Speaking Through Pictures: Images vs. Icons

Xiaojuan Ma, Jordan Boyd-Graber, Sonya Nikolova and Perry Cook
Princeton University
35 Olden Street, Princeton, NJ 08540 U.S.A.
{xm,jbg,nikolova,prc}@cs.princeton.edu

ABSTRACT
People with aphasia, a condition that impairs the ability to understand or generate written or spoken language, are aided by assistive technology that helps them communicate through a vocabulary of icons. These systems are akin to language translation systems, translating icon arrangements into spoken or written language and vice versa. However, these icon-based systems have little vocabulary breadth or depth, making it difficult for people with aphasia to apply their usage to multiple real world situations. Pictures from the web are numerous, varied, and easily accessible and thus, could potentially address the small size issues of icon-based systems. We present results from two studies that investigate this potential and demonstrate that images can be as effective as icons when used as a replacement for English language communication. The first study uses elderly subjects to investigate the efficacy of images vs. icons in conveying word meaning; the second study examines the retention of word-level meaning by both images and icons with a population of aphasics. We conclude that images collected from the web are as functional as icons in conveying information and thus, are feasible to use in assistive technology that supports people with aphasia.

Categories and Subject Descriptors
H.5.2 [**User Interfaces**]: Evaluation/methodology

General Terms
Experimentation, Human Factors

Keywords
Aphasia, Visual Communication, Computerized Visual Communication, VIC, C-VIC

1. INTRODUCTION
In the course of a day, an active individual has many reasons to communicate: reading a newspaper, visiting a doctor, shopping, or debating sports. Daily communication not only enriches peoples' lives but also expands their vocabulary and enhances their language skills. However, for people with aphasia, approximately one million in North America alone [16], who are afflicted to varying degrees in their ability to produce or comprehend verbal or written languages [9], getting information from the external world is extremely difficult. Any written or spoken communication, from small talk to news on TV or the Internet, is challenging for aphasia afflicted individuals. It is therefore not surprising that people with aphasia tend to withdraw from social activities and suffer from isolation and depression. Many devices (e.g. Lingraphica [12], DynaVox [6], and PRC devices [17]) have been developed to assist individuals for communication and language rehabilitation. However, their efficacy is limited by the size of their icon-based vocabulary.

ASSETS'09, October 25–28, 2009, Pittsburgh, Pennsylvania, USA.

Icons are traditionally used in these assistive technologies because of their simplicity and clarity. Because they are painstakingly designed and engineered, icon-based systems remain unable to handle a large number of topics to the depth that a user might want. For example, the current icon-based systems are not flexible enough to handle the ever-changing topics of politics nor deep enough to accommodate the local and regional terms that appear in newspapers. It is also hard for other people to use these structured collections of icons to interact with people with aphasia; a caregiver may want to introduce a new medicine, but find it impossible to describe the medicine with existing icons and therefore need to add new icons to the system. This can lead to a frustrating and slow exchange as caregivers first search for desired icons and then create and add new ones.

Web images which are ubiquitous, abundant, and cheap to access can possibly assist information delivery to people with aphasia. The abundance and diversity of web images guarantees that a unique one can be found that captures the specific entity that is wanted for an image-based communication. For example, to present a new medicine to an individual with aphasia, the caregiver can perform a web search for images of the product as well as the symptoms that it targets. However, compared to the icons which were carefully crafted to support communication, web images have inherent shortcomings. Their complex content may evoke several meanings, and different people might interpret them differently. Their quality and resolution varies significantly, and relatively few are labeled meaningfully.

To investigate this use of images, we conducted two studies to explore whether web images are as effective as a set of icons created specifically for people with aphasia. We concentrated on an aphasic individual's perception of the pictorial representations instead of how they would use such representations in communicating, because, in practice, caregivers or family members involved in the care of the patients are expected to be the individuals who will select the images from the web and add them to the aphasic's assistive language system.

The first study using 24 senior citizens, served as a pilot study for comparing the overall efficacy of web images to icons. The study confirmed our hypothesis that the performance of web images was statistically similar to that of specially-designed icons for a population with similar demographics to our target group. It provided a valid foundation for testing the use of web images further with people with aphasia. The second study was conducted using 50 people with aphasia, which is the largest (by an order of magnitude) subject pool of this population compared to related research (e.g., [3], [2], and [22]). Results showed that web images are comparable to purposely designed icons in illustrating daily concepts for people with aphasia. Our findings therefore indicate that web images can be beneficial in assistive communication as an extension of traditional iconic collections.

The paper is organized into eight sections. In Section 2, we use existing research to support our premise that web images can be used to supplement communication systems for people with aphasia. Section 3 defines our first experiment comparing the efficacy of images vs. icons. Section 4 analyzes the results from Experiment 1 and leads to Experiment 2, which examines the efficacy of icons vs. images for people with aphasia (described in Section 5). A results section immediately follows in Section 6. Finally an overall discussion of the implications for the design of assistive technology for aphasics is presented in Section 7 followed by a conclusion in Section 8.

2. BACKGROUND WORK

Research shows that aphasic individuals retain abilities such as identification, sequencing, generalization, and meaning association, which can be used to improve communication via visual prompts [21]. Because of these findings, pictorial representations of concepts have been widely used in the design of assistive devices for people with language disorders.

