Reading and Learning Smartfonts

Danielle Bragg
University of Washington
Seattle, USA

Shiri Azenkot
Cornell Tech
New York, New York

Adam Tauman Kalai
Microsoft Research
Cambridge, MA

ABSTRACT

Reading text on small screens can be difficult because many people have trouble focusing their eyes on small text. However, small screens are becoming increasingly pervasive with the advent of personal computing devices, and small text makes the best use of screen real estate. We design multiple different scripts for displaying English text, legible at small sizes even when blurry, for small screens such as smartphones and smartwatches. These "smartfonts" redesign visual character presentations to improve the reading experience. Like cursive, Grade 1 English Braille, and ordinary fonts, they preserve orthography/spelling. They have the potential to enable people to read more text comfortably on small screens, e.g., without wearing their reading glasses. We also consider the difficulty of learning to read smartfonts fluently and observe a learnability/readability trade-off. We artificially blur images and evaluate their readability using paid crowdsourcing. For blurry text, our most learnable font can be read smaller than half the size of the traditional Latin (i.e. "English") script, and can be read at roughly half the speed of regular text after two thousand sentences.

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation (e.g. HCI): User Interfaces

Author Keywords

Fonts; Reading; Learning; Scripts; Presbyopia.

INTRODUCTION

More and more people use smartphones and smartwatches for a variety of activities. Reading text is a primary component of our interaction with these devices. However, their small screens can make it difficult to perceive letters and words. Can we design radical new scripts for these small screens? Or are the letters Romans inscribed in stone millenia ago, which we still use today, optimal? These new "smartfonts," modern analogs of cursive or (Grade 1) English Braille, could offer several potential advantages over traditional letters. First, consider *presbyopia*, the inevitable and irreversible decrease in the eye's ability to focus with age, resulting in blurred near vision for virtually everyone by age 51 [18, 1, 7]. Smartfonts that

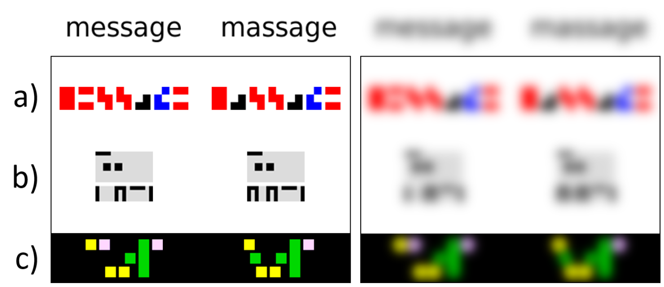


Figure 1: The words *message* and *massage* clear (left) and blurred (right) in our smartfonts: (a) Tricolor (b) Logobet (c) Polkabet (on black). Words have been sized to have equal areas.

are robust to blur could be easier to read for presbyopes and others. Even for someone with 20/20 vision, out-of-focus text appears blurry, e.g., text on a smartphone navigation system will appear blurry when one is focusing on the path in front of them. Second, text that is robust to blur might also be readable at smaller sizes.

Our smartfonts employ blocks (not merely strokes) of color to improve visibility at small sizes and with blur. Some smartfonts may be displayed perfectly with only six pixels per letter. Fonts that are easier to read in small sizes can be used either to display an equal amount of text more clearly, or more text in the same area with equal legibility. Other potential benefits include increased privacy (e.g., for obfuscating embarrassing personal messages that may pop up on smartwatches), reading speed, aesthetics, personalization, and comfort (e.g., fatigue and motion-sickness). Because software can render text in smartfonts as easily as existing fonts, the value of a smartfont does not hinge on large-scale adoption. For instance, one of the authors has been wearing a smartwatch with firmware updated to display all text in a smartfont. SMS senders do not know their messages are read in a smartfont.

For simplicity, in designing our smartfonts we consider only distinct renderings of the twenty-six letters so the user can read the text, letter for letter, without changes in orthography. In particular, we do *not* consider spelling changes or shortenings such as reading without vowels, though they could be used together with smartfonts. We focus on English, but similar ideas may be applicable to other languages.

Our questions are: (a) what are effective font designs to improve these ancient twenty-six letters for display on small screens, and (b) how difficult would they be to learn? If the answers were negative – if optimal scripts require significant learning and only offered marginal improvements over the Latin forms – then smartfonts could still serve as important icons and sources for healthy scholarly debate like the Dvorak keyboard and Esparanto language.

We use paid crowdsourcing to facilitate the rapid development and evaluation of smartfonts. Presbyopia is simulated to this remote crowd by applying a Gaussian blur to the image. Our data suggests that it is possible to design smartfonts that, compared to the traditional Latin A-Z, are more readable when blurry or, equivalently, can be displayed at smaller sizes with equal clarity. In particular, the smartfont in Figure 1b can be read smaller than half the size of Latin text when blurry, *without training*, by crowd members. This increased readability can help people read smartphones or smartwatches at a glance, even when not wearing reading glasses.

The second key factor we consider is the difficulty learning to read a smartfont, which is similar to learning cursive after print. We show that our smartfonts, to varying degrees, can be read fluently with a reasonable amount of practice. We also find a learnability/readability trade-off: certain scripts, especially ones that resemble the Latin alphabet, are easier to learn but perform worse with blur. Our Tricolor script offers a reasonable compromise in that it is relatively easy for many people to learn to read quickly.

Evaluating and optimizing unfamiliar smartfonts is challenging. Even with tutorial and practice, a person's comfort with their native character set will be far greater. However, if a reader can make out an unfamiliar smartfont more clearly than a familiar font, it is likely that the smartfont is more readable. The increased familiarity that comes with practice would only make the smartfont more readable. Thus, our testing methodology is to use crowdsourcing to compare the identifiability of random strings of artificially blurred letters at various sizes (measured by *area*) in our smartfonts to that of Latin text. This methodology enables us to compare and optimize our designs without having to train someone to read fluently at each iteration.

The key contributions of this paper include: (a) raising the theoretical question of how much one might improve over ancient scripts for display on screens by radically redesigning characters, (b) introducing and demonstrating how one can design and optimize (based on data) smartfonts for learnability/readability under specific reading conditions, in our case varied size and blur, and (c) providing a methodology for evaluating readability before teaching people to read fluently.

RELATED WORK

Related work spans fields such as HCI, psycholinguistics, design, perception, and economics. Due to space limitations, we discuss some of the most related work.

