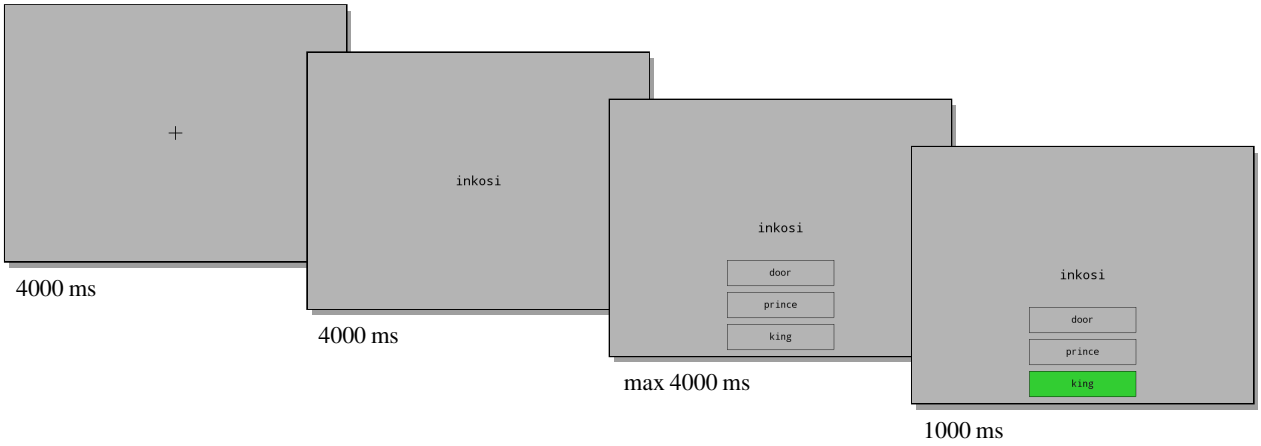

Figure 1. The four stages of a rehearsal trial: fixation, prompt, answering, and feedback. Test trials followed the same procedure but omitted feedback.

cipants, regardless of differences in baseline pupil size. If the learner's answer to an item is correct, the difficulty score for that trial is equal to the maximum PCPS value that was measured during the trial. This value is used because it represents the peak mental effort. If the answer is incorrect, the difficulty score is the maximum PCPS value multiplied by 1.1. In this way we ensure that an incorrectly answered item receives a small boost in its difficulty score, which makes it more likely to be repeated than a correctly answered item with the same PCPS.

EXPERIMENT

Participants

10 students from the University of Groningen (4 female, age range 21-31 years, mean age 25.2 years) took part in the experiment without compensation. All participants were non-native English speakers and had no previous experience with Zulu.

Stimuli and design

In the experiment participants learned the English translations of fifteen Zulu nouns. The translations were well-known words in English. The experiment consisted of two blocks of trials: a rehearsal phase and a test phase. Trials in either phase contained one of the fifteen Zulu words, along with three potential translations in English, only one of which was correct. Two additional distractors were drawn randomly from a different set of English nouns. The Zulu words had a length of five to nine letters, and had differing similarity to their English counterparts. Some were quite similar (e.g., *womuntu – human*) while others were not (e.g., *thulile – restaurant*). The order of item presentation in the rehearsal block was determined on the fly by the algorithm described in the previous section. The rehearsal phase took 20 minutes (except for two participants, who took part in an earlier version of the experiment with a 10-minute rehearsal phase). The rehearsal phase was followed by a test phase, in which each item was presented only once in a random order.

Apparatus and setup

The experiment was built in OpenSesame and ran on Windows Vista. It was presented to the participants on a 20.1 inch monitor with a resolution of 1600 by 1200 pixels. Pupil measurements were made using an SR Research Eyelink 1000 eye tracker, which was located directly under the screen. Participants were seated on a chair in front of the monitor, their head being held in position by a desk-mounted chin rest. Participants used a computer mouse to make their responses. Before the start of the experiment the chair and chin rest were adjusted to fit the participant, and a nine-point calibration of the eye tracker was performed, followed by a nine-point validation.

Procedure

At the start of the experiment participants saw the task instructions on screen. They were told that a Zulu word would appear, for which they had to recall the translation, and that after a few seconds three answer options would appear, from which they had to select the correct one within a limited time. Feedback would be given afterwards. Once these instructions had been read, the rehearsal phase began.