2.1 Perception of Use of Icons

Icons are a type of practical, minimalist art. When the task of building visual vocabularies was one of simply moving from concept to image, with no ready-made alternatives available, they were a logical choice for presenting concepts. Icons retain contours, which are important for the perception of shapes but omit visual cues such as texture and color that are not as relevant to interpretation. Research has shown that line drawings depict the essentials of size, shape, and location of their subjects [19].

Developed in the eighties, VIC (VIsual Communication) is an icon-based language for individuals with aphasia. It consists of a set of black and white symbols drawn on index cards. Each symbol represents a meaningful concept that a person with aphasia might wish to express. VIC was later enhanced and implemented on a computer based interactive system known as C-VIC (Computerized VIsual Communication System) [20]. One computer application based on C-VIC is Lingraphica [12], a commercially available system for aphasic individuals. The experiments presented in this paper use icons from Lingraphica.

Lingraphica's vocabulary, like C-VIC, includes proper and common nouns, prepositions, conjunctions, interjections, adjectives, location, and activity particles as well as request forms like "Wh" words ("where," "what," "when'" etc.) and some additional symbols (e.g. "?"). Unlike C-VIC, Lingraphica also includes animated icons for verbs. It has a vocabulary of over 5000 words. Users search through a hierarchy to construct sentences and practice speaking the written translation of the icons. Results from an evaluation of VIC and C-VIC usage found that people with aphasia communicate more effectively using the symbol language than using English [20]. However, Lingraphica has limitations in that a vocabulary of 5000 symbols is not enough to represent the breadth and depth of concepts people need. Lingraphica developers have realized this limitation and have introduced mechanisms to allow users to upload their own photos to represent the concepts of an aphasic's daily life.

2.2 Perception of Use of Images

Images are realistic and provide more interpretation cues than icons. Image attributes such as color, contrast, and segmentation influence the perception of the size and depth of an object. The distinguishing property of many real-world objects is also their distinctive color (e.g. lemon vs. lime). Likewise, texture is useful in visualizing a surface; orientation, density, contrast, and size of the texture pattern are effective cues for recognizing the image. The presence of color, texture, and visible discontinuities help observers identify basic features of the viewed image such as shape, orientation in space, light source, and size without the need for training or written annotation [11].

Images are used to help people cope with memory, language, or speech deficiencies that accompany aging. Investigations show that communication is enhanced by pictures, especially realistic pictures, not only for people functioning at normal cognitive levels but also for people with cognitive degeneration or impairment. Danielsson [4] suggests that using images for communication can lead to significant improvements in learning for people with cognitive disabilities. Research as well as anecdotal evidence also shows that people with aphasia already use photographs to share memories, experiences, and information with friends, family, and strangers [14]. This suggests that pictures might be a viable replacement for the icon-based images currently used by those with aphasia.

2.3 Integrating Web Images

Although many systems for assisting individuals with aphasia have introduced photos as a supplement to their icon-based vocabulary (e.g. [3], [22]), they only provide a limited set of additional images and, thus, the key limitations with using icons as the core vocabulary remain. Furthermore, the images employed in the existing assistive devices are mostly photographs taken by the system users, which, similar to having artists design icons, has only shifted the burden of generating the needed language representations from the icon designers to the system users. One alternative to breaking this "effort" bottleneck is to utilize the abundance and diversity of internet images. In this case, images still need to be found, but this is faster than image generation. In addition, the abundance of web images available makes selecting the most appropriate image easier.

For example, online image repositories such as Flickr and Picasa provide hundreds of thousands of images with associated word tags. Other resources that provide image-text associations which can be used to provide images for an aphasic support system database include image search engines like Google Images. A problem with using the tagged images in Google is one of search context. Image-keyword associations that are brought up by search engines are based on surrounding text and may not accurately depict the tag. Using images tagged via the ESP Game [23] might be better candidates for our purposes, especially since the game ensures that each image has elicited the associated word multiple times from different people.

ImageNet [5] is an image data\base that has organized tens of millions of web images in WordNet [7] hierarchies. This collection could also be used for content in an aphasia support system database. On average, 1,000 images are assigned to each of the selected 100,000 synonym sets (mainly nouns) in WordNet. ImageNet provides a quick and free access to numerous web images that have a precise association with concepts. An online multimedia language assistant that uses a similar structure to that being used for people with aphasia is already under development by the authors of this paper [24]. The system allows users to look up pictorial representations of words in a pop-up dictionary. This could be adapted to a system that supports people with aphasia.

2.4 Comparison of Icons and Images

Both anecdotal reports from speech language therapists plus systematic investigation suggest that, unlike written words, cultural differences do not prevent people from identifying images (photographs) and icons (line drawings) in a similar way, although different cultures may resolve ambiguities differently [15]. This suggests that no training is required for understanding images; in fact, experiments have found that infants are able to understand images without training [10].

Ryan et al. [18] compared the speed of recognition of four different modes of representation, realistic line drawings, cartoon like line drawings, shaded drawings, and photographs to assess the effectiveness of different ways of presenting objects and poses. His results showed that cartoons were perceived in the shortest time whereas line drawings took the longest, while photographs and shaded drawings took about equal amounts of time. Currently, very little research has been done on comparing the effectiveness of different visual illustrations when concepts rather than concrete objects are represented, e.g., "craving," "thinking hard," "disdain," etc.