A motivating starting point for our study is the work on human perception by Pelli et al. [19] who compared the "efficiency" of letter identification across various languages. They also compared random block patterns of varying sizes, and found 3×2 block patterns, surprisingly, to be three times as efficient as traditional alphabets. Efficiency was measured by how well individual letters could be identified in the presence of random noise, which is different but possibly related to blur. They also found that a few thousand training examples sufficed to teach someone to identify unfamiliar letters fluently. This work motivates our use of 3×2 blocks as the basis of several scripts and informs our understanding of learning scripts.

Traditional alphabets have several properties of interest. Some scripts, such as the Korean alphabet Hangul, are *featural* meaning the shapes of the letters encode phonological features of the sounds they represent. It has been found that *mature* scripts, say those that have been in use for over 350 years, have many fewer mirror-image letter pairs, such as the lower-case Latin pair b/d, than younger scripts [22]. Motivated by a variety of factors, numerous creative scripts have been constructed by artists and hobbyists,[1] though we are unaware of any rigorous studies of their learnability or readability.

Legge et al. conducted a series of studies exploring various aspects of reading, including contrast [12] and low-vision [13].

Regarding adoption of new technologies, a notable relevant example has been the debate about the Dvorak keyboard's adoption "failure": Economists used early studies to claim that it is 20-40% faster than the QWERTY keyboard, and thus the low adoption rate of the significantly "superior" Dvorak keyboard proves how difficult it is to change behaviors [5], while others used later studies that Dvorak is only 2% faster [15]. Later input techniques garnered higher adoption on PDAs and smartphones [9, 24], highly influenced by user preference. Smartfonts, like virtual keyboards, require no additional hardware and can be personalized.

Crowdsourcing has been used to understand perception. For instance, Demiralp et al. explored the use of crowdsourcing to evaluate the perceptual similarity of different shapes and colors, and developed perceptual kernels to quantify crowd-learned similarity. [6] They found crowdsourcing to be an inexpensive, rapid, and efficient means to gather data on human perception.

Font design has been shown to strongly impact the reading experience for people with vision conditions. Prior studies on color-grapheme synethesia, where people have strong associations between letters and colors (see, e.g., [4]), have shown that reading books with colored letters suffices to passively learn and create strong perceptual associations between letters and colors. In her dissertation, Bessemans explored font design for children who are low-vision and just learning to read. [2] Children with visual impairments are at a disadvantage in comparison to their normally sighted peers in learning to read. Bessemans designed a font especially for low-vision children, based on her exploration of the effects of font design parameters on legibility for children.

HCI techniques proposed for improving digital reading digitally include RSVP[2] [10], Froggy [23], ClearType [8], and

[1]A collection of constructed scripts can be found at http://omniglot.com.

[2]RSVP has recently received attention due to http://spritzinc.com and its inclusion on the Microsoft Band smartwatch.

visual syntactic text formatting [21], among others. These techniques could be combined naturally with smartfonts. Similarly, to ease learning one could adapt existing teaching techniques such as software that gradually teaches a user a language by introducing new words over time [20].

OUR FONTS

We initially designed three smartfonts to be easily readable, even at small sizes and out of focus. We leveraged three main techniques: 1) using simple characters characterized by high contrast and blocks of color; 2) using color to distinguish between characters; and 3) radically reducing the space between adjacent characters. Shape and color are primary dimensions that distinguish two-dimensional shapes. Our fonts Visibraille and Polkabet each leverages one of these dimensions. Text is not simply a collection of individual characters; these characters form words and sentences, and pairs of adjacent characters require differing amounts of spacing between them. Our font Logobet drastically reduces space between characters.

Visibraille

Both theoretical [14] and experimental [19] work suggests that simple characters, characterized by high contrast and thick line strokes, are most easily recognizable. Out of a range of established and made-up alphabets, Pelli et al [19] showed that 3×2 grids (a visual analog of Braille) are most easily recognizable. Our first font, Visibraille, shown in Figure 2, is based on Pelli et al.'s findings. It maps 26 3×2 blocks onto the 26 letters of the English alphabet. We selected the 3×2 blocks to be similar in shape to the English characters they represent. Because of its simple design and similarity to Latin characters, we expect this font to be both readable and learnable.

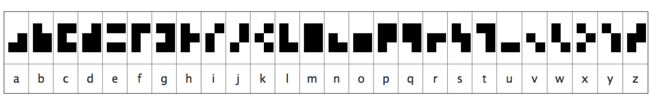


Figure 2: Visibraille alphabet

Polkabet

Color can also help to distinguish between characters. For example, if two characters have similar shapes, making one yellow and the other blue makes them clearly distinguishable. However, there is a trade-off between the number of colors that a font uses and the distinguishability of its characters. If many colors are used, then the colors are pushed closer together in color space and become hard to differentiate. Blocks of color are also resilient to blur. When an image is blurry or out of focus, each pixel appears to be a mixture of nearby pixels. Solid blocks of color are highly robust to this type of blurring, because many nearby pixels are likely to have the same color.

Our font Polkabet, shown in Figure 3, leverages color to yield 26 differentiable characters. It uses five colors: red, yellow, green, blue, purple, and white. These colors were chosen for readers who are not colorblind, but could readily be tailored to various color-blindness. We chose to make two characters rainbows to avoid adding another color. Polkabet is designed

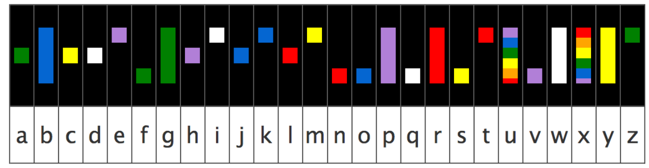


Figure 3: Polkabet alphabet

to be read on a black background, and consequently is uniquely suited for small, personal devices like the smartwatch.

Figure 4: Mnemonics for Polkabet's small square characters.

We developed a mnemonic system, shown in Figure 4, to help people learn this smartfont. If a character uses a small square of color, the reader can think of the associated item from Figure 4 to remember which letter it represents. The first letter of that item is the letter that the character represents. Squares of color at the top are linked to foods, middle squares link to animals, and bottom squares are associated with miscellaneous items. For example, suppose a reader encounters a red square at the top of the line and does not remember what that means. He/she would think, "This character uses a red square at the top. So think of the red food... Tomato! 'Tomato' starts with 't', so that's a 't'!". Characters that are solid blocks of color represent the first letter of that color (with the exception of X).