Figure 1 shows an example rehearsal trial. Each trial started with a 24 point black fixation cross on a grey background, at the centre of the screen. After four seconds the cross was replaced by a Zulu word. The word was rendered in black in a 22.5 point monospaced font. Four more seconds passed before the three answer options appeared in a column below the Zulu prompt. Each answer was presented in 16.5 point black monospaced font and was surrounded by a black rectangular outline that formed a button for the participant to click. Once the participant selected one (or after a four-second time limit had passed) the feedback appeared. The button corresponding to the correct answer was always coloured green, regardless of the participant's response. In case of an incorrect response, the button of the selected answer was coloured red. The feedback was shown for one second, after which the next trial started. This process continued until the rehearsal time limit of 20 minutes had passed.

After the rehearsal phase, the instructions for the test phase were shown. The procedure was identical to the rehearsal phase, except that now there was no feedback after a trial. Once each of the fifteen items had been presented, the final test score was shown, and the participants were thanked for their participation.

Measurement and preprocessing of pupillary data

Pupil measurements were recorded during each trial. During the last 250 ms of the fixation cross and throughout the four

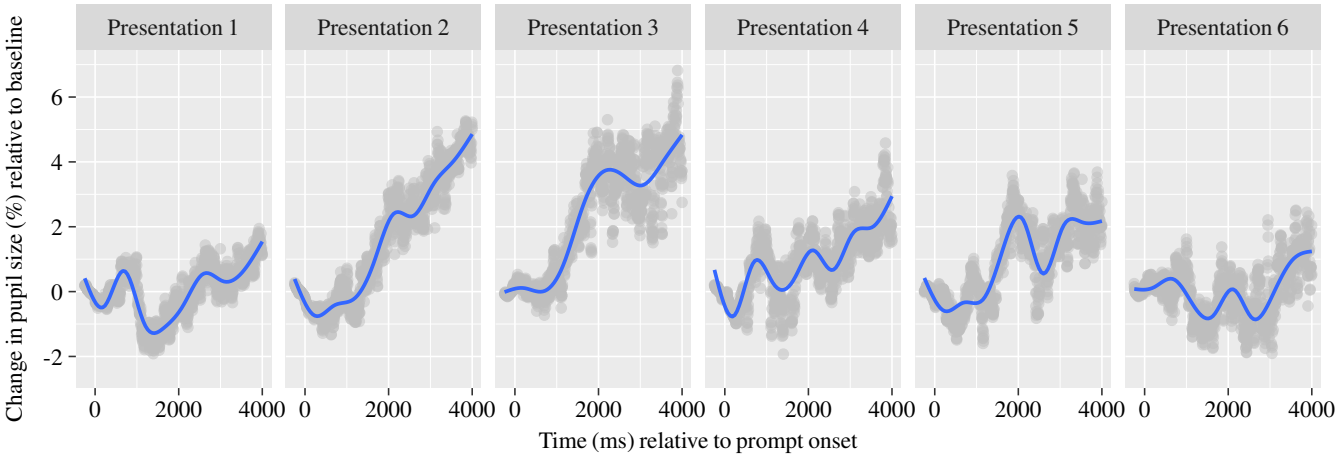

Figure 2. Mean percentage change pupil size (relative to each trial's baseline) across all rehearsal trials and all participants, separated by the number of presentations the item has had. The grey points represent the mean value at each time step. The solid blue line is a smoothed representation of the data.

seconds in which the prompt was presented, the eye tracker measured the size and location of the left pupil at a frequency of 250 Hz. Using these measurements, the system logged the pupil size at a frequency of 50 Hz. Blinks were detected and removed automatically by the eye tracker. To filter out remaining artefacts resulting from blinks, the top 2.5% and bottom 2.5% of the pupil size measurements logged during a trial were discarded.

Measurement of performance data

Whenever a participant clicked on the correct translation of a prompt within the time limit, their response was deemed correct. Selecting an incorrect translation, or selecting no answer at all, was considered an incorrect response. Response times were measured from the moment at which the answer options appeared on screen to the moment when a mouse click on one of the answer options was registered.

Results

Of the ten students who participated, one did not complete the whole experiment and was therefore not included in further analyses. Another participant was excluded for not responding in any of the rehearsal trials. As mentioned previously, two participants had a shorter rehearsal phase of 10 minutes, compared to a 20-minute rehearsal phase for all other participants. Where differences exist, the results of these two groups are presented separately.