Despite the increased use of images for communication, no research has focused on how the irregular quality and inconsistent information complexity can influence perception of an arbitrary selection of images from the web in contrast to carefully and deliberately designed icons. In order to provide a basis for using web images instead of icons as the principal visual representation in assistive technologies for people with aphasia, we first need to conduct a set of studies illustrating that the web images are, on the whole, as effective as icons for communication. Thus, in this work, we have directly compared the effectiveness of images and icons for communication with our target population, 50 aphasic individuals. This large subject pool is uncommon in research on assistive technologies because it is much more time consuming to run studies using language deficient subjects and also more difficult to obtain subjects. Nevertheless, it is important to run comparisons such as these with a larger number of subjects for two reasons. First, our aim is to investigate if images are equal to icons in efficacy. As such, we need to look at the power of the test when conducting statistical inferences that show no difference. This approach demands a larger N. In addition, because of the wide variability in the impact of the traumatic event that has caused aphasia and also in the subjects themselves who may be affected cognitively in other ways, a larger N is needed to handle this variance.

Unlike earlier studies, the vocabulary we examined was not restricted to concrete nouns, but extended to illustrate different parts of speech, including concrete and abstract nouns, verbs, and adjectives. Moreover, our source of images is relatively random. It is compiled from the Internet and not painstakingly annotated. The hope is that this work will demonstrate that this cheap and plentiful source of images is "good enough" to use in communication systems developed for aphasics.

3. EXPERIMENT 1 DESCRIPTION

The goal of Experiment 1 was to compare the efficacy of web-based images to Lingraphica icons. To do so, we asked senior citizens to interpret visually illustrated words and phrases in which either images or icons were used. We used senior citizens reasoning that they paralleled our target population in cognitive skills and computer experience. Certainly we can say this definitively about computer experience, since people with aphasia typically come from this group. However, we are not as confident about similar cognitive skills which a stroke might have impacted.

Our procedure for eliciting the meanings of phrases worked as follows. We collected a set of sentences relevant to our target group, transformed them into a mixture of either words and images or words and icons and then asked subjects to reformulate the original intended meaning. We then coded the effectiveness of the reformulation and compared the scores of imagery to icon-based representations.

3.1 Turning Sentences into Pictures

To run the experiment, we chose 25 phrases: 14 from blogs of senior citizens featured on the Ageless Project [1] (Table 1, with target words underlined) and 11 (Table 2) from daily phrases suggested by speech-language pathologists (SLPs). For the first set, we often had to paraphrase the sentences to make the transformation into pictures easier. While the paraphrased sentences are by design, transactional, i.e., carrying message content information, the originals often employ idiosyncratic, interactional content that promotes a closer connection with the reader. While some information is lost, this study focused on the information-carrying ability of pictures rather than the emotional component. Paraphrasing also helped to increase the ratio of pictures to words in the phrases because we replaced only words (nouns, verbs, and adjectives) that had both icons from Lingraphica and images from the ESP Game database. This preference, thus, biased us toward more of the common terms that were in both resources.

Table 1. Sample Paraphrased Sentences from Blogs

1. I've wished and <u>dreamt</u> of having a talent to <u>perform</u>.
2. We <u>started</u> to have <u>car</u> <u>trouble</u>, but we weren't too <u>far</u> from a <u>service station</u>.
3. My <u>eye</u> <u>drops</u> <u>fell</u> in the <u>toilet</u>.
4. I'm <u>buying</u> a <u>vacation</u> <u>home</u>.

Table 2. Sample Sentences Suggested by SLPs

1. Can you <u>help</u> me?
2. Can you <u>write</u> that for me?
3. Where is the <u>bathroom</u>?
4. Where is the <u>bus</u> <u>stop</u>?

For each of the selected words, we chose the appropriate image using the following process. We grabbed all ESP images that were tagged with the phrase word and its synonyms in the labels, ruled out images where the word represented a different sense than that conveyed by the image or where the image was of low quality, and then had three judges vote separately for the best image. Online image search results were added if not enough images were retrieved from the ESP data. Likewise, the best icons from Lingraphica were chosen. All pictures were enlarged to 124 by 164 pixels for icons and an average of 200 by 160 pixels for images without damaging the resolution to compensate for the possible visual degeneration of our elderly subjects.

Sentences spanned a range of lengths. The shortest sentence was three words and the longest, 15. Substitution varied with the ratio of the number of pictures to the sentence length ranging from 0.17 to 0.57, thus ensuring that even in the most difficult sentence there remained a written context for the subjects to use.

3.2 Experiment 1 Procedure

The phrases were presented to individuals in a booklet form with one phrase per sheet. We asked the participants to write down their guess of the sentence's meaning in the given space on the bottom half of the paper. Participants were also asked to rank how hard it was to reconstruct each phrase (very hard, hard, medium, easy, and very easy). Figure 1 shows examples of image and icon testing sheets, respectively. At the top of each sheet is a place for the user to record the perceived difficulty, the middle of the sheet contains the picture substituted phrase, and the bottom has a field for the user to write the sentence in English.