Logobet

Logobet is a font that visually resembles a logography, like Chinese, but is in fact an alphabet that can be sounded out, somewhat like Korean. Logobet radically reduces the spacing between letters. Kerning refers to changing the amount of space between adjacent letters in proportional fonts. A proportional font is one where the space allotted to a character is proportional to the character size. For example, an "m" will be allotted more horizontal space than an "l." However, sometimes it is desirable to reduce the spacing between particular pairs of characters. For example, a capital "T" allows a short subsequent letter, like an "m," to shift left under the "T" 's umbrella.

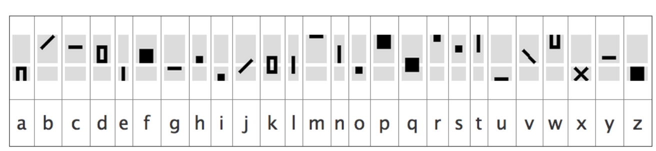


Figure 5: Logobet alphabet

(a) The Logobet alphabet without kerning

(b) The Logobet alphabet with kerning

Figure 6: Example of aggressive kerning with the alphabet.

Our font Logobet, illustrated in Figure 5, has characters that are designed so that they allow subsequent characters to be entirely shifted underneath. This means that text is read first top-to-bottom, then left-to-right. For example, Figure 6 shows the reduction in space due to kerning in Logobet for the alphabet.

OPTIMIZATIONS

While we designed Visibraille, Polkabet, and Logobet to be readable at small sizes and out of focus (and to be learnable), further optimizations can be made. In particular, we present two optimized fonts: Visibraille 2, which is made of 3×2 blocks chosen to be minimally confusable; and Tricolor, which leverages both color and familiar 3×2 character shapes of Visibraille. Using paid crowdsourcing, we generated a confusion matrix on 3×2 shapes which is used to optimize Visibraille 2's shapes and Tricolor's use of color.

Visibraille 2

Because of Visibraille's close resemblance to the English alphabet, we hypothesized that it would be easy to learn and remember. However, its characters are not necessarily the most distinguishable set of 3×2 blocks. Here we present an alternate font Visibraille 2, comprised of the least confusable 26 3×2 blocks.[3] Figure 7 shows the 26 selected shapes in black and the remaining 16 in gray.

We determined the confusability of these 42 characters using paid crowdsourcing on Amazon's Mechanical Turk platform.[4] Our experimental setup mimicked a Snellen eye chart test, a familiar test routinely used in eye exams, to the extent possible. A typical worker was shown a sequence of rows of decreasing size, one row at a time, with 1-9 characters at the precise heights indicated in Figure 8. Participants were asked to transcribe the observed character(s) using a virtual keyboard consisting of 7 shapes. To ensure that the shapes on the chart appear on keyboard, we first picked 7 random characters from 42 for the keyboard, and then sampled from those 7 characters with replacement to generate the target sequence of appropriate length, rendered at the appropriate size.

In total, we collected 4022 evaluations over 548 people. Each character was shown between 379 and 500 times, with an average of 442.1. Each character appeared with each other

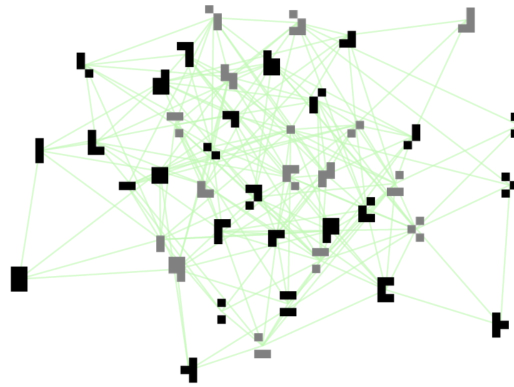


Figure 7: The 26 selected shapes (black) of the 42 considered. Edges denote highly confusable shapes. Although the confusability matrix appears to be high-dimensional (as measured by eigenvalues), the two-dimensional graph generated using D3's force-directed graph layout [3] is able to display many confusable pairs near one another.

character as a transcription choice at least 33 times, with an average of 64.7. Since experiments were conducted remotely through web browsers, we did not control for display conditions or factors such as retinal angle. However, this enables us to assess the relative confusability of different shape pairs "in the wild," across a wide variety of display types and people. To avoid pixelation artifacts, participants were instructed to keep their web browsers at the default 100% zoom.

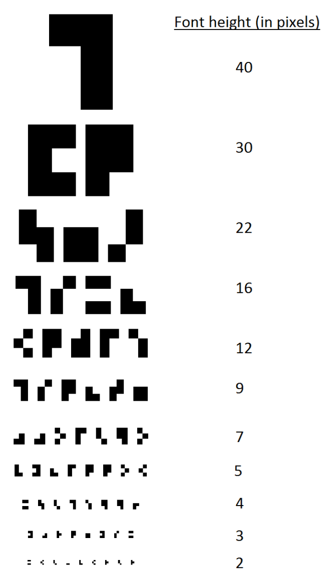


Figure 8: Our confusion matrix was generated by testing a series of rows of random characters, like a Snellen chart.

To compute the confusability matrix C, for each pair of shapes i and j, the confusability score c_{ij} is the number of times shape j was transcribed when i was shown, divided by the number

[3] Of the $2^6 = 64$ possible 3×2 configurations of black and white squares, we considered a subset of 42 characters to reduce labor. In particular, two configurations were considered to have the same shape and likely to be confused if the black patterns were translations of one another.

[4] http://www.mturk.com

of times that j was available as a transcription choice when i was shown as a target. The confusability of a shape with itself c_{ii} is similarly defined to be the fraction of times that i was transcribed when i was shown. The confusability values are plotted in Figure 9.

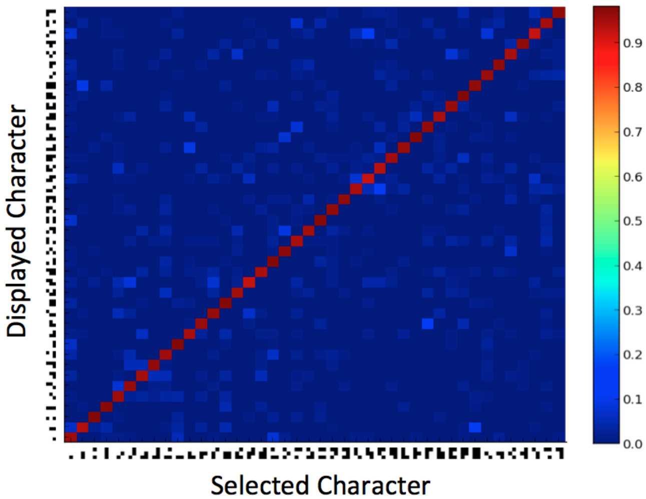
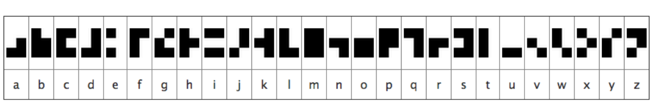


Figure 9: 3×2 block confusability matrix.