Performance summary

Participants who rehearsed for 20 minutes completed 94 rehearsal trials on average (SD = 3.10), while those with a 10-minute rehearsal completed 41.5 rehearsal trials on average (SD = 6.36). Participants with 20-minute rehearsals had a mean response accuracy during rehearsal of 85.3%, versus 94.4% in testing. Predictably, the participants with a 10-minute rehearsal had a lower mean accuracy: 72.3% during rehearsal and 80.0% in testing. Of all errors, 17.1% were caused by a failure to respond within the time limit; the other 82.9% being attributable to the selection of an incorrect answer. In the rehearsal phase participants had a median response time of 1609 ms (IQR = 930 ms). During the test phase responses were about half a second slower, with a median value of 2160 ms (IQR = 899 ms).

Item repetition

As expected, the number of times each item was repeated varied quite strongly, due to the algorithm's prioritisation of more difficult items. Repetitions followed a heavily right-skewed distribution, meaning that most items were only repeated a few times, while a small number of them were presented much more often. For participants with a 10-minute rehearsal phase, the median number of presentations for a particular item was 2.5 (IQR = 1.75). Because there was more time for repetition, the 20-minute rehearsal group had a higher median of 6 presentations per item (IQR = 6).

The effect of repetition on pupil response

Based on previous research [2], we hypothesised that a prompt would elicit progressively weaker pupil responses as it was presented more. The repetition of an item was expected to strengthen the corresponding memory trace, making it easier to retrieve the answer. Figure 2 appears to bear out this expectation. It shows the PCPS across the four seconds in which a prompt was presented, averaged over all rehearsal trials from all participants, separated on the basis of the number of times the prompt had been shown. The figure suggests that there is indeed a relationship between the number of presentations an item has had, and the magnitude of the pupil response it evokes. The response to the first presentation is relatively weak, which may be explained by the fact that at that moment the participant has never seen the Zulu prompt before, and would therefore not even attempt to retrieve a translation. From the second presentation onwards, when the participant could attempt a retrieval of the translation from memory, there appears to be a downward trend in the height of the pupil response. This trend would mean that repeated exposure to a prompt would indeed reduce the mental effort needed to retrieve its translation in a measureable way.

Prioritising difficult items

An item that elicits a stronger pupil response in a particular user should be regarded as more difficult for that user, and the system should consequently dedicate more rehearsal time to it. Conversely, easier items should command less time. Figure 3 suggests that the system succeeded in this prioritisation. It

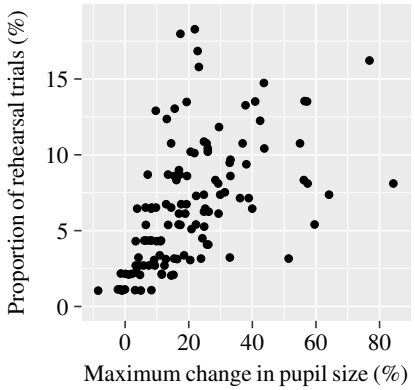

Figure 3. Relationship between a participant's maximum percentage change pupil size (relative to each trial's baseline) across all rehearsal trials of a particular item, and the proportion of trials dedicated to that item. Each point represents one item for one participant.

depicts, for each item and each participant in turn, the relation between the item's overall difficulty for that participant (calculated here as the maximum PCPS across all trials in which it occurred) and the proportion of rehearsal time spent on it. The figure seems to show quite a clear trend: when the overall difficulty of an item is higher for a participant, it tends to occur in a larger proportion of their rehearsal trials. This relationship is confirmed by a strong Spearman's correlation (corrected for ties), $r_s = 0.652$, $n = 120$, $p = 7.189 * 10^{-16}$.

Adapting to individual differences in difficulty

Figure 4 provides a closer look at the data from Figure 3 by showing the same relationship for six of the fifteen items individually. Once again the maximum PCPS across trials of an item is plotted against the proportion of trials dedicated to that item. Each letter represents a participant. The figure shows two things. Firstly, it confirms that the difficulty of an item can indeed differ quite dramatically between participants, and that the system adapts its rehearsal frequency because of this. For example, the prompt *isinkwa* (bread), shown in the top-right panel, elicited a larger pupil response for participant *c*, and also took up much more of the rehearsal time, than for participant *a*. Secondly, the figure shows that differences between participants seem to be larger for some items than for others, which reflects the notion that some items (such as *womuntu* (human); shown in the bottom-right panel) might be inherently easier for everybody.