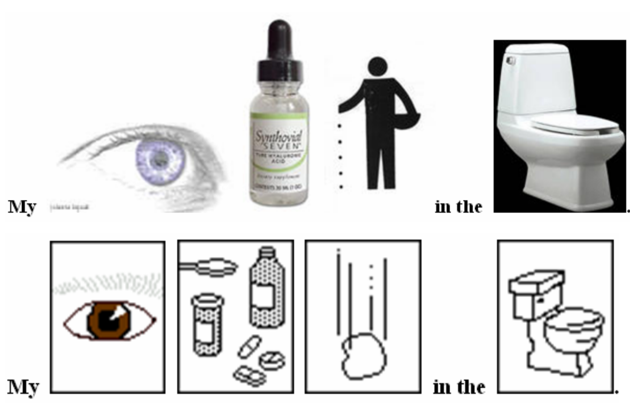

Figure 1. Sentence with images (above) and icons (below).

Twenty-five sheets with different phrases, all with either icons or images, were stapled into one booklet. Twenty-four booklets were made, half of which had icons while the other half had images. All phrases were translated into both icons and images to form these booklets. In addition, no sentence appeared twice for a subject (i.e., as an icon-based sentence and an image sentence). Thus, the experiment was a between subjects design. Subjects did not experience both images and icons to prevent potential bias toward one of the modalities from affecting their response. The order of the phrases in each booklet was randomly shuffled. At the beginning of the experiment, every participant picked a number (1-24) from a hat, and the booklet with the corresponding number was given to the participant. When finished, the investigator sealed the booklet so that only the answer part could be viewed.

3.3 Experiment 1 Participants

We recruited 24 participants from a senior center. The group consisted of seven male and 17 female subjects over the age of 60. There were five study participants in the 60 - 69 age range, eight in the 70 - 79 age range and 11 in the 80 - 89 age range. We also used a between subjects design because senior participants tire readily requiring their testing period to be short. Their attendance at the senior center was also sporadic, making it difficult to schedule them for another session.

4. RESULTS, EXPERIMENT 1

4.1 Automatic Scoring

In evaluating how well pictures conveyed the meaning of an intended sentence, we need to quantify the degree to which two pieces of text differ. This is also a problem in evaluating outputs in machine translation, so we initially used General Text Matcher (GTM) [8], which gives high scores for long runs of identical words in the source (in this case, the original sentence) and target texts (here, the sentences provided by the subjects). Considering each sentence separately and calculating the median of the GTM score for each of the sentences allows us to determine those phrases in which the image and icon interpretations were different. An analysis of variance showed significant disparity between the icon and image representation for only one phrase (Table 2, SLP 4) ($F(1, 22) = 8.358$, $p = 0.008$, $\eta^2 = 0.275$).

As in machine translation, simply using the number of words that matched is not a viable measure of the fine-grained distinctions between the meanings of words. Because the responses often included synonyms (e.g. "automobile" for the "car" pictures), hyponyms, or hypernyms, we decided to refine our investigation using a more precise error metric based on human judgment.

4.2 Human Scoring

To run the human evaluation, each of the responses was transcribed and placed into a plain text file by one of the coders without viewing the picture stimuli. Then two coders assigned scores both to every word replaced by a picture as well as the overall meaning of the sentence based on how close the subject came to recapturing the original meaning. Subjects did not have to provide a single word for each picture; it was acceptable to provide any phrasing that captured the meaning in the sentence. The coders used the scale shown in Table 3. If the coders differed by more than a point, the difference was adjudicated by a third party. Approximately 4% of word ratings and 4% of sense of sentence ratings had to be arbitrated. The median score was used in these cases. For the two coders, the exact same rating was provided 76% of the time.

Table 3. Human Scoring Scheme (example: "dog")

Score	Description	Example
5	Matches perfectly	"canine"
4	Somewhat good	"animal"
3	Moderately good	"cat"
2	Poor, but not ruled out	"fish"
1	Missing or completely wrong	"hammer"

4.3 Word-level Effects

There were 33 concrete nouns, 17 abstract nouns, 10 adjectives and 24 verbs tested in our study. Given that icons are usually pictures of things, it was fairly easy to select an icon to convey the meaning of nouns, but we often had difficulty selecting appropriate icons for adjectives and for verbs. This difficulty was also reflected in how subjects responded to images. There is a significant interpretation accuracy difference within part of speech ($F(3, 41) = 4.027$, $p = 0.013$, $\eta^2 = 0.228$), but the interaction with visual mode (icon vs. image) is not significant. While subjects did very well in identifying nouns, they had more difficulty with adjectives and verbs (see Figure 2). Images were better for understanding nouns while, on the whole, subjects performed slightly better with icons in all other categories.

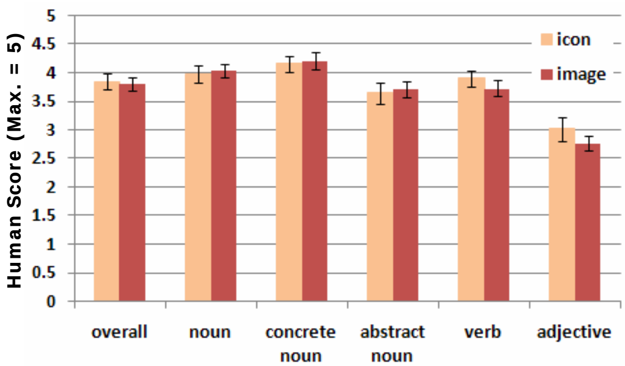

Figure 2. Comparison of image vs. icon representation for different parts of speech (y axis is quality of match).