From this matrix, finding the set of 26 most distinct (least "confusible") characters was modeled as choosing the set S of size 26 so as to minimize $\sum_{i,j \in S, i \neq j} c_{ij}$. This problem is NP-hard, but a branch-and-bound search was used to quickly find the exact optimum among the $\binom{42}{26} \approx 10^{1}1$ possible solutions. The set S found by our branch-and-bound algorithm are visualized in Figure 7. The 26 selected letters, shown in black, minimize the sum of edge weights between nodes in the selected letters.

We then mapped these 26 shapes to a-z, as illustrated in Figure 10. The permutation was chosen so as to ease learning[5] by heuristically mapping the shapes to the Latin shapes that we felt they most closely resembled (upper- or lower-case).

Figure 10: Visibraille 2 alphabet



Tricolor

Since large blocks of colors were expected to be robust to blur, and since 3×2 shapes were shown to be highly "efficient" to identify [19], we tried a colored version of 3×2 shapes called Tricolor, shown in Figure 12. As a compromise between learnability and resilience, we followed the easily-learned shapes of Visibraille and assigned colors to maximally distinguish easily confused pairs of letters. We chose to use 3 colors. Here, we re-used our confusion matrix, C, to solve the following problem: partition the set of letters into three disjoint sets $S = S_1 \cup S_2 \cup S_3$ so as to minimize $\sum_{k=1}^{3} \sum_{i,j \in S_k, i \neq j} c_{ij}$. This NP-hard search over $\approx 4 \times 10^{11}$ partitions succumbed easily to exact optimization again using branch-and-bound search. The alphabet is shown in Figure 12. The selection, visualized in Figure 11, minimizes the sum of inter-color edge weights. The two-dimensional layout aims to reflect, as well as possible, this high-dimensional data set.

Consistent with prior work [22], the data indicates that pairs of "mirror image" symbols were confusable, and the optimization assigned different colors to each such pair. For example, "a" and "n" are mirror images of one another and are colored black and blue respectively. Other highly confusable pairs, such as "l" and "n," are also colored differently.

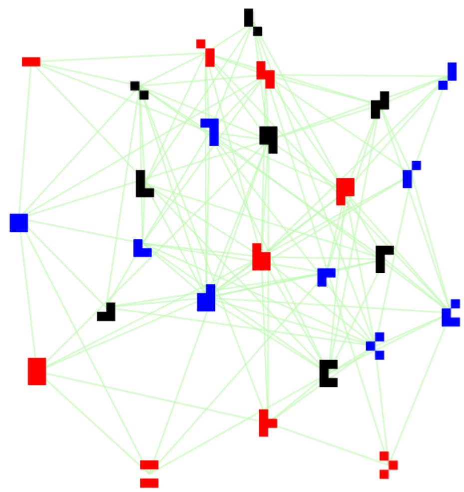


Figure 11: A force-directed layout, generated by D3 [3], of the Tricolor smartfont, attempts to locate similar pairs of letters near one another and also illustrates the colors chosen by our optimization algorithm.

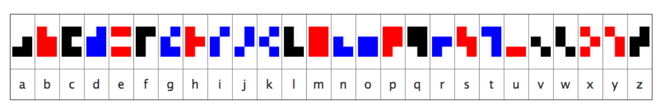


Figure 12: Tricolor alphabet

Because the characters of this font closely resemble the characters of the English alphabet, we hypothesized that it would be easy to learn and remember. Because it uses both shape and color to distinguish between characters, we hypothesize that it will also be highly readable at small sizes and blurry.

READABILITY

As we shift to using smaller screens on personal devices like smartwatches and health bands, readability at small sizes becomes increasingly important. Being able to read text that is blurry, or out of focus, is also important for both young people with certain vision problems and the aging population.

Evaluating the readability of new smartfonts is difficult because nobody currently knows how to read them. Evaluating

[5] Since evaluation was performed on random letter sequence, and since participants were not asked to learn this mapping, the selected mapping is of no consequence to the numerical results.

smartfont readability in comparison to Latin is further complicated by the fact that the test population are highly experienced at reading and identifying Latin characters. Even with training, we cannot reasonably expect our test population to accumulate a comparable amount of experience with a new smartfont over the course of a study. Our evaluation method does not require training people to read smartfonts.

Experimental Setup

Our readability experiments consist of showing participants a target string, and asking them to select the matching string from a list of strings. The targets were random strings of length five, roughly the average word length for English. Each question came with four possible answer choices. One of the answer choices was the same five-letter string as the target. The other three answer choices matched the target in four out of five characters, with one random replacement.

Because we are particularly interested in readability at small sizes and with blur, we simulated reading conditions that varied in terms of size and blurriness. Instead of screening with complex vision tests and asking people with specific blurry vision conditions to read normally printed text, we blurred text and asked people with any type of vision to read it. Whether the blurring occurs on the screen or in the visual system, the perceptual effect is similar.

Size was manipulated by rendering strings of characters at different sizes. Determining a metric for font size applicable to a diverse set of fonts is challenging. When evaluating text size, vision scientists typically use visual angle, which refers to the angle formed by the bottom of the text, the viewer's eye, and the top of the text. Typographers prefer to use the physical print size of characters. [11] Quantifying size is further complicated by variance in both height and width of characters within fonts. Because we are comparing radically different fonts, we use text area to measure text size. Text area includes the white space between and around characters that is required to render the text and cannot be occupied by surrounding text.

Blurriness was manipulated by applying a Gaussian blur filter, which replaces each pixel with a weighted average of nearby pixels. A large radius creates highly distorted images and mimics severe presbyopia for normally-sighted people, while a small radius leaves images largely intact.

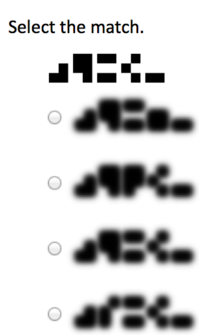


Figure 13: Sample readability task with font Visibraille.

We used a within-subject design with each participant answering matching questions for a single smartfont and for Latin. The target image was presented at decreasing sizes, with three questions at each size for both fonts. The blur radius was fixed throughout each experiment.