DISCUSSION

This study evaluated the feasibility of using pupil dilation to inform item spacing in a fact memorisation task. We designed a system that used changes in pupil size evoked by the presentation of items as an indicator of their difficulty. A stronger pupil response to a prompt was taken to signal that retrieving its translation was more mentally taxing. The order in which items were rehearsed was determined on the fly by an algorithm that prioritised items that evoked the biggest response. This approach was found to be effective in a task in which participants learned translations of Zulu words. The results showed that the system could identify the more difficult items among the easier ones, and spent more time rehearsing

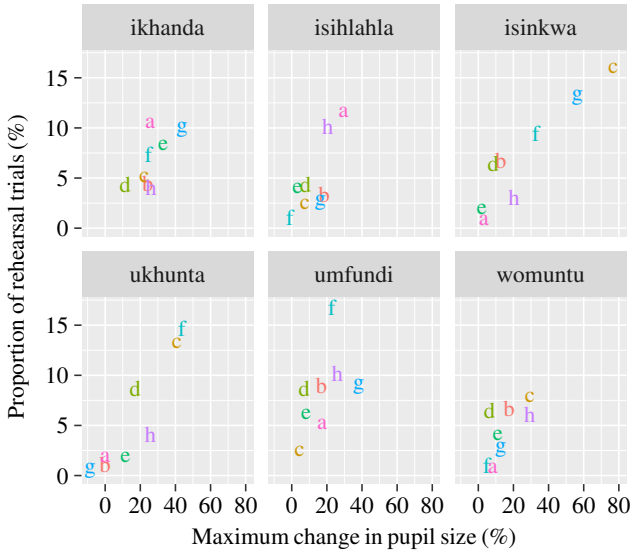

Figure 4. Relationship between the maximum percentage change pupil size (relative to each trial's baseline) across all rehearsal trials of a particular item, and the proportion of trials dedicated to that item, for six of the fifteen items. Each letter represents a participant.

them. It did this on a purely individual basis, and adapted its assessments of difficulty continuously to the latest measurements. These findings suggest that pupil dilation can indeed be used for scheduling items in a computer-assisted learning system.

Future work

A logical next step in validating this approach is to compare retention rates of users of this system with those of other study methods, such as the flashcard method or the various activation-based methods described in [3]. A further improvement might be made by adapting the length of the rehearsal to the individual, which might prevent unnecessarily long study sessions. For example, the system might end the rehearsal as soon as the change in pupil size evoked by an item falls below a certain threshold for all items. Finally, it may be beneficial to the accuracy of the system's difficulty estimation to use behavioural measures such as response time and accuracy alongside pupil dilation. Following [3], a quick response could be taken to mean that the answer was easy to retrieve from memory, while an error would indicate that the memory trace was too weak to be retrieved.

REFERENCES

1. Chapman, L. R., and Hallowell, B. A Novel Pupillometric Method for Indexing Word Difficulty in Individuals With and Without Aphasia. *Journal of Speech Language and Hearing Research 58*, 5 (2015), 1508.
2. van Rijn, H., Dalenberg, J. R., Borst, J. P., and Sprenger, S. A. Pupil Dilation Co-Varies with Memory Strength of Individual Traces in a Delayed Response Paired-Associate Task. *PLoS ONE 7*, 12 (2012), e51134.
3. van Rijn, H., van Maanen, L., and van Woudenberg, M. Passing the test: Improving learning gains by balancing spacing and testing effects. In *Proceedings of the 9th International Conference of Cognitive Modeling*, vol. 2 (2009), 7–6.

WHO DID WHAT?

- Programming the system: Maarten
- Supervising the actual experiment:
  - first session: Marten, Kim
  - second session: All
- Compiling the list of Zulu words: Marten
- Experiment design: All
- Testing the system: All
- Recruiting participants: Marten, Kim
- First presentation:
  - Slides: Marten
  - Presenting: All
- Final Presentation:
  - Slides: Kim, Maarten
  - Presenting: All
- Analysis R-script, used to generate plots in the presentations: Maarten