To identify the concepts where icons and images had disparate effectiveness in communicating the underlying meaning, we created a median score for each word across all subjects for both icon and image codings. Figure 3 presents the words found to have significant disparities in icon and image representation using an analysis of variance. There were several reasons why people misinterpreted these words. "First, the icon for the noun "vacation" was interpreted as "beach," and that for "home" was interpreted as "house." In contrast, the image for the verb "help" used such a complex scene that many participants failed to focus on the action. For other words, no real difference was found between the icon and image scores.

Both images and icons performed equally poorly in capturing the essence of adjectives. Adjectives like "white", "first", and "little" were rarely correctly interpreted and seemed to be the most likely cause for confusion in interpreting images, as subjects tried to cast adjectives into nouns ("I am a confused eater." instead of "I am hungry."). Abstract nouns like "trouble" also presented difficulty for both images and icons.

4.4 Sentence-level Effects

Significant differences were found between those who saw images and those who saw icons in two of the 25 phrases (Figure 3: SLP1 and SLP4). In SLP1, people misinterpreted "help" using a wider range of interpretations for the image than the icon. In SLP4, the image for the noun "stop" was often interpreted as a verb so that the phrase "Where is the bus stop?" became "Where does the bus stop?" Overall, no significant differences were found between images and icons at the sentence level. Sentence properties such as length, number of pictures embedded, and picture-to-word ratio were also not found to differ significantly.

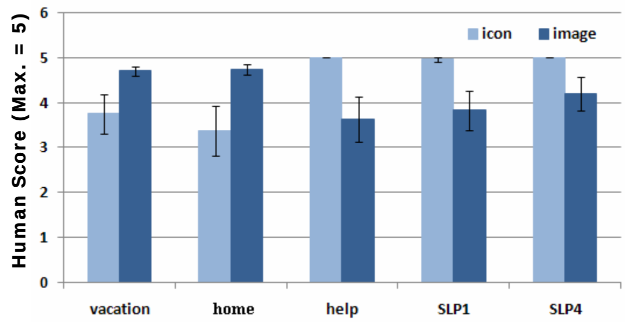

Figure 3. median coding scores for words and phrases with significant differences between icon and image

The reader should note that at this point, we are not claiming that no significant differences exist, only that they have not been found in this study. However, the results were promising and suggested that we continue with Experiment 2 in which we run our target population, increase the number of subjects and look at the effect size of the study.

5. EXPERIMENT 2 DESCRIPTION

In the second study, we presented both Lingraphica icons and web images to individuals affected by aphasia and asked them to choose, from a list of five words, the word that best represented the meaning conveyed by the picture. Because the extreme variability in impairment prevented us from making comparisons across individuals, we showed both icons and images to each participant and conducted a within subjects design. Since the first study found that neither icons nor web images performed well when illustrating verbs and adjectives, we used only nouns in the second study. There was also a bias towards icons because the aphasic subject group had experience using software that uses icons for language rehabilitation.

5.1 Experiment 2 Participants

Fifty individuals with aphasia were recruited from a local aphasia center for the study, including 15 females and 35 males. The average age of the participants was 60 and ranged from seven individuals under 50 to two over 90. While slightly over half of the participants had some computer experience before they had aphasia, all were taking computer lessons at the center. The coordinators of the center provided three broad groupings of impairment via computer-based assessments: 27 of the subjects were high functioning, meaning that they had little difficulty in linking pictures to words; 18 were medium functioning, meaning that they could understand some of the words, but not the entire sentence; and 5 were low functioning, meaning that they had trouble even understanding single written words.

5.2 Experiment 2 Methods and Interface

In this study, we presented both icons and images to individuals affected by aphasia and asked them to choose, from a list of five words, the word that best represented the meaning conveyed by the picture. Each word came with an embedded sound of its pronunciation assisting participants with impaired reading comprehension (a mechanism derived from Lingraphica). This experiment is different from Experiment 1 in that subjects are recognizing words not recalling them. The Experiment 1 task is much more difficult for people with aphasia. Experiment 1, thus, helped us in selecting images and icons for this task.