We recruited participants through Amazon's Mechanical Turk platform.[6] We ran two main experiments. Our first experiment compares all five smartfonts at a fixed blur of 3.5. We had 154 participants: 69 female, 81 male, and 4 other. Ages ranged 18-72, with an average of 35. 32 evaluated Polkabet; 30 evaluated Visibraille 2; 31 evaluated Visibraille; 33 evaluated Tricolor; and 28 evaluated Logobet. 69 were wearing glasses during the study, and 85 were not.

Our second experiment compares Tricolor at three different blurs. It involved 104 participants: 37 female, 64 male, and 3 other. Ages ranged 20-69, with an mean of 36 years old. Of these, 36 saw a blur of 2.5; 33 saw a blur of 3.5; and 35 saw a blur of 4.5. 41 participants were wearing glasses during the study, and 63 were not.

Results

It is not obvious how to compare the readability of smartfonts, especially since experiments were performed "in the wild" with users with varied screens and software. To address this, we compare, for each participant, the smallest sizes they can read the Latin font and the smartfont which was being evaluated. Since each experiment involved the Latin font and a single smartfont, this enabled us to compare, on an individual-by-individual basis, the Latin letters to those of the smartfont under the same conditions.

Since participants have years of experience reading Latin text and no experience reading our smartfonts, we can conclude little if they can read the Latin text better than the smartfonts. However, if, for some reason, they can read the smartfonts better, this strongly suggests that with practice the smartfonts would be even more readable than Latin text. We also caution that this data offers little value in comparing two different smartfonts, since again one of them might benefit significantly more from training than the other.

We define a *Minimal Reading Area* (MRA) for font f, MRA_f, which is specific to the participant (and blur). Note that in our experiment we asked three questions for each font at each size. As we decrease the size, the first size for which a participant makes a majority of errors (2 out of 3), we say that they failed to read at that size and define the MRA to be just-larger area used in the three questions prior to that size. Although participants were asked to continue attempting to read at smaller sizes, this further data was not used in any way in the analysis because it typically reflected random guessing. We also exclude the data from participants who failed to read at the largest size, since it is likely they did not understand the instructions or were guessing. It will be convenient to consider the log-MRA since a constant difference in log-MRA reflects a constant factor change in size. The Empirical Distribution Function of the log-MRA is shown in Figure 14.

[6] http://www.mturk.com

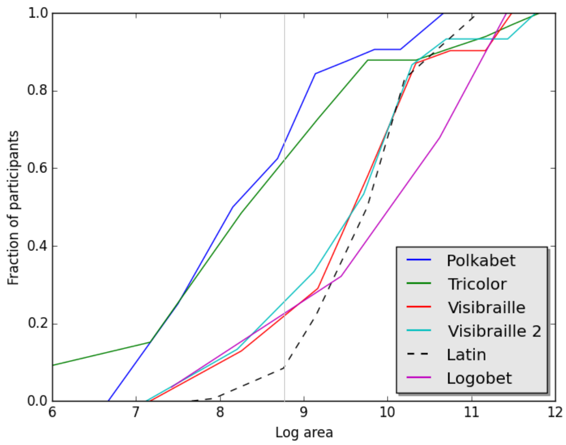


Figure 14: The (smoothed) Empirical Distribution Function of log Minimum Reading Area (at blur of radius 3.5 pixels) for the fonts: y is the fraction of participants whose log Minimum Reading Area (in that font) was larger than x.

It is hard to interpret the meaning of an area due to the wide variation in parameters. Nonetheless, to get some intuition, a Facebook post on a Chrome desktop[7] web browser today appears in a font whose full ascender-to-descender height is 13 pixels, which corresponds to a log area 8.75 (see the vertical line in Figure 14) in our experiments. With the interpolation in Figure 14, this suggests that only 8% of participants could read this size text, at our blur, in the Latin font while over 60% of the participants could read Polkabet and Tricolor fonts.

To quantify performance by a meaningful number, bounded by a confidence interval, we define the *log-score* (LS), for each experiment to be the logarithm of the ratio of the MRA for Latin to the MRA for the smartfont f in question, or equivalently,

$$\mathrm{LS}_{Latin,f} = \lg \frac{\mathrm{MRA}_{Latin}}{\mathrm{MRA}_f} = \lg(\mathrm{MRA}_{Latin}) - \lg(\mathrm{MRA}_f),$$

where lg denotes base 2 logarithm. A log-score of 0 means that the participant read Latin and the smartfont at the same size, a log-score of 2 would correspond to the participant being able to read the smartfont at $1/4$ the size of Latin. Note that our experiment is inherently one-sided: upper bounds on log-score do not bound the readability of the smartfont *after training*.

For the Tricolor font, 26 of the 33 participants (79%) had positive log-scores, meaning that they read the smartfont at a smaller size than Latin, and 18 of the 33 (55%) had log-scores greater than 1, meaning they read the smartfont at least twice as small as Latin. The sample mean log-score was 1.28. A histogram of the log-scores is displayed in Figure 15. The wide variance in this histogram means that some users might benefit significantly more than others from adoption.

[7]Most of our mechanical turk participants were using desktop, not mobile, browsers. See `http://facebook.com` and `http://google.com/chrome`.

Font	CI lower-bound	Mean log-score
Polkabet	0.78	1.30
Tricolor	0.62	1.28
Visibraille	-0.23	0.14
Visibraille 2	-0.32	0.14
Logobet	-0.56	-1.03

Table 1: Mean and 95% simultaneous (one-sided) confidence interval lower-bounds. A log-score of 1 corresponds to readability at half the size of Latin.

For *simultaneous* 95% post-hoc confidence intervals for 5 fonts, we choose what would normally be 99% confidence intervals bounding each (the union bound on the 1% failure probability of each estimate then implies 95% confidence). Since our test are inherently one-sided, as mentioned, we use simultaneous one-sided confidence intervals, based on mean and standard deviation. The results are displayed in Table 1. Only the confidence intervals for Tricolor and Polkabet are entirely positive.

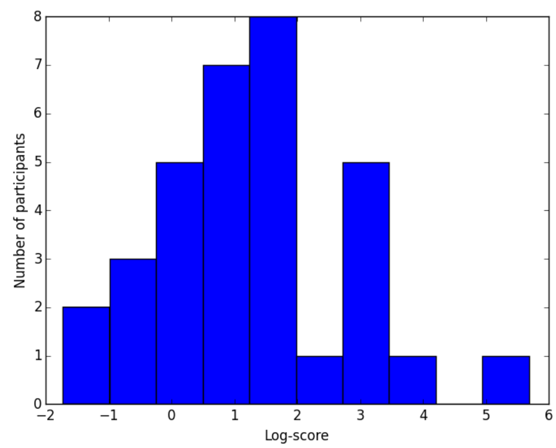


Figure 15: Histogram of log-scores for the Tricolor at blur of radius 3.5 pixels. A log-score of 3, for instance, indicates that Tricolor was readable 8 times smaller than the Latin font, for that participant.