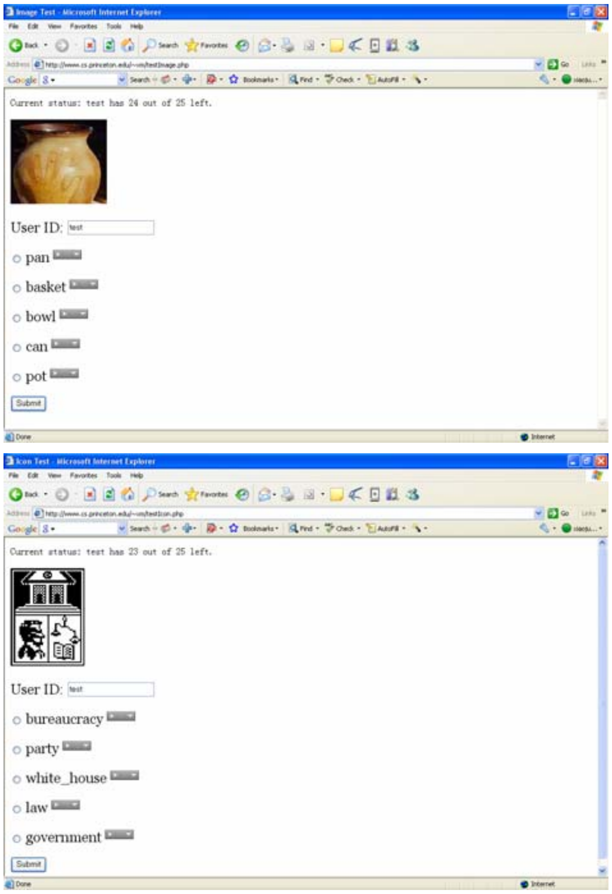

Figure 4. Exp. 2 screens: images (above) and icons (below)

The study was conducted on a web-based interface (Figure 4). Note that the layout favored the left-hand side of the screen. This prevents subjects with right-field neglect (which commonly co-occurs with aphasia) from ignoring elements on the right. This also allowed us, through the use of embedded recorded speech sound files, to offer consistent auditory clues for subjects who are unable to read individual words. We ensured that each page did not exceed a single screen, thus, eliminating the need to scroll. Users were allowed to signal their decisions by directly using the mouse, pointing at the screen, or saying the answers out loud.

5.3 Experiment 2 Data

Twenty-five nouns, of which nine are abstract and 16, concrete, were selected from the sentences used in Experiment 1. For each word, an icon or an image representing the chosen noun was displayed along with five possible word options for the subject to choose from. All the icons and images came from Experiment 1 except for those nouns that were found to cause difficulties.

The alternatives for each noun were chosen in order using the following scheme. 1) WordNet: sister terms (other direct hyponyms of its direct hypernym), part meronym, uncle terms (sister terms of its direct hypernym). 2) ESP labels. 3) Words selected from erroneous data collected in Experiment 1.

We chose alternatives that could readily replace the target in a sentence without violating grammatical form or general meaning, that is, the alternatives could also be used in the same or similar context. No complicated or uncommon words that aphasics would have difficulty with were included.

5.4 Experiment 2 Task

Each participant was instructed to provide answers for all 25 nouns, half of which were represented by icons and the other half by images. This is a departure from our first study; while in the first study we wanted to guard against internal bias, with aphasics we felt it more important to account for the extreme variance in ability by having subjects attempt both modalities. The order of the nouns was kept the same. The first half was either represented by icons or images and the latter half by the opposite. Subjects were randomly assigned to receive the icons or the images first in a balanced study. Each subject selection was verbally confirmed and pointed to by the experimenter after the subject had indicated a selection. It was then recorded as was the time it took to answer each question.

6. RESULTS, EXPERIMENT 2

This section seeks to examine differences between the users' response to pictorial stimuli in our trial. We first explore various methods of gauging the accuracy of responses and, finding little difference, then examine if the different inputs have discernibly different patterns with regards to the time in which it takes users to determine the appropriate word. Lastly, we examine differences by restricting our data to subgroups of the population.

6.1 Measuring how pictures evoke meaning

We first used error rate, which is the proportion of words in which the selected answer does not match the target answer. The error rate of icons is higher than that of images in 13 out of 25 nouns. Figure 5 shows the words that were found to have a significant difference in error rate between web images and icons.

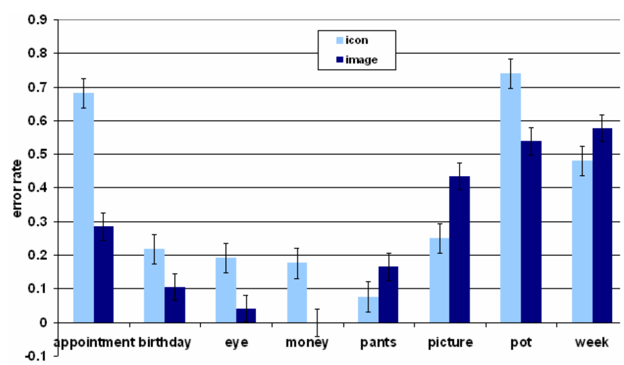

Figure 5. Nouns for which significant differences were found between icons and images

However, error rate is not a satisfying measure of how well pictures evoke words. Because we hand-selected the options for the pictures, some "wrong" choices are closer to the targets than others. A uniform distribution over responses means that the picture did not shape the user's choice and a peaked distribution shows that one word was preferred over the alternatives. Another type of comparison would capture this.