To see how performance would vary as we change the blur parameter, we compared Tricolor versus Latin at three different blur radii. The results at radii 2.5 pixels, 3.5 pixels, and 4.5 pixels, were all greater than 0 with statistical significance, though the differences were not statistically significant. The mean log-scores of 1.17, 1.28, and 1.41, respectively, suggest a possible increasing trend.

LEARNABILITY

In order for our smartfonts to be usable, they must be learnable. To evaluate their learnability, we designed an online learning system and tracked partcipants' progress learning our fonts. The results showed that after reading a couple thousand sentences, our fonts are read at speeds that are the same order of magnitude as that of Latin.

Learning Site Design

Our online learning tool provides a tutorial about the font, flashcards for drilling the meaning of individual characters, and simple yes/no questions in the font.

Tutorial

We provided a brief tutorial explaining each font. The tutorial presented 1) the mapping of characters from the new font to Latin (i.e. "English") characters, 2) a description of the organization of the new font's characters, and 3) examples of words in the new font with their Latin equivalents.

The tutorial was presented in the main menu of our learning site. The main menu welcomed participants each time they visited the site, first introducing them to the font and refreshing their memory each time they logged back in. Participants could return to the main menu at any point to view the description. The main page also provided a chart of the participant's performance for over time for self-tracking.

Yes/No Questions

Our site provided short yes/no questions to help participants practice and learn their new font. These questions were generated via crowdsourcing, including questions from MindPixel[17] augmented with questions we gathered from Amazon Mechanical Turk workers. We screened the questions for inappropriate content before releasing them. In total, we had 2739 questions: 1245 with a positive answer and 1494 with a negative one. The questions were generally fun and entertaining. Examples include "Is the moon out at night?" and "Are you a celery?".

Figure 16: Sample yes/no question for font Logobet.

The learning site asked the yes/no questions in the new smartfont, as demonstrated in Figure 16. After receiving an answer, the site showed the question in both the smartfont and plain English so that they could verify their reading of the question. It also gave feedback on correctness.

We provided a "cheatsheet" that participants could use while answering the yes/no questions. The cheatsheet showed the mapping of the new font's characters to standard Latin characters. The Polkabet cheatsheet also provided mnemonics. To view the cheatsheet, a participant could click on a link above the question. The cheatsheet overlayed the yes/no question page, so that the participant could not continue answering questions while viewing the cheatsheet. This design forced participants to rely on their memory for answering questions with prompting from the cheatsheet, rather than visually "looking up" each character with the cheatsheet.

Flashcards

To further help participants memorize their smartfont, we provided flashcards of the font's characters. The flashcards present a single encoded character at a time, and prompt the participant for the Latin character equivalent. Mistaken characters are repeated until the participant is no longer making any mistakes. Participants were free to make use of the flashcards at any point during the study.

Experimental Setup

To evaluate our smartfonts' learnability, we recruited people to use our site to learn smartfonts through Amazon's Mechanical Turk platform.[8] We had 23 participants in total. Each participant was assigned randomly to a single font: 8 to Polkabet, 6 to Tricolor, and 9 to Logobet. Varying numbers for each font are due to participant dropout during the study. Participants chose how long they spent on our site. They typically 2-3 hours per day on our site over the course of about a week, and were compensated for the practice questions they answered.

When participants first visited the site, they set up an account so that they could return to use the site at any time. The site was in operation for about one week. Participants were compensated for the yes/no questions that they answered, but were free to make use of the flashcards, cheatsheet, or main menu at any time. One in every 10 yes/no questions was displayed in Latin characters for comparison. We recorded the time it took participants to answer the yes/no questions in both their smartfont and in Latin characters. We also recorded their use of the cheatsheet and flashcards throughout the study. Participants were free to provide open-ended feedback through a form on the site at any point during the study.

Results

To evaluate a participant's speed reading a smartfont, we calculated the ratio of the time it took them to answer each yes/no question in the smartfont to the average time it took them to answer our control questions in the Latin font. A value of 1 means that it takes the person the same amount of time to answer questions in the encoding as it does with Latin characters, a value of 2 means it takes them twice as long, and so on. All participants held over 95% accuracy in answering the encoded questions, i.e., they were not guessing.

Trends Across Fonts

Figure 17 shows the general trends across our smartfonts. Tricolor exhibits the easiest learning curve, followed by Logobet and then Polkabet. After $2,000$ questions, participants learning Tricolor were reading a median of 2.1 times slower than they did in Latin; participants learning Logobet were 5.2 times slower; and participants learning Polkabet were 6.7 times slower. We ran an unpaired t-test to determine whether the differences in response time across fonts was significant after 2000 questions. We found a statistically significant difference between each pair of fonts: Polkabet and Logobet ($t(8498) = 10.6623, p < 0.0001$), Polkabet and Tricolor ($t(7998) = 4.0640, p < 0.0001$), and Logobet and Tricolor ($t(8498) = 2.6588, p < 0.008$).

[8] http://www.mturk.com

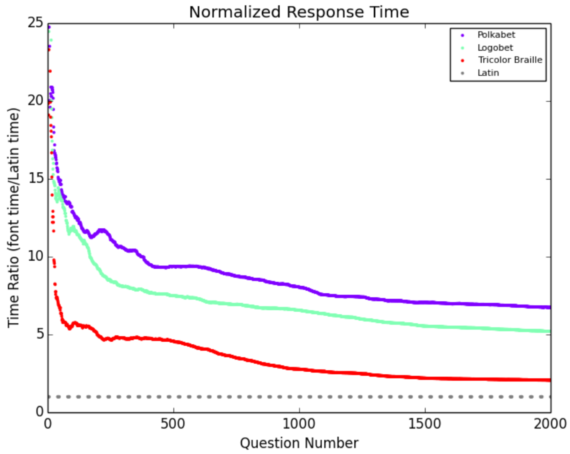


Figure 17: Response times to yes/no questions over time, normalized by each person's average response time in Latin characters. Each point is the median of a sliding window of averages across all participants to remove outliers.

Individual Learning Curves

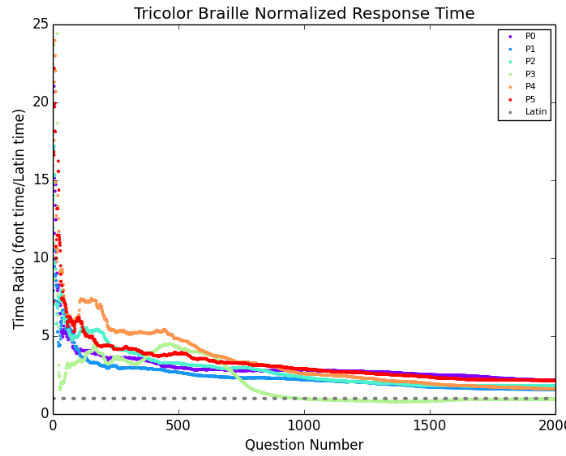
There was some variation in learning curves between participants learning the same smartfont, as hown in Figure 18. Notable outliers are P3 for Tricolor and P6 for Logobet. These two participants learned their respective fonts extremely quickly–they became as quick at reading the smartfont as they were at Latin after only around 1,000 questions. Their quick learning curves suggest the benefit of personalization–some may prefer to learn smartfonts that would challenge others, and these preferences may be individual.

Learning colors

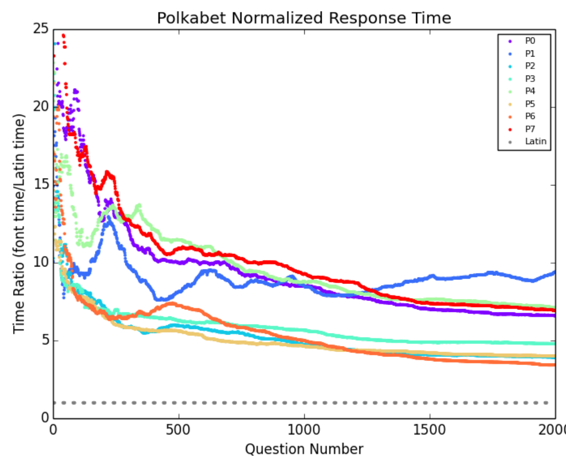
Tricolor could be read ignoring the colors. The redundant coloring, however, may be helpful when text is out of focus. Since the characters resembled the Latin alphabet, one might be concerned that people learn to read ignoring colors and then when faced with blurry text, will not remember the colors. This did not seem to be the case. Consistent with prior work on training color-grapheme synesthesia by reading books with colored letters [4], we find that participants remembered the colors of common words. In a post-test administered to seven readers of Tricolor three days after the system was shut down, we asked them to correctly identify the coloring of five common words (like "the") each on a multiple-choice question with four choices (three random colorings of the same shaped letters). The aggregate accuracy was 28/35 (80%) over these four questions, strongly indicating that they had remembered at least some of the colors.

Learning Resources

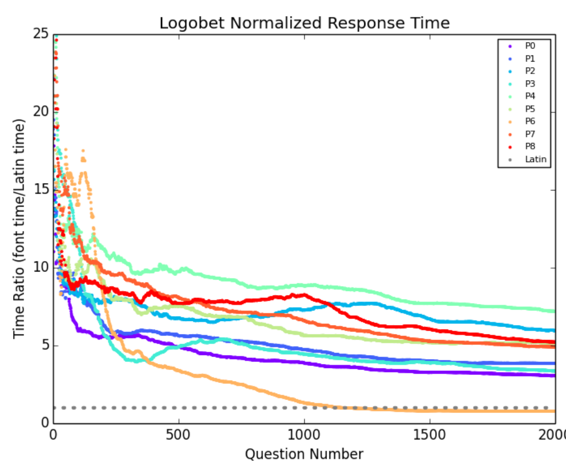
Participants used both the flashcard and cheatsheet as they learned our smartfonts. Participants learning Tricolor made more use of the learning resources than participants learning Polkabet or Logobet. It is possible that participants learning Tricolor relied more on the learning resources because their font tutorial did not include additional information beyond



(a) Normalized learning progress for Tricolor participants.



(b) Normalized learning progress for Polkabet participants.



(c) Normalized learning progress for Logobet participants.

Figure 18: Learning curves for individual participants, separated by font.

its mapping to Latin characters. We provided mnemonics for Polkabet, which likely helped Polkabet participants recall more characters independently, if slowly. Similarly, we provided a lengthy tutorial for Logobetdetailing the character organization, that likely helped participants remember the character representations. Because of its relative simplicity, Tricolor had no mnemonics or details about font organization.

Qualitative Feedback

We gave participants the opportunity to provide open-ended feedback. Their responses indicated that they largely enjoyed learning and reading our fonts. The majority remarked that their experience was "fun." Several participants compared the process of decoding and answering questions to solving puzzles. One participant explained, "I thought this was extremely fun and interesting because I love puzzles, especially ones that deal with words." Another wrote, "someone should find a way to turn this into an Android game. ." At the end of the experiment, one participant contacted us, asking if they could continue using our site to practice their font. The positive experience that our participants describe suggests that people would enjoy continuing to learn, read, and use smartfonts.

Participants also indicated that they felt they were learning. One participant explained, "It was a lot to take in at first, but I felt my responses becoming more intuitive." Another explained, "I thought this was super hard in the beginning but on the last couple I actually was reading them as though I was seeing the letters." Coupled with our learning curves, this qualitative feedback suggests that at least some people can learn to read smartfonts fluently.

DISCUSSION AND FUTURE WORK

There are several limitations to the current work. First, we were unable to control for screen type, screen resolution and distance from the viewer to the screen. This was done because crowdsourcing enabled us to rapidly experiment with a number of different fonts. Hence, it would be beneficial to reproduce these results in a laboratory setting with users with presbyopia. Second, we do not currently offer users who learn a smartfont the ability to use it in any meaningful way. This could be crucial to adoption.

Privacy is an additional benefit of smartfonts. "Substitution ciphers" which encrypt text by replacing each letter with a symbol, have been used by da Vinci in mirror-writing [], by Union prisoners in the Civil War, and by children using Captain Magneto encoder rings and other creative encodings of their own design. Privacy can be especially valuable on smartwatches, where potentially embarassing personal communications may pop up without warning, visible to anyone sufficiently close.

Costs and durability may also be affected by font. For instance, the seven-segment display of digits, common among digital alarm clocks and other electronics, is less expensive and has fewer pieces that may fail than a high-resolution screen. Aesthetics are, of course, a major consideration which we leave to future work since they are difficult to design.