One such measure is entropy, which gives low scores if users agree on any word and high scores for distributions that are more spread out. It also removes the assumption that one word is correct and allows images and icons to elicit different words without penalty (although this was not the case). The entropy of choice distribution across five options for each noun were calculated with icon, image, and overall responses computed separately using the standard entropy equation,

$$H(p) = -\sum_i p_i \log_2 p_i$$

The probabilities of selection were smoothed via Laplace smoothing, which adds one to each of the counts for all words. The higher the entropy is, the less people agreed with each other on the concept represented in the picture. ANOVA results (for mode: $F(1, 9) = 1.748$, $p = 0.219$, $\eta^2 = 0.163$; for mode and parts of speech interaction: $F(1, 9) = 2.292$, $p = 0.164$, $\eta^2 = 0.203$ implied that the difference between images and icons in entropy is not significant in general or when compared across concrete and abstract nouns. Table 4 gives the means for accuracy and entropy showing that images slightly outperformed icons although not significantly. However, there is a big effect size (> 0.14), meaning that a larger N might find significance. Entropy had nearly perfect (0.95) correlation with error, so subsequent comparisons will be made with this easier to understand metric.

Table 4. Accuracy and entropy values for icons and images

	Accuracy		Entropy	
	Mean	Std. dev	Mean	Std. dev
icon	0.76	0.22	0.41	0.12
image	0.77	0.19	0.39	0.10

Given the similar entropy between images and icons, the next step is to ask whether the distribution over responses is different. Figure 6 shows the top eight words in which the distributions varied the most. Although for "birthday" and "pot" there was a significant difference between icons and images, taken in aggregate the data do not support a clear superiority of one or the other. The five nouns (four are abstract) for which both pictorial representations were the most confusing were "appointment", "government," "pot," "trouble," and "week." For both icons and images, abstract nouns are harder to interpret precisely.

6.2 Speed and Accuracy

In order to run comparisons across subjects, we took the time logged by the system for each noun and computed the variance-normalized time that was spent on viewing the picture before making a selection decision. Thus, for scores above zero, subjects took longer selecting the word. For icons, it took longer for people to figure out "week" (1.16.), "government" (1.14.), "appointment" (1.07), "pot" (0.78), and "trouble" (0.53); for images, more time was spent on "trouble" (1.63), "pot" (0.74), "government" (0.47), "picture" (0.37), and "leg" (0.28). Although most of the time people got a concept from an icon faster than from a web image, ANOVA results ($F(1, 48) = 0.759$, $p = 0.388$, $\eta^2 = 0.016$) did not find significant differences.

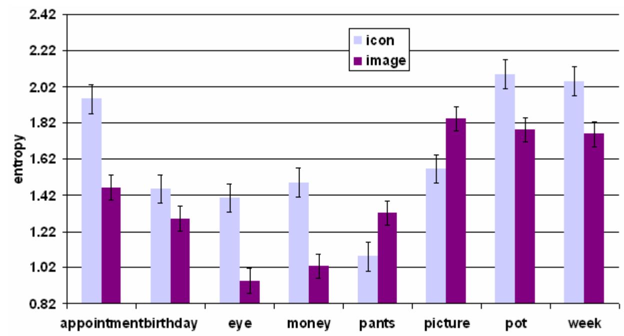

Figure 6. Nouns which differed the most on entropy, that is, subject agreement on a selected word

Subjects either grasped the concept represented by a picture immediately or spent more time on interpreting a picture but still did not comprehend it. The correlation between error rate and standardized time is 0.86 for icons and 0.84 for images.

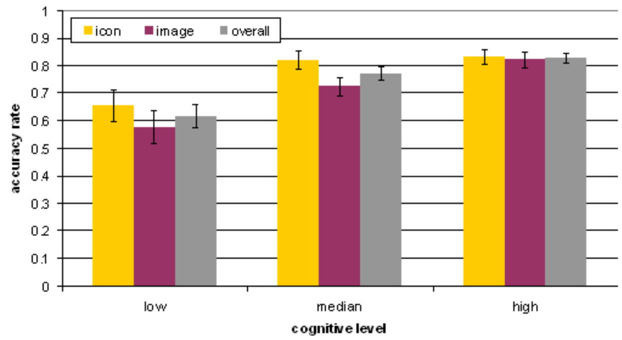

Figure 7. Image vs. Icon at different cognitive levels.

We also observed that people with high and medium cognitive levels were faster with images, while those with low cognitive level were faster with icons. In terms of accuracy, images and icons worked equivalently well for people with high cognitive level, while for the others, icons did a better job (Figure 7). The differences, however, are not significant. In addition, the metric of cognitive level we use is based on the informal evaluation of our subjects' cognitive abilities by the coordinators in the aphasia center. Since we did not have access to results from a formal evaluation, our metric is not precise, but the results may have implications for further investigation in our future work.

7. Discussion

Experiment 1 results suggest that images, despite an uncertain provenance and lacking uniform conventions, may be as effective as icons in conveying meaning. Experiment 1 was also useful in uncovering ambiguous images and problems with parts of speech helping to improve the stimuli used in Experiment 2.

As with any study with individuals with aphasia, our target population is so diverse that it makes drawing conclusions about a "typical" individual difficult, despite our relatively large sample size. Experiment 2 was not designed to draw any conclusion about speed and accuracy. An individual who takes longer on the test could be slowed by serious expressive and/or motor disability, determination and nervousness, or by listening intently to the associated pronunciations multiple times.

Although we provided a magnifying glass in the experiments, the low resolution and poor contrast of some images might impact people's interpretation. Web images which are not standardized in quality may introduce heavier processing loads. Processing images via special visual enhancement tools before using them in assistive technologies could address such problems.