In the future, smartfonts could be tailored to an individual's eyesight or display screen. Each person is unique, and a wide variety of vision conditions exist. We imagine a system that evaluates a person's vision and generates optimized smartfonts on-the-fly. Such a system would require learning a model of how vision relates to font readability. Just as many Southeast Asian scripts have rounded letters because straight lines would tear the palm leaves on which they were written [16], smartfonts could also tailor to the screens that display them.

Smartfonts could also be generalized to other character systems besides Latin. For example, we can develop smartfonts for the Hebrew alphabet or Chinese characters. Many East Asian scripts are read top-to-bottom, so any smartfont involving kerning would need to support combining adjacent characters vertically. The size of character sets can also vary enormously. For example, there are over 50,000 Chinese characters. A smartfont for such a large character set would likely need to take advantage of language or character structure.

CONCLUSION

In this work, we introduce smartfonts, scripts that completely redesign the written alphabet with the purpose of improving the reading experience. We do not claim to have created the best smartfonts or even optimal smartfonts for reading blurry text, but we have hopefully demonstrated that there is room for improvement over the millenia-old letters in use today.

Smartfonts have many potential benefits: improved readability for various reading conditions, increased privacy, heightened aesthetics, and a "cooler" reading experience. Allowing interested users to opt-in, smartfonts do not require alphabet reform. They also do not require new hardware or software, but are deployable on existing platforms. As our experiments showed, it is possible to learn to read them with a reasonable amount of practice.

We developed a set of smartfonts to be readable at small sizes and with blur. As we move into an age of personalized electronics, screen sizes shrink and enabling people to read on small screens becomes increasingly important. Similarly, making it possible for people to read text that looks blurry is also important. As people age, their eyes lose the ability to focus, and glasses are not always convenient or available. Similarly, a variety of low-vision conditions exist and cannot be corrected with glasses. Our readability experiments provide evidence that our smartfonts are indeed more readable for a range of small sizes and with varying amounts of blur.

We also presented experimental designs for evaluating 1) the learnability and 2) the readability of smartfonts under various reading conditions. We evaluated learnability by teaching smartfonts through an online system that provided a tutorial, encoded yes/no questions with a cheatsheet, and flashcards, and tracking yes/no question response times. We evaluated readability through a novel experimental setup that allowed us to evaluate readability under various reading conditions without training people to read the smartfont.

REFERENCES

1. M Alpern. 1969. Accommodation and the Pupil. *The Eye* 3 (1969).
2. Ann Bessemans and others. 2012. *Letterontwerp voor kinderen met een visuele functiebeperking*. Ph.D.

Dissertation. Academy of Creative and Performing Arts, Faculty of Humanities, Leiden University.

3. Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D^3 data-driven documents. *Visualization and Computer Graphics, IEEE Transactions on* 17, 12 (2011), 2301–2309.

4. Olympia Colizoli, Jaap MJ Murre, and Romke Rouw. 2012. Pseudo-synesthesia through reading books with colored letters. *PLoS One* 7, 6 (2012), e39799.

5. Paul A David. 1985. Clio and the Economics of QWERTY. *The American economic review* (1985), 332–337.

6. Cagatay Demiralp Demiralp, Michael S Bernstein, and Jeffrey Heer. 2014. Learning Perceptual Kernels for Visualization Design. *Visualization and Computer Graphics, IEEE Transactions on* 20, 12 (2014), 1933–1942.

7. Alexander Duane. 1912. Normal values of the accommodation at all ages. *Journal of the American Medical Association* 59, 12 (1912), 1010–1013.

8. Joyce Farrell, Jiajing Xu, Kevin Larson, and Brian Wandell. 2009. 47.2: Visual Preference for ClearType Technology. In *SID Symposium Digest of Technical Papers*, Vol. 40. Wiley Online Library, 702–705.

9. David Goldberg and Cate Richardson. 1993. Touch-typing with a stylus. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. ACM, 80–87.

10. Marcel A Just, Patricia A Carpenter, and Jacqueline D Woolley. 1982. Paradigms and processes in reading comprehension. *Journal of experimental psychology: General* 111, 2 (1982), 228.

11. Gordon E Legge and Charles A Bigelow. 2011. Does print size matter for reading? A review of findings from vision science and typography. *Journal of Vision* 11, 5 (2011), 8.

12. Gordon E Legge, Gary S Rubin, and Andrew Luebker. 1987. Psychophysics of readingâĂŤV. The role of contrast in normal vision. *Vision research* 27, 7 (1987), 1165–1177.

13. Gordon E Legge, Gary S Rubin, Denis G Pelli, and Mary M Schleske. 1985. Psychophysics of readingâĂŤII. Low vision. *Vision research* 25, 2 (1985), 253–265.

14. William Lidwell, Kritina Holden, and Jill Butler. 2010. *Universal principles of design, revised and updated: 125 ways to enhance usability, influence perception, increase appeal, make better design decisions, and teach through design*. Rockport Pub.

15. Stan J Liebowitz and Stephen E Margolis. 1990. The fable of the keys. *Journal of Law and Economics* (1990), 1–25.

16. C.P. Masica. 1993. *The Indo-Aryan Languages*. Cambridge University Press. https://books.google.com/books?id=J3RSHWePhXwC

17. Chris McKinstry, Rick Dale, and Michael J Spivey. 2008. Action dynamics reveal parallel competition in decision making. *Psychological Science* 19, 1 (2008), 22–24.

18. Robert A Moses. 1987. Accommodation. *Adler's Physiology of the Eye* (1987), 291–310.

19. Denis G Pelli, Catherine W Burns, Bart Farell, and Deborah C Moore-Page. 2006. Feature detection and letter identification. *Vision research* 46, 28 (2006), 4646–4674.

20. Andrew Trusty and Khai N. Truong. 2011. Augmenting the Web for Second Language Vocabulary Learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 3179–3188. DOI: http://dx.doi.org/10.1145/1978942.1979414

21. Stan Walker, Phil Schloss, Charles R Fletcher, Charles A Vogel, and Randall C Walker. 2005. Visual-syntactic text formatting: A new method to enhance online reading. *Reading Online* 8, 6 (2005), 1096–1232.

22. Alexandra Wiebelt. 2004. Do symmetrical letter pairs affect readability?: A cross-linguistic examination of writing systems with specific reference to the runes. *Written Language & Literacy* 7, 2 (2004), 275–303.

23. Chen-Hsiang Yu and Robert C Miller. 2010. Enhancing web page readability for non-native readers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2523–2532.

24. Shumin Zhai and Per-Ola Kristensson. 2003. Shorthand writing on stylus keyboard. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 97–104.