Experiment 2 results were consistent with those reached in Experiment 1 showing that icons were no better than images for illustrating nouns. Despite moving from a free response test to a multiple choice test, replacing unimpaired elderly users with disabled individuals, and moving away from a sentence context, our results were analogous. Despite doubling the number of subjects and working with a cognitively challenged group, we were not able to find an advantage of icons over images with the exception that icons may be easier to process visually.

8. CONCLUSION AND FUTURE WORK

The results of the two studies suggest that icons are no more effective than images in conveying noun concepts visually for people with aphasia and that they perform well for representing other parts of speech. Although no significant differences were found, the large effect size implied that with a bigger N, images may outperform icons in conveying meaning, but this effect may be countered by a person's cognitive level. As mentioned in the rationale for our work, images do overcome the shortcomings of icons in breadth and depth and therefore may have the potential to support the vocabulary needs of our target population.

In future work, we will extend the exploration of effective visual representations to other parts of speech. Our initial results have led us to build a pictorial dictionary database that maps collections of web images and other online multimedia to concepts. This database will serve as the basis for creating a rich automated visual vocabulary that extends the traditional icon-based vocabularies. Such a vocabulary could radically expand the breadth of communication available to users with aphasia. Using web images for communication could also decrease the cost of assistive devices, giving more patients access to such devices.

9. Acknowledgements

First and foremost, we thank the members of the Princeton Senior Center and the Adler Aphasia Center for their time in this study. We additionally wish to thank Marilyn Tremaine, Christiane Fellbaum, Barbara Purves, Joanna McGrenere, and Dan Osherson for valuable comments and advice. This research was supported by a grant from Microsoft, and a gift from the Kimberley and Frank H. Moss '71 Research Innovation Fund of the Princeton School of Engineering and Applied Science.

10. REFERENCES

[1] Ageless Project. http://jenett.org/ageless/. 2006

[2] Allen, M., McGrenere, J., and Purves, B. The Field Evaluation of a Mobile Digital Image Communication Application Designed for People with Aphasia, ACM Transactions on Accessible Computing, 1(1), 1-26, 2008

[3] Boyd-Graber, J., Nikolova, S., Moffatt, K., Kin, K., Lee, J., Mackey, L., Tremaine, M., and Klawe. M. Participatory design with proxies: Developing a desktop-PDA system to support people with aphasia. In *Proc. CHI 2006*, 151–160. ACM Press, 2006.

[4] Danielsson, H. and Jonsson, B. Pictures as language. *In Proc. International Conference on Language and Visualization 2001*, Stockholm, Sweden, 2001.

[5] Deng, J., Dong, W., Socher, R., Li, L-J., Li, K. and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *Proc. CVPR 2009,* IEEE Press, 2009.

[6] DynaVox. http://www.dynavoxtech.com/. 2006

[7] Fellbaum, C. WordNet: Electronic Lexical Database, A semantic network of English verbs. MIT Press, 1998.

[8] General Text Matcher. http://nlp.cs.nyu.edu/GTM/. 2006

[9] Goodglass, H., Kaplan, E. and Barresi, B. The assessment of aphasia and related disorders (3rd Ed.). Philadelphia: Lippincott Williams & Wilkins, 2001.

[10] Hochberg, J. and Brooks, V. Pictorial recognition as an unlearned ability: a study of one child's performance", American Journal of Psychology, 75, 624-628. 1962

[11] Kjelldahl, L. A Survey of Some Perceptual Features for Computer Graphics and Visualization. In *Proc. Linkoping Electronic*, 2003

[12] Lingraphica. http://www.lingraphicare.com/. 2005

[13] Lucia, C. Toward Developing a Model of Music Therapy Intervention in the Rehabilitation of Head Trauma Patients. *Music Therapy Perspectives*. 4, 34-39.

[14] Moffatt, K., McGrenere, J., Purves, B., and Klawe, M. The Participatory Design of a Sound and Image Enhanced Daily Planner for People with Aphasia. In *Proc. CHI 2004*, 407-414. ACM Press 2004.

[15] Nadel, S. A field experiment in racial psychology. British Journal of Psychology, 28, 195-211. 1937

[16] National Aphasia Association. http://www.aphasia.org.

[17] Prentke Romich Company. http://www.prentrom.com/. 2006

[18] Ryan, T, and Schwartz, C. Speed of perception as a function of mode of representation. American Journal of Psychology, 69, 193-199. 1956

[19] Rubin, E. Figure and ground. In DCW Beardslee, M. (Ed.), Readings in perception. New York: Van Nostrand, 1958.

[20] Steele R., Weinrich M, Wertz RT, Kleczewska MK, Carlson GS. Computer-based visual communication in aphasia. Neuropsychologia; 27(4):409-26. 1989

[21] Thorburn, L., Newhoff, M., and Rubin, S. Ability of Subjects with Aphasia to Visually Analyze Written Language, Pantomime, and Iconographic Symbols. American Journal of Speech Language Pathology, 4(4): 174-179, 1995

[22] Van de Sandt-Koenderman, M., Wiegers, M., and Hardy, P. A Computerized Communication Aid for People with Aphasia. *Disability Rehabilitation*, 27(9): 529-533, 2005

[23] Von Ahn, L. and Dabbish, L. Labeling Images with a Computer Game. In *Proc. CHI 2004*, 319-326. ACM Press, 